Multiple Retropseudogenes from Pluripotent Cell-specific Gene Expression Indicates a Potential Signature for Novel Gene Identification*

Received for publication, December 16, 2004
Published, JBC Papers in Press, January 6, 2005, DOI 10.1074/jbc.C400587200

Debleena Pain*, Gung-Wei Chirn*, Christopher Strassel¶, and Daniel M. Kemp**

From the ‡Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 and the ¶Diabetes & Metabolism Disease Area, Novartis Institutes for BioMedical Research, Cambridge, Massachusetts 02139

Oct4, Nanog, and Stella are transcription factors specifically expressed in embryonic stem (ES) cells and germ lineage cells that impart critical functions in the maintenance of pluripotency. Here, we report the excessive frequency and apparent selectivity of retropseudogenes for Oct4 cell-specific genes. Six highly homologous pseudogenes for Oct4, 10 for Nanog, and 16 for Stella were identified by nucleotide BLAST (basic local alignment sequence tool) searches against the respective gene mRNA transcripts. Of 15 non-ES cell-specific transcription factor genes, only one had a single pseudogene hit in our screen, emphasizing the apparent selectivity. We present a hypothesis whereby retropseudogenes of ES or germ cell-specific genes may reflect an innate predisposition. This is based on the increased probability of germ-line transmission when retropseudogenes occur at a very early stage of development within cells known to contribute to the germ cell lineage. The parental genes for Nanog, Stella, and another embryonic gene, GDF3 are all located on chromosome 12p13 of the human genome, and on chromosome 6 in mouse. Here, we identified an Oct4 pseudogene at the same respective loci in both human and mouse genomes, suggesting functional relevance and indicative of epigenetic regulation. We tested whether the apparent selectivity for ES cell-specific gene retropseudogenes may be extrapolated to a more unified phenomenon, such that a bioinformatic approach may represent a potentially novel strategy for identification of genes with embryonic cell-specific functionality. A preliminary investigation indeed revealed a single gene, previously demonstrated to be responsible for multiple retropseudogenes via germ cell-specific expression in Xenopus.

Embryonic stem cells are pluripotent and able to differentiate into all cell types of the body, including germ cells (1).

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

** To whom correspondence should be addressed: Diabetes & Metabolism Disease Area, Novartis Inst. for BioMedical Research, 100 Technology Square, Cambridge, MA 02139. Tel.: 617-817-0105; Fax: 617-871-7051; E-mail: Daniel.kemp@pharma.novartis.com.

Identifying the molecular determinants by which pluripotency is maintained is a critical pursuit that will ultimately result in expedient therapeutic application. To date, four transcription factors have been identified that exhibit essential roles in maintenance of the pluripotent state. These are Oct4 (2), Nanog (3, 4), Stella (5), and GDF3 (6). The expression level of all four of these genes is down-regulated coincident with ES cell differentiation, and loss of expression leads reciprocally to differentiated cell states. Beyond maintenance of pluripotency, the functional roles of these genes have not been clearly defined, and the prospective identification of other novel ES cell genes will undoubtedly contribute to this endeavor. Notably, Oct4 is detected in adult stem cells such as bone marrow-derived mesenchymal stem cells (7) and thought to play a role in maintaining the multipotency of these cells also. Recently, 10 pseudogenes of Nanog were reported within the human genome (8), likely a result of retropseudogene events. Characterized by extensive alignment and significant sequence identity with the Nanog mRNA transcript, these pseudogenes were shown to be located on multiple chromosomes. Functional relevance of the pseudogenes of Nanog remains undetermined, although they appear not to be expressed as transcripts in human ES cells (8).

In the present study, we identified extensive retropseudogenes as a potential signature event of ES cell-specific genes. Of the four known ES cell-specific genes, we found that Nanog, Oct4, and Stella, each possess multiple pseudogenes. Our data suggest that retropseudogene frequency of ES cell-specific genes far exceeds that of non-ES cell-specific genes. Moreover, we propose that this “pseudogene signature” may prove useful in identifying novel ES cell-specific genes in future.

MATERIALS AND METHODS

The basic local alignment sequence tool (BLAST)1 was used to compare the cDNA of Nanog, Oct4, Stella, and GDF3 with the NCBI human genome data base (9). An E value limit was set to 1 × 10-5 for these searches. BLAST was used to generate a map depicting alignments of the pseudogenes to the parent genes. cDNA sequences of transcription factors, critical to the phenotypes of pancreatic β cells and skeletal muscle cells, were screened against the human genome in a similar BLAST search. The Ensembl genome browser was used to confirm all blast hits and was used to create a map showing the chromosomal locations of the hits for Nanog, Oct4, and Stella.

For a given set of pseudogene sequences, a BLASTTable data base was created using formatdb. A BLAST search was then run to compare the original gene sequence against the pseudogene, with the default BLAST parameters, except for an E value cutoff of 1 × 10-5. The results of the search were run through an in-house developed tool to transform the results into an image, allowing the user to view the lengths and positions of all high scoring pairs, relative to the query sequence. This tool was built using open source BioJava tools for handling the BLAST data and the Batik toolkit for generating the images.

Vista and BLAST were employed to identify sequence alignments between chromosome 12p13 and chromosome 6 of the human genome. Sequence homologies ≥80% and at least 1000 base pairs in length were selected for further analysis. The Celera genome browser was used to identify these sequences. One million base pairs of DNA segment at chromosome 12p13 was investigated. The computer program Repeatmasker was applied on the sequence to mask out the repeat sequences. Then the masked sequence was split into 3000-bp units and BLASTed against the human chromosome 6 to look for homologous segments. The filter criteria was set for homolog search to 200 base pairs, with 95% sequence identity.
identity. All segments that matched the criteria were further filtered by NCBI Refseq to remove known genes. Finally, the public human EST database was searched for evidence of expression, as was the Entrez nucleotides data base.

RESULTS

Four human ES cell-specific genes have been identified to date, such that their expression is strictly required to maintain the phenotype of ES cells (2–6). Nanog, Oct4, Stella, and GDF3 have been characterized as genes, whose function is known to maintain pluripotency. We observed by BLAST analysis that three of these four genes possess multiple, highly conserved pseudogenes that are dispersed throughout the human genome (Fig. 1). Briefly, mRNA transcript sequences of these genes were individually BLASTed against the NCBI human genome data base, and multiple hits were generated that were identified as pseudogenes for Nanog, Oct4, and Stella. Hits that represented the parent genes contained the expected intron/exon gene organization, but all other sequences were contiguous, such that the conserved alignments reflected the mRNA transcript sequence rather than the genomic sequence. This provided strong evidence that these conserved sequences likely resulted from retrotransposition events. We observed no conserved hits for GDF3 within the limits of experimental stringency. As a control experiment, we searched for pseudogenes using mRNA sequences of a selection of distinct cell type-specific genes. Specifically, we focused on genes that encoded transcription factors with known specificity to ES cells, pancreatic β cells, and skeletal muscle cells were BLASTed against the human genome. The stringency (E value) was set to 0.0001. BLAST hits of high homology indicated the presence of pseudogenes. All hits were characterized to ensure that they were not limited to conserved domains of a protein but reflected full-length retrotransposition events.

The BLAST program was used to create a map of relative alignment of the pseudogenes for each ES cell-specific gene (Fig. 2). Notably, certain regions of the Nanog transcript were highly conserved among pseudogenes, whereas other regions were less conserved. Specifically, the Nanog coding region, spanning positions 217–1134 of the mRNA transcript (Fig. 2A), was found in 70% of the pseudogenes with sequence identity conserved at ~90% or greater. In contrast, the 3′-untranslated region of the Nanog mRNA appeared to have undergone significant sequence degradation, although sufficient identity remained to confirm its initial transposition. Pseudogene 8, located on chromosome 15 showed greatest similarity to Nanog, while pseudogenes 3, 6, and 11 displayed the least identity.

We identified six pseudogenes of Oct4 (Fig. 2B). Interestingly, pseudogene 3 was found to be located on chromosome 12p13 of the human genome. This locus is significant as it also harbors the Nanog, Stella, and GDF3 parent genes (Fig. 3). The BLAST hit for this Oct4 pseudogene equated with an E value of 0, indicating that the random chance of such a sequence match within the genome is insignificant. Moreover, this pseudogene contained the entire coding region of Oct4 with a sequence identity of 94%. The exact location of this Oct4 pseudogene was traced to a position within an intron of another gene, designated CLECSF6, between exons 3 and 4. CLECSF6 is a member of the C-type lectin/C-type lectin-like domain (CTL/CTL) superfamily. Astonishingly, Nanog, Stella, GDF3, and Oct4 pseudogene 3 collectively span 445 kb of human genomic sequence, which represents a region of only 1/10,000 of the entire human genome (Fig. 3).

Through analysis of the mouse genome, we also found that an Oct4 pseudogene was located in the same locus as Nanog, Stella, and GDF3 on chromosome 6. Moreover, the relative positional order of these genes was conserved between the mouse and human genomes. By BLASTing the EST data base we found that this mouse pseudogene is likely transcribed, as an exact sequence hit was generated (data not shown). This suggests that the mouse oct4 pseudogene, which colocalizes with Nanog, Stella, and GDF3 is transcriptionally functional.

Our search for Stella pseudogenes generated 16 hits with remarkable sequence homology (Fig. 2C). Chromosomes 14 and X contain Stella-like genes with predicted expression. The parent Stella gene on chromosome 12 contains 97% sequence homology with a 1058-bp region on chromosome 14 and 96% homology with a 1068-bp region on chromosome X. Moreover, the Gnomon gene prediction method predicts that this region on chromosome X corresponds to a gene whose cDNA consists of 482 base pairs, and the region on chromosome 14 contains a Stella-like gene whose cDNA consists of 589 base pairs.

A preliminary effort to determine whether additional pseudogene-related sequences exist within Ch12p13 was carried out. We blasted a region of one million nucleotides of this locus against the entire length of chromosome 6. Chromosome 6 was selected because it appeared from previous searches to represent a preferred retrotransposition target. Besides the expected hits described previously, this screen yielded a conserved 387-bp region of >90% identity between these two chromosomes (Fig. 3). This same sequence was also shown to display significant homology with another genomic region on chromosome 9 and also mapped to eight publically available ESTs. Within chromosome 12p13, this 387-bp sequence was located proximal to DKFP566B183, a gene of unknown function. Further analysis confirmed that the actual source of the sequence was the gene elongation factor 1a (eEF-1a) on human chromosome 6. Notably, this gene was previously identified to have undergone extensive retrotransposition to form multiple pseudogenes within human, mouse, and pig genomes (11, 12).

DISCUSSION

We report the existence of multiple pseudogenes for three of the four known markers of pluripotency, Nanog, Oct4, and Stella. The evident frequency of germ line-transduced retrotransposition of these genes is extensive, relative to that of non-ES cell-specific genes (Fig. 1). Furthermore, the degree of sequence conservation to the parent genes suggest functional utility and possible evolutionary implications. For example, Nanog pseudogenes 2, 4, 5, 7, 9, and 10 appear to have retained extensive homology within the Nanog coding region but show significant degeneration of sequence identity in peripheral sequences such as the 3′-untranslated region of the parent mRNA sequence. Whether these inconsistencies are related to the age of retrotransposition events, to the susceptibility of the
respective loci to sequence degradation, or to advantages in evolutionary selection remains to be defined.

A rational explanation for the surprising frequency of ES cell-specific pseudogenes may be drawn from lessons in developmental biology. It is assumed that retrotransposition of a transcript may only result in a *bona fide* pseudogene if the "event" cell subsequently contributes to the germ line. Note that within this paradigm, the event cell denotes a cell in which a retrotransposition event occurs. Importantly, the probability of contribution to the germ line is greatly enhanced when the event cell is an ES cell (ES-event cell) or a germ lineage cell. As ES cells are known contributors to the germ cell lineage (1) and given that they exist at the earliest developmental stages, the probability of an ES-event cell contributing to the germ line far exceeds that at later stages of development. Notably, within this relatively permissive environment of the blastocyst, the ES-event cell expresses genes including Nanog, Oct4, and Stella at relatively high levels. Therefore, the probability of

![Alignment score](image)

**Fig. 2.** Pseudogene alignments with parent gene mRNA transcripts. A–C, pseudogene sequence alignments of Nanog (A), Oct4 (B), and Stella (C). Across the top of each alignment panel is the query cDNA sequence of the respective genes. Black arrowheads identify the borders of coding regions. The degrees of homology of pseudogenes to the parent gene are represented by the color code depicted in the legend, which signifies conserved length of sequence. The relative direction of sequence is indicated within the homology bars.
Fig. 3. Location of ES cell-specific genes on human chromosome 12p13. All of the transcription factors currently known to maintain pluripotency of ES cells can be mapped to the human chromosome 12p13. Oct4 pseudogene 3 maps to the intronic sequence of CLECSF6. Additional genes at this locus are indicated alphabetically: a, APOBEC1; b, CLEC5F7; c, SLC2A14; d, SLC2A3; e, FHX; f, C3AR1; g, CLECSF6. The asterisk represents the 387-bp sequence that mapped with chromosome 6 with >90% identity. CLECSF6 is expanded below to show numbered exons (1–6) and the Oct4 pseudogene located within intron 3. This entire sequence spans ~500 kb.

these genes contributing to a retrotransposition event is relatively high. We propose that our data support this hypothesis such that ES cell-specific genes are more susceptible to retrotransposition and subsequent pseudogene generation than are somatic cell-specific genes. That this disparate distribution of pseudogenes is observed in favor of ES cell-specific genes is suggestive of retrotransposition being the most likely mechanism rather than chromosomal rearrangement. This is because a process of gene duplication by chromosomal rearrangement would likely not be selective for ES cell-specific genes. More importantly, all of the sequences we identified in this study were intron-less with respect to the parent gene, providing convincing evidence that the pseudogenes originated from retrotransposition events.

Nanog, Stella, GDF3, and an Oct4 pseudogene all map to a region of 445,088 base pairs on chromosome 12p13. This region represents only 0.01% of the genome, reflecting an impressive concentration of clustered ES cell-specific genes. A potential advantage of such localization can be hypothesized given the simultaneous regulation pattern of these genes. Moreover, a mechanism of epigenetic regulation would be consistent with their common function, to maintain ES cell pluripotency. Recently, it was demonstrated that DNA methylation status is closely linked to the chromatin structure of the Oct4 gene and that this epigenetic mechanism plays a functional role in the developmental stage- and cell type-specific expression of Oct4 (13). Whether the Ch12p13 hotspot for ES cell-specific genes is also epigenetically regulated remains to be determined. We observed that a similar gene localization pattern is found in the mouse genome, whereby Nanog, Stella, GDF3, and an Oct4 pseudogene map to chromosome 6 with the same relative orientation as in human. Notably, the NCBI data base predicts that the mouse genome contains multiple such pseudogenes. Given that this equivalent pseudogene appears to exist in the human and mouse genomes, we surmise that the appearance of this pseudogene occurred prior to the evolutionary diversification of mouse and human, and the fact that these inter-species sequences remain conserved to such a degree, suggests functionality and evolutionary relevance.

Nanog, Oct4, and Stella were initially identified as ES cell-specific genes by observations that their expression was significantly down-regulated during ES cell differentiation. In fact, restricted expression analysis is currently the primary method to identify genes that are essential for ES cell specificity. We suggest that a novel method for identifying ES cell-specific genes may emerge from our observation that these genes in particular are predisposed to retrotransposition. Specifically, we hypothesize that intra-genome BLASTing may identify regions of sequence homology and, with appropriate filtering mechanisms, aid in the search for novel ES cell-specific factors. In a preliminary study, we identified a ~400 base pair region on human chromosome 6 with >90% identity to a sequence within chromosome 12p13. Further analysis revealed that this sequence derived from the human eEF-1α, previously identified as the origin of multiple retropseudogenes. In fact, retropseudogenes have been shown to constitute the major part of the human eEF-1α gene family (11). Moreover, it was determined that the derivation of eEF-1α retropseudogenes in *Xenopus* laevis was generated from germ cell-specific expression of the parent gene (14), consistent with our findings and proposed hypothesis.

Two recent studies recently exemplified the process of retrotransposition in embryonic cells, based on lineage tracing of retrotransposition events in transgenic mice (15, 16). Notably, in both cases, retrotransposition occurred at a very early stage in the development of the founder mice and displayed high frequency of chromosome-to-chromosome retrotransposition within the male germ line of these animals. Consistent with our observations in the present study, these data further validate our theory.

Finally, it is well established that retrotranscription correlates with genomic evolution (17). Creation of new genes with new functions arise predominantly via retrotransposition and chromosomal rearrangements. The implication for retrotransposition of embryonic genes such as Nanog, Oct4, and Stella, in an evolutionary context, is an attractive notion. Further efforts to annotate these retropseudogenes and determine their functional relevance will reveal whether such an evolutionary mechanism is apparent.

Acknowledgments—We thank Drs. En Li, Susan C. Stevenson, and Thomas E. Hughes for helpful discussions.

References

1. Geijnsen, N., Horoschak, M., Kim, K. Gribnau, J., Eggen, K., and Daley, G. Q. (2004) *Nature* 427, 148–154
2. Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Kleva-Nebenius, D., Chambers, I., Scholer, H., and Smith, A. (1998) *Cell* 95, 379–393
3. Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003) *Cell* 113, 643–655
4. Mitsui, K., Tokuzawa, Y., Ishi, H., Segawa, K., Murakami, M., Takashashi, K., Maruyama, M., Maeda, M., and Yamakawa, S. (2000) *Cell* 113, 631–642
5. Payer, B., Saitou, M., Barton, C. S. Thresher, R., Dixon, J. P., Zahn, D., Colledge, W. H., Carlton, M. B., Nakano, T., and Surani, A. M. (2003) *Curr. Biol.* 13, 2110–2117
6. Caricase, A. A., van Schuij, R. H., Zeinstra, H. M., Wierikx, C. D., van Gurp, R. J., van den Pol, M., Looijenga, L. H., Oosterhuis, J. W., Per, M. F., Ward, A., deBrujin, D., Kramer, P., deJong, F. H., and Van den Eijnden-van Raaij, J. A. J. (1998) *Hum. Mol. Genet.* 16, 95–103
7. Pochampally, R. R., Smith, R. J., Ylostalo, J., and Prokop, D. J. (2004) *Blood* 103, 1647–1652
8. Booth, H. A., and Holland, P. W. (2004) *Genomics* 84, 229–238
9. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403–410
10. Clark, A. T., Rodriguez, R. T., and Bodnar, M. S. (2004) * Stem Cells 22*, 169–179
11. Madsen, H. O., Poulsen, K., Dahl, O., Clark, B. F., and Hojrt, J. P. (1999) *Nucleic Acids Res.* 18, 1513–1516
12. Land, A., Knudsen, S. M., Vissing, H. Clark, B., and Tommerup, N. (1996) *Genomics* 36, 359–361
13. Hattori, N., Nishino, K., Ko, Y. G., Hattori, N., Ohgane, J., Tanaka, S., and Shintz, K. (2006) *J. Biol. Chem.* 279, 17063–17069
14. Abdalla, B., Hourdy, J., Krieg, P. A., Denis, H., and Mazabraud, A. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 9277–9281
15. Ostertag, E. M., deBerardinis, R. J., Goodman, L. J., Zhang, Y., Yang, N., Gerton, G. L., and Kazazian, H. H., Jr. (2004) *Proc. Natl. Acad. Sci. U.S.A.* 100, 1832–1837
16. Devor, E. J., and Moffat-Wilson, K. A. (2003) *Hum. Biol.* 75, 661–672