



OPEN

CCIVR facilitates comprehensive identification of cis-natural antisense transcripts with their structural characteristics and expression profiles

Tatsuya Ohhata^{1,3}, Maya Suzuki^{1,3}, Satoshi Sakai¹, Kosuke Ota¹, Hazuki Yokota¹, Chiharu Uchida², Hiroyuki Niida¹ & Masatoshi Kitagawa¹

Cis-natural antisense transcripts (cis-NATs) are transcribed from the same genomic locus as their partner gene but from the opposite DNA strand and overlap with the partner gene transcript. Here, we developed a simple and convenient program termed CCIVR (comprehensive cis-NATs identifier via RNA-seq data) that comprehensively identifies all kinds of cis-NATs based on genome annotation with expression data obtained from RNA-seq. Using CCIVR with genome databases, we demonstrated total cis-NAT pairs from 11 model organisms. CCIVR analysis with RNA-seq data from parthenogenetic and androgenetic embryonic stem cells identified well-known imprinted cis-NAT pair, *KCNQ1/KCNQ1OT1*, ensuring the availability of CCIVR. Finally, CCIVR identified cis-NAT pairs that demonstrate inversely correlated expression upon TGF β stimulation including cis-NATs that functionally repress their partner genes by introducing epigenetic alteration in the promoters of partner genes. Thus, CCIVR facilitates the investigation of structural characteristics and functions of cis-NATs in numerous processes in various species.

Natural antisense transcripts (NATs), first discovered in bacteria as early as 1981¹, are the transcripts encoding complementary sequences to other RNA transcripts^{2,3}. In contrast to trans-NATs whose partner genes are transcribed from different genomic loci, cis-NATs fully or partially overlap their partner genes but are transcribed from the opposite DNA strand, and some of them function in the regulation of gene expression^{4,5}. Cis-NATs regulate gene expression at different levels. At the level of transcriptional regulation, a cis-NAT negatively regulates its partner gene by interfering with recruitment of RNA polymerase II to its overlapping region (e.g., *Airm*⁶ and *qrf*⁷), by depositing repressive epigenetic modifications on the promoter of its partner gene (e.g., *Tsix*^{8,9}), and by recruiting epigenetic repressors, such as G9a, PRC2, and PRC1 (e.g., G9a and PRC2 by *Kcnq1ot1*¹⁰; PRC2 by *ANRIL*¹¹; PRC1 by *ANRIL*¹²). In contrast, cis-NATs positively regulate their partner genes by forming RNA-DNA-DNA triplexes that recruit active epigenetic regulators to the regulatory elements of the partner gene (e.g., *KHPS1*¹³ and *TCF21*¹⁴). At the level of post-transcriptional regulation, cis-NATs positively regulate their partner genes by forming an RNA duplex, which stabilizes its partner gene to mask RNase and miRNA degradation (e.g., *BACE-ASI*¹⁵ and *Sirt1 AS*¹⁶). At the level of translation, cis-NATs negatively regulate their partner genes by masking ribosomal pairing (e.g., *MAPT-ASI* and *MIR_NATs*¹⁷). In contrast, cis-NATs positively regulate their partner genes by forming duplex RNAs that recruit the partner transcript to heavier polysomes (e.g., *AS Uchl1*¹⁸ and *SINEUPs*^{18,19}).

Cis-NATs can be divided into four types according to their structural characteristics. In the “embedded type”, the entire transcription unit of the antisense gene is embedded in the transcription unit of the sense gene. In contrast, in the “fully-overlapped type”, the transcription unit of the antisense gene covers the entire sense gene. In the “head-to-head type”, sense and antisense genes partially overlap only at their 5’ ends, while in the “tail-to-tail type”, the partial overlap is only at their 3’ ends. To investigate the function of cis-NATs, it is important

¹Department of Molecular Biology, Hamamatsu University School of Medicine, Hamamatsu, Shizuoka 431-3192, Japan. ²Advanced Research Facilities & Services, Preeminent Medical Photonics Education & Research Center, Hamamatsu University School of Medicine, Hamamatsu, Shizuoka 431-3192, Japan. ³These authors contributed equally: Tatsuya Ohhata and Maya Suzuki. ✉email: ohhata@hama-med.ac.jp

to determine their structural characteristics, including the distance between the promoters of the sense and antisense genes and whether the antisense transcription unit contains a regulatory element of the sense gene, such as its promoter, enhancer, miRNA targeting sequence, or ribosome binding sites.

Therefore, to elucidate the function of cis-NATs, it is critical to investigate both structural characteristics and expression profiles of cis-NAT pairs simultaneously. RNA-seq has become a common technique to investigate genome-wide gene expression, and genome-wide sequencing data is accumulating in databases, such as those curated by ENCODE²⁰ and FANTOM²¹. To date, whole genomes and transcriptomes from more than 2,000 species, including subspecies and strains, are deposited in Ensembl²² and NCBI. The transcriptome data include the locational information of each gene, including chromosome location, strand direction, transcription start sites (TSS), and transcription termination sites (TTS). Using the locational information of each gene in transcriptome data, it is theoretically possible to simultaneously investigate expression profiles and structural characteristics of cis-NATs. Identification of comprehensive cis-NATs with their original pipelines has been reported from multiple species including Arabidopsis²³, rice²⁴, maize²⁵, and sugarcane²⁶, as well as three kinds of mammals such as human, mouse, and rat²⁷, and 10 different species²⁸; however, the source codes for the computational program are not available for researchers. In contrast, the source code of some of the bioinformatics tools is available: NASTI-seq²⁹, written in R, allows reliable detection of cis-NATs using variable error rate of the strand-specific protocol; NATpipe³⁰, written in Perl, allows systematical discovery of NATs from de novo assembled transcriptomes; BEDTools³¹, written in C++, allows identification of overlapped cis-NAT pairs based on genome annotation. While these tools offer a reliable method for such analysis, they are not amenable to identifying the cis-NATs with their expression profiles and structural characteristics including “embedded”, “fully-overlapped”, “head-to-head”, and “tail-to-tail”.

Here, we developed a simple and convenient program termed CCIVR (comprehensive cis-NATs identifier via RNA-seq data) that enables the identification of total cis-NAT information with their structural characteristics based on its locational information with or without expression profiling obtained from processed RNA-seq data. CCIVR provides a novel tool to simultaneously investigate the function and structural characteristics of cis-NATs in numerous processes in various species.

Results

Overview of CCIVR and its principles of operation. To simultaneously investigate genome-wide structural characteristics and expression profiling of cis-NAT pairs, we developed CCIVR. Four types of cis-NAT, embedded (EB), fully-overlapped (FO), head-to-head (HH), and tail-to-tail (TT) are defined in CCIVR according to the criteria shown in Fig. 1A. Previous studies did not separate EB and FO cis-NATs because they defined “paired gene sets” as cis-NATs^{3,4}. Here, we defined “antisense transcripts” as cis-NATs and this is why we separated the types as described above. Furthermore, in this study, the criteria of identified cis-NATs were based on their structural characteristics only, and were not related to their RNA type, such as protein coding, long non-coding RNA (lncRNA), miRNA, and pseudogene.

The CCIVR process runs in a step-by-step manner (Fig. 1B). The input file contains every gene’s locational information, including chromosome location, strand direction, TSS, and TTS, obtained from the Ensembl database, as well as expression profiles, such as FPKM (fragments per kilobase of exon per million mapped reads) or TPM (transcripts per kilobase million), obtained from processed RNA-seq data using a peer-reviewed tool such as RSEM³² (for details, please see README.md file placed at <https://github.com/CCIVR/ccivr>). The input file is first divided into two groups of data sets according to whether the genes are on the plus (+) or minus (-) strand. Subsequently, the four cis-NAT types, EB, FO, HH, and TT, are sequentially extracted from plus-to-minus and from minus-to-plus strands to generate transient data that list each type of cis-NAT. Finally, all of the cis-NATs are combined to generate an output file that contains all cis-NAT data.

As an example from eight CCIVR processes, extraction of FO cis-NATs from minus-to-plus strand is shown in Fig. 1C. A mouse dataset that contains 55,146 genes (Ensembl GRCm39) was subjected to the process. The *Xist* gene was chosen as an example target gene on the minus strand, and the *Tsix* gene was chosen as an example of an identified FO cis-NAT on the plus strand. Since the *Xist* gene is on the X-chromosome, only the genes on the minus strand of the same chromosome were selected (selection 1: chromosome, Fig. 1C). The selected genes on the minus strand (1,352 genes) were subjected to the next screening that matched the criteria for FO screening of the *Xist* gene [condition: (AS-TSS \leq 102,503,972) & (102,526,860 \leq AS-TTS)] (selection 2: location, Fig. 1C). *Xist* and *Tsix* information was combined as paired data with their structural relationship information as “FO”, and all of the identified cis-NAT pairs were integrated as an FO cis-NAT list. With mouse datasets, a total of 317.4 million gene-to-gene comparisons were performed to accomplish the CCIVR analysis.

A comprehensive study of cis-NATs in model organisms. Subsequently, we attempted total cis-NAT identification from multiple model organisms with CCIVR. Ensembl (release 105), Ensembl plant (release 52), and Ensembl fungi (release 52), contain data for 311, 119, and 1,506 species (including sub-species and strains), respectively. From these, we chose 11 genetically well-studied representative model organisms for CCIVR analysis (Fig. 2, Supplementary Dataset File 1). We found that the percentage of cis-NAT-containing genes tended to increase from lower to higher organism complexity among each of fungi (*S. cerevisiae*, *N. crassa*, and *S. pombe*), invertebrate (*C. elegans* and *D. Melanogaster*), and vertebrate species (*D. rerio*, *X. tropicalis*, *G. gallus*, *M. musculus*, and *H. sapiens*) (Fig. 2: please note that this was not the case for *D. rerio*). Interestingly, there tended to be an inverse correlation with the percentage of protein-coding genes (Fig. S1), indicating that the existence of non-coding RNA is a reason to increase the percentage of cis-NATs. However, CCIVR analysis with only protein-coding genes also showed this tendency (Fig. S2, Supplementary Dataset File 2), indicating that lncRNA is not the only reason for the positive correlation between the percentage of cis-NATs and evolutionary com-

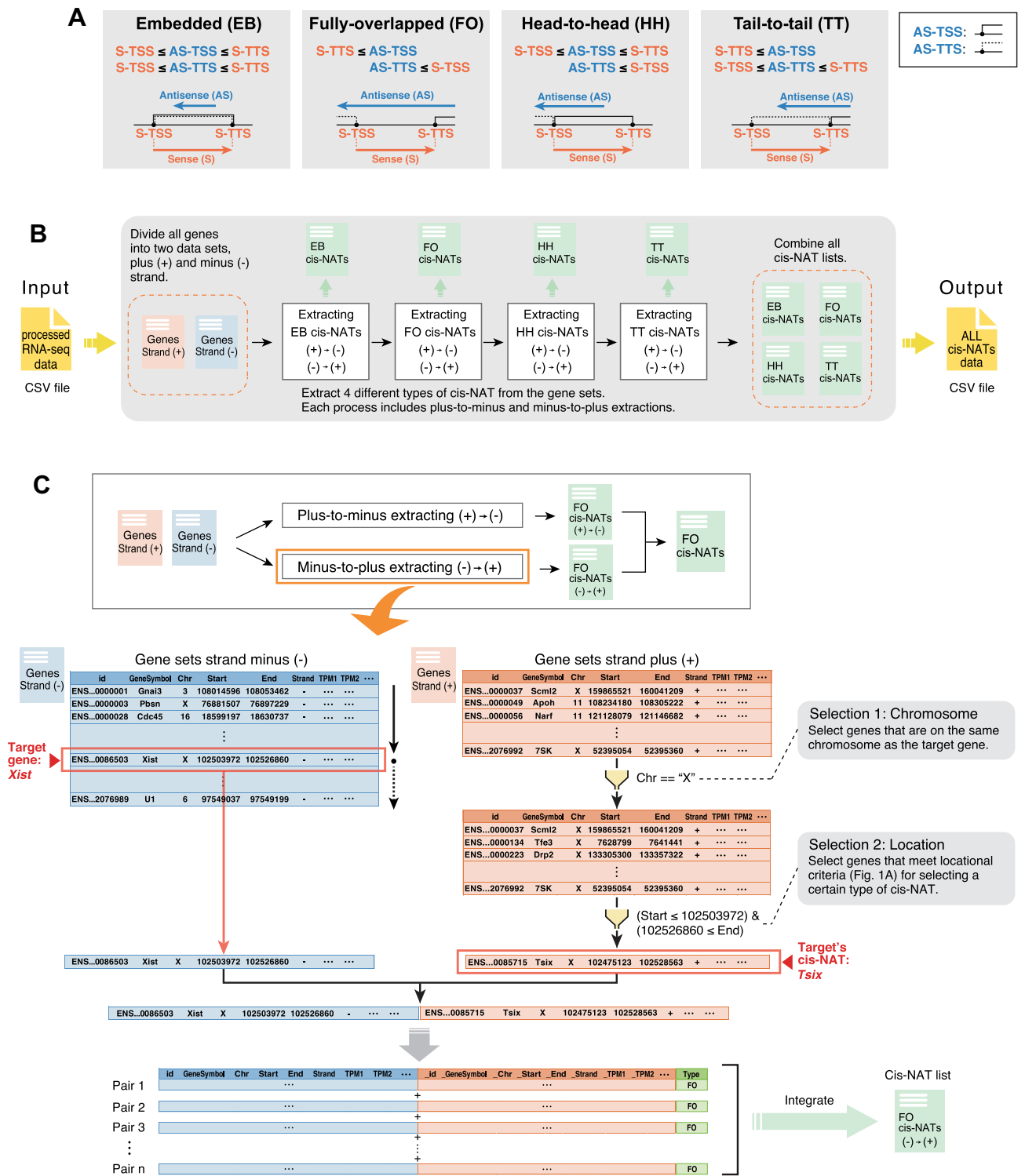


Figure 1. Overview of CCIVR and the processing steps of its pipeline. (A) Definition of four types of cis-NAT analyzed in this study, “embedded”, “fully-overlapped”, “head-to-head”, and “tail-to-tail”, which can be identified by CCIVR. TSS: transcription start site; TTS: transcription termination site. (B) Overview of CCIVR. The program generates a list of all cis-NAT types from a processed RNA-seq data. The processed RNA-seq data must contain five different gene annotation columns including “id”, “Chr”, “Strand”, “Start”, and “End”. Attaching information for expression profiling obtained from RNA-seq analysis is also available such as “TPM”, “FPKM”, “fold-change”, and “padj”. For details, please see README.md file placed at <https://github.com/CCIVR/ccivr>. (C) An example of the process for a target gene, *Xist*, and identification of its cis-NAT, *Tsix*, during the process that identifies fully-overlapped cis-NATs from minus to plus strand.

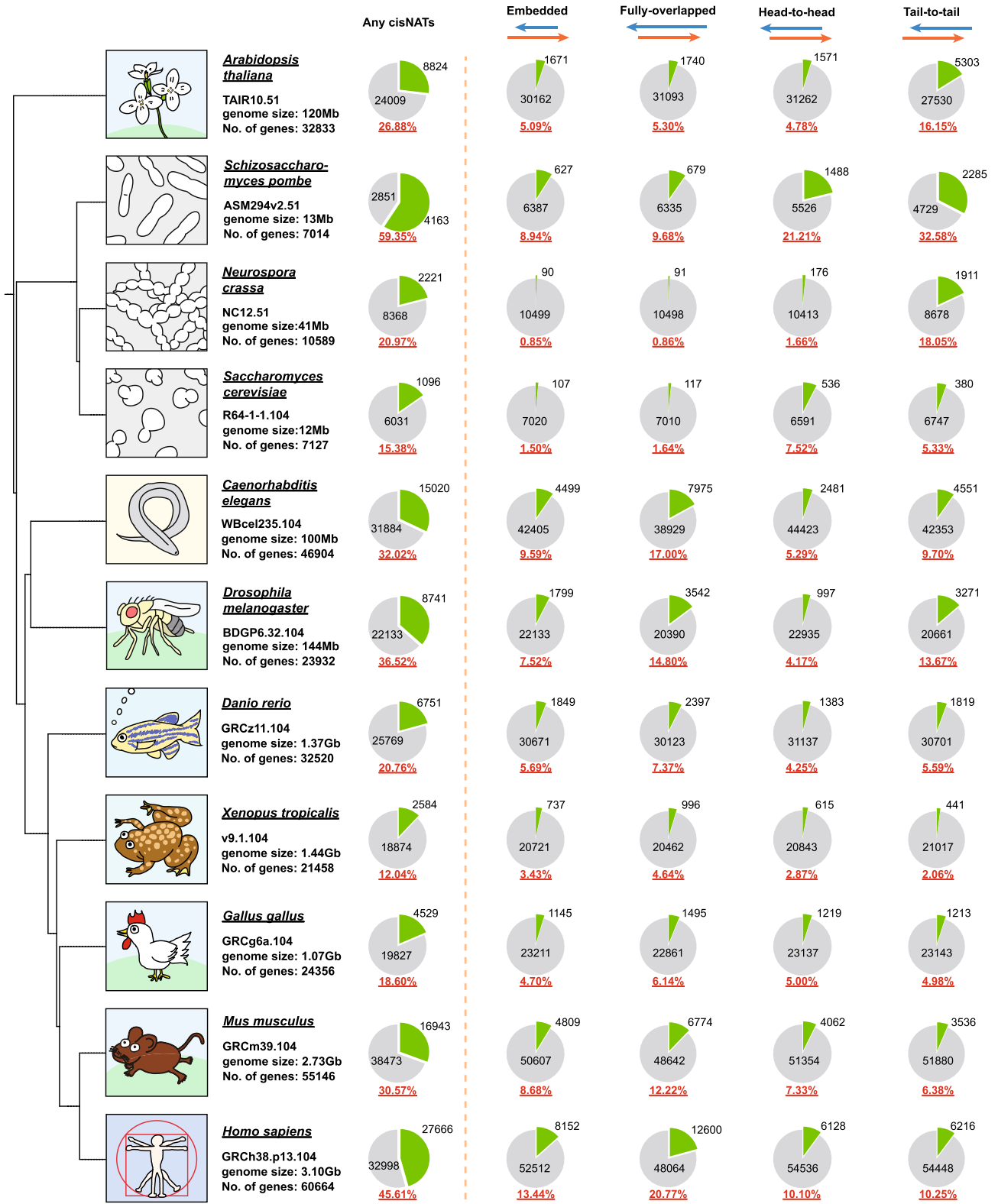


Figure 2. CCIVR enables identification of total cis-NATs: demonstration with representative model organisms. The percentage and number of genes that possess each type of cis-NAT (any cis-NAT, embedded, fully-overlapped, head-to-head, or tail-to-tail) from eleven different model organisms are shown. Red and blue arrows represent sense transcripts and their antisense transcripts, respectively. The genome data version from Ensembl/Ensembl_plant/Ensembl_fungi is shown below the name of each species. The phylogenetic tree was generated by phyloT_v2 (<https://phyloT.biobyte.de>).

plexity. Although the completeness of these databases might vary among species, which reflects the number of cis-NATs identified, these results indicate that the percentage of cis-NATs and evolutionary complexity might be somehow correlated. Intriguingly, the positive correlation between the percentage of cis-NATs and evolutionary complexity could be also confirmed by the data from a previous study that attempted to identify total cis-NATs from different species²⁸; among invertebrates, the percentage of cis-NATs was 6.8% and 22.8% in worm and fly respectively, and among vertebrates, the percentage of cis-NATs was 5.2%, 6.7%, 9.7%, 28.6%, and 36.2% in zebrafish, frog, chicken, mouse, and human, respectively. The percentage of the cis-NATs tended to increase in all species in our study compared to the previous one. It may reflect the improvement of gene annotation and genome information of the moment. In summary, CCIVR facilitates the comprehensive identification of all types of cis-NAT pairs from numerous different species.

Identification of cis-NATs that demonstrate parental-biased expression in embryonic stem cells.

A well-known process that cis-NATs are involved in is genomic imprinting, which is an epigenetic phenomenon whereby identical alleles of genes are expressed in a parent-of-origin-dependent manner³³. We attempted to identify cis-NAT pairs that demonstrate parentally biased expression to evaluate whether the CCIVR program can identify known and/or novel imprinted cis-NAT pairs. To this end, we used published RNA-seq data from human parthenogenetic embryonic stem cells (pESCs³⁴) and androgenetic ESCs (aESCs³⁵), which possess only maternally or paternally inherited gene sets, respectively (Fig. 3A). (Please note that strand-specific RNA-seq data is preferable for CCIVR analysis because the origin of sequence reads from an overlap region becomes apparent. Nevertheless, non-strand-specific RNA-seq samples are applicable; indeed, this pESC and aESC RNA-seq data is non-strand-specific.) The RNA-seq samples were verified by principal component analysis (PCA) (Fig. 3B; each pESC or aESC sample was spatially gathered) and volcano plot analysis (Fig. 3C; well-known maternally and paternally imprinted genes could be identified). We defined differentially expressed genes (DEGs) as those whose difference in expression between pESCs and aESCs was statistically significant ($\text{padj} < 0.05$) and that the difference was more than twofold; 1,661 and 1,311 DEGs were identified in pESCs and aESCs, respectively (Fig. 3C). The processed RNA-seq data were then subjected to CCIVR analysis (Supplementary Dataset File 3), and the numbers of each type of cis-NAT pair for genes that were differentially expressed in pESCs (maternal expression or “mat”) or aESCs (paternal expression or “pat”) were counted (Fig. 3D). Then, we counted the number of all types of cis-NAT pairs that show positive or negative correlation, and 164 and 29 cis-NAT pairs were found, respectively. (please note that some of the cis-NAT pairs were redundantly represented in Fig. 3D. For example, the cis-NAT pairs showing embedded type with mat-pat expression were the same as the cis-NAT pairs showing fully-overlapped type with pat-mat expression) For all types of cis-NAT pair, the correlations with overlapping genes tended to be positive (mat-mat or pat-pat) rather than negative (mat-pat or pat-mat), consistent with previous studies^{27,36}. Although positive correlation is interesting because it might indicate positive regulation of gene expression in the imprinted gene clusters, we rather focused on the negative correlation because some cis-NATs have been reported to act as negative regulators of partner genes in the imprinted gene clusters^{6,10}. We analyzed all 29 cis-NAT pairs that showed negative correlation from all types of cis-NATs by heatmap analysis (Fig. 3E). Notably, among the embedded type of cis-NAT pairs, we identified the well-known functional cis-NAT, *KCNQ1OT1*, and its partner gene, *KCNQ1*, which tended to be expressed from paternal and maternal alleles, respectively (Fig. 3E), consistent with their previously reported expression patterns³⁷. Taken together, we conclude that CCIVR is an effective program for identifying functional cis-NAT pair candidates from numerous deposited RNA-seq datasets.

Identification of cis-NATs upon TGF β stimulation.

To further evaluate CCIVR-mediated identification of cis-NATs involved in a biological process, we chose to examine the TGF β signaling pathway, which induces epithelial-mesenchymal transition (EMT) and apoptosis³⁸. To this end, we performed RNA-seq of the human hepatocellular carcinoma cell line, Huh-7, with or without TGF β stimulation for 12 h and 48 h (Fig. 4A and S3A), with EMT confirmed by morphological examination (Fig. S3B and C). These RNA-seq samples were prepared as “strand-specific” to improve the accuracy of mapping at the overlap region. Reproducibility between duplicated samples was confirmed by PCA analysis (Fig. S3D) and DEGs were identified by volcano plot analysis (Fig. S3E and F). We defined DEGs as those whose difference in expression between sample groups was statistically significant ($\text{padj} < 0.05$) and that the difference was more than 1.5-fold for up-regulated genes and less than 0.67-fold for down-regulated genes. Then, following GO analysis (Fig. S3G) and confirmation of epithelial/mesenchymal marker gene expression (Fig. S3H), which indicates proper TGF β responses, the processed RNA-seq data were subjected to CCIVR analysis (Supplementary Dataset File 4) and the numbers of each type of DEG constituting cis-NAT pairs were counted (Fig. 4B). These genes were subjected to GO analysis, and we found that TGF β signaling-related genes, including genes involved in EMT, were enriched in up-regulated genes and that cell growth-related genes were enriched in down-regulated genes (Fig. 4C). This indicated that cis-NAT genes that are differentially expressed by TGF β stimulation are involved in TGF β -related biological processes.

We subsequently attempted to discover novel cis-NATs that regulate the expression of their partner genes. The murine *Tsix* gene is a cis-NAT that coordinates the initiation of X chromosome inactivation by negatively regulating its partner gene, *Xist*^{39–41}. We have studied the mechanism of *Tsix* action and found that a histone modification, H3K36me3, accompanied by *Tsix* transcription is required for *Xist* repression^{8,9}. We therefore focused on the *Tsix*-like regulation system from among the multiple kinds of regulation system that cis-NATs possess. We selected cis-NAT pairs whose expression was negatively correlated as down-up [the expression of sense transcript is down-regulated whereas its antisense transcript (cis-NAT) is up-regulated upon TGF β stimulation], and they were subjected to heatmap analysis (Fig. 4D; please note that the tail-to-tail group was omitted owing to limited space). Given that *Tsix* transcription running through the *Xist* promoter is required for its function⁴²,

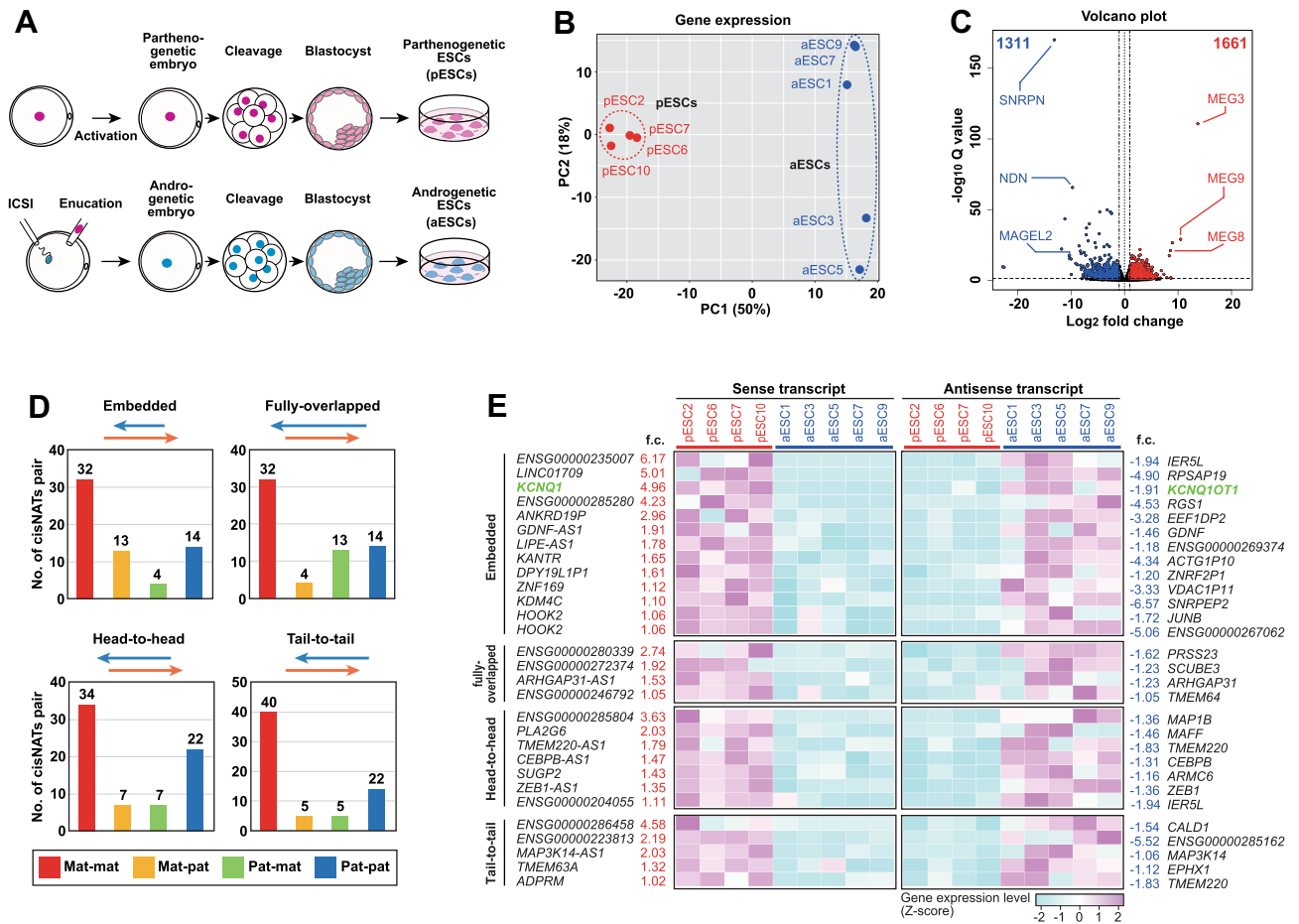


Figure 3. Example of CCIVR analysis I: identification of cis-NATs that demonstrate parentally-biased expression in ESCs. (A) Schematic representation of the procedure for establishing parthenogenetic ESCs (pESCs³⁴) and androgenetic ESCs (aESCs³⁵). (B) Principle component analysis of gene expression levels in pESCs and aESCs. The percentage of explained variation is indicated in parentheses. (C) Volcano plot of RNA-seq data displaying the gene expression values for pESCs compared with aESCs. Highly differentially expressed genes (DEGs) with statistical significance ($\text{padj} < 0.05$) in pESCs (fold change ≥ 2) and aESCs (fold change ≤ 0.5) are highlighted in red and blue, respectively. The total number of DEGs is shown. Selected known imprinted genes³⁵ are also indicated. (D) CCIVR analysis with DEGs. The number of four kinds of cis-NAT pairs (embedded, fully-overlapped, head-to-head, tail-to-tail) with different expression patterns: maternally-highly expressed gene (mat) paired with mat (Mat-mat), mat paired with paternally-highly expressed gene (pat) (Mat-pat), pat paired with mat (Pat-mat), and pat paired with pat (Pat-pat). The numbers of each cis-NAT pair are also indicated. (E) Heatmap analysis of four kinds of cis-NAT pairs with inverse-correlated gene expression (Mat-pat). *KCNQ1/KCNQ1OT1*, highlighted in green, is a known cis-NAT pair showing imprinted expression in an inverse-correlated fashion. f.c.: fold change (\log_2).

we focused on only fully-overlapped and head-to-head cis-NATs because their transcription runs through the promoters of their partner genes. These genes were subjected to further analysis to investigate whether the down-regulation observed in the sense genes was dependent on SETD2 histone methyltransferase, which catalyzes H3K36me3 modification (Fig. S4: please note that *BCAR3-AS1* and *AC046134.2* replaced *AL109613.1* and *AC097103.2*, respectively, after revisiting the latest version of Ensembl, Human GRCh38.p13). The efficiency of SETD2 knockdown was confirmed by decreased levels of its mRNA (Fig. 4E) and protein (Fig. 4F), and reduced catalysis of H3K36me3 modification (Fig. 4G). Based on the expression of 32 cis-NAT pairs, we chose nine cis-NAT pairs (*BCAR3-AS1/BCAR3*, *RBP2/AC097103.2*, *THNSL1/ENKUR*, *SERBP1P5/FRAS1*, *ADH6/AP002026.1*, *GPX2/CHURC1*, *EGFR-AS1/EGFR*, *CFAP97/SNX25*, and *UROD/HECTD3*), and investigated their dynamic expression upon TGF β stimulation and then after SETD2 depletion. We also assessed alteration of H3K36me3 accumulation in the promoters of sense genes following SETD2 depletion. We identified two cis-NAT pairs, *BCAR3-AS1/BCAR3* and *GPX2/CHURC1*, that demonstrated statistically significant alterations in these experiments (Fig. 4H–L). Interestingly, both of these cis-NAT types were “fully-overlapped”, which is the same as *Tsix*. Furthermore, their structures resembled *Tsix/Xist*⁴⁰ in that the TSS of its sense transcript is at the 3' end of its cis-NAT (Fig. 4H). Dynamics analysis revealed that antisense and sense transcription was symmetrically altered upon TGF β stimulation (Fig. 4I). While expression of the sense genes was significantly decreased (Fig. 4J) and that of the cis-NATs was significantly increased (Fig. 4K) upon TGF β stimulation, H3K36me3 modification

was significantly increased at the promoter regions of sense genes (Fig. 4L). When SETD2 was depleted, the accumulated H3K36me3 was significantly reduced (Fig. 4L) accompanied by derepression of the sense genes (Fig. 4J). Importantly, the derepression was not because of the decrease in their cis-NAT expression (Fig. 4K). Taken together, these results indicate that the transcription of cis-NATs, *BCAR3* and *CHURC1*, negatively regulate their partner genes, *BCAR3-AS1* and *GPX2*, by promoting H3K36me3 modification within the promoters of each partner gene.

BCAR3 is involved in anti-estrogen resistance in breast cancer cells⁴³. Although stable overexpression of *BCAR3* does not lead to a typical EMT phenotype, it results in down-regulation of cadherin-mediated adhesion and augmentation of fibronectin expression⁴⁴, suggesting that it positively regulates part of the EMT phenotype. In contrast, the function of *BCAR3-AS1* is obscure. It would be interesting to elucidate whether *BCAR3* promotion of the EMT phenotype is through repression of *BCAR3-AS1*. *CHURC1* is a zinc finger transcriptional activator⁴⁵. Its cis-NAT, *GPX2*, encodes a glutathione peroxidase (GPX) that possesses glutathione-dependent hydrogen peroxidase reducing activity⁴⁶. *GPX2* is known as a negative regulator of apoptosis⁴⁷; therefore, investigation of its involvement in the progression of apoptosis by TGF β stimulation is warranted. In summary, we used CCIVR to identify novel cis-NATs that regulate the expression of their partner genes. CCIVR analysis can therefore be used to screen and identify cis-NATs that possess specific mechanisms of action among multiple kinds of gene regulation.

Discussion

In this study, we demonstrated that CCIVR can contribute to the identification of cis-NATs involved in the regulation of transcription. In contrast to transcription, some cis-NATs are involved in regulation at the level of translation^{17–19}. Although CCIVR uses transcriptome data, such as from RNA-seq, it can also be applied to proteome data such as from quantitative mass spectrometry⁴⁸. Therefore, CCIVR enables functional studies of cis-NATs in both transcriptional and translational regulation. Some antisense RNAs are transcribed from sequence that is upstream of the promoter of its partner gene (in a strict sense, these genes are not cis-NAT pairs because they do not overlap), and some antisense transcripts may have a role in regulating expression of their partner genes through modulating the action of their enhancers. The CCIVR program is open-source and can be readily customized for specific purposes; therefore, identifying such antisense RNAs is also practicable.

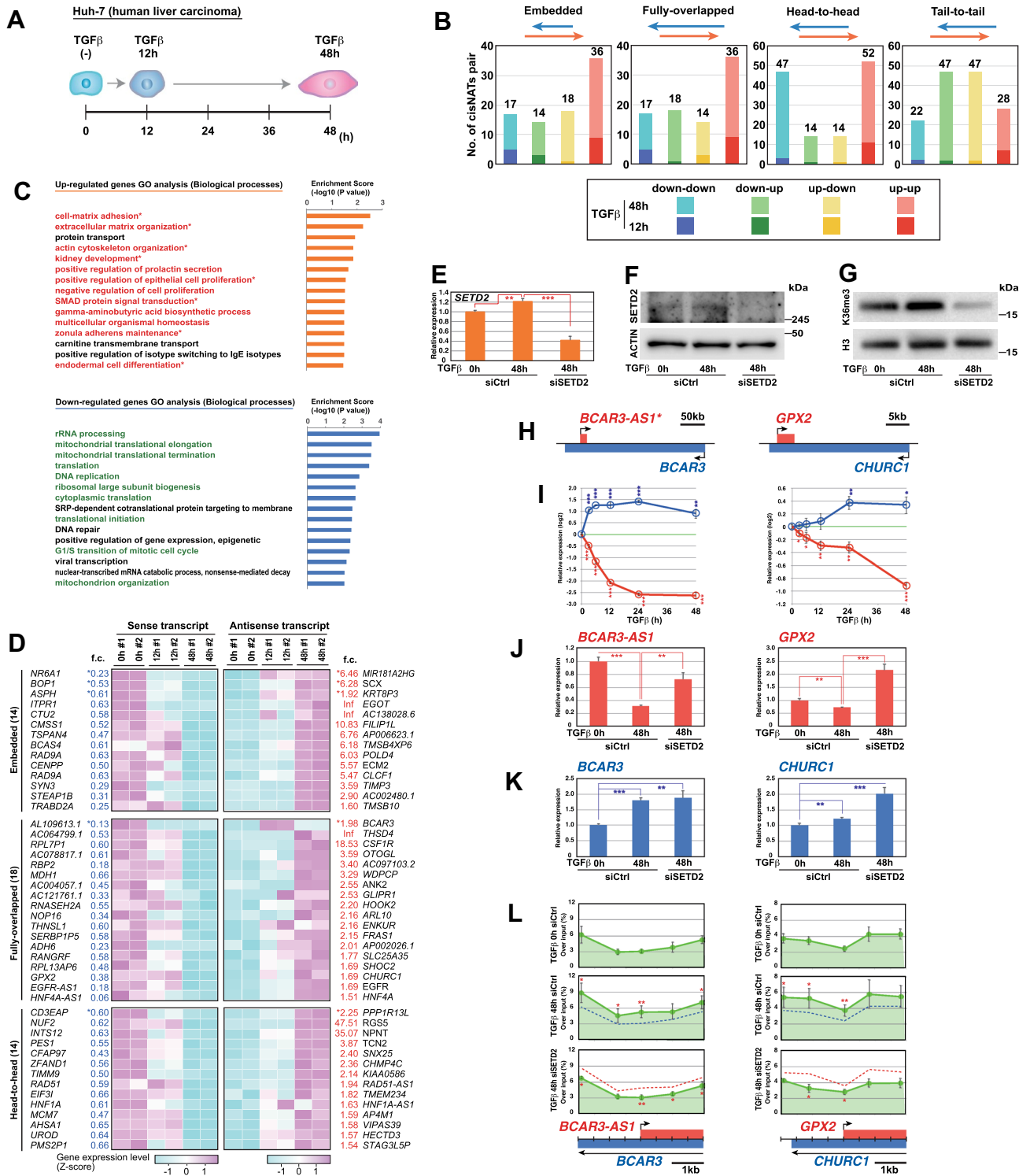
Many cis-NATs involved in human diseases have been reported^{49,50}, and some of them are therapeutic targets. Therefore, CCIVR can contribute biomedically by identifying novel cis-NATs involved in human diseases. Compared to previous studies attempting comprehensive identification of cis-NATs using *Arabidopsis* genome data²³, human microarray data²⁷, and EST data from 10 different species²⁸, our study has two advances: firstly, we updated the results by utilizing the latest genome datasets and, secondly, CCIVR is a simple, convenient, and open-source program that allows investigation of all RNA-seq and genome datasets from more than 2,000 species deposited in the NCBI and Ensembl databases. For predicting the composition of various cis-NATs by CCIVR, it depends on the accuracy of gene annotation including their structure and strand direction deposited in the databases. Furthermore, for performing the expression profile analysis of cis-NATs by CCIVR, it uses processed RNA-seq data using third-party programs such as STAR⁵¹ for mapping, RSEM³² for expression profiling, and DESeq2⁵² for statistical analysis; i.e., CCIVR scripts do not cover the full pipeline of CCIVR analysis. They are the limitations of CCIVR analysis at the moment and further improvements are required in the future.

Here, we introduced an original program termed CCIVR that simultaneously analyzes the structure of cis-NAT pairs and their expression profiles. We believe that CCIVR will drive the study of cis-NATs to elucidate their mechanisms of action and functions in numerous processes in various species.

Materials and methods

CCIVR analysis of model organisms. All of the gtf files from 11 species were downloaded from Ensembl plant (<https://plants.ensembl.org/index.html>; *Arabidopsis thaliana*: TAIR10, release 51), Ensembl Fungi (<https://fungi.ensembl.org/index.html>; *Schizosaccharomyces pombe*: ASM294v2, release 51, *Neurospora crassa*: NC12, release 51, *Saccharomyces cerevisiae*: R64-1-1, release 104), and Ensembl (<https://www.ensembl.org/index.html>; *Caenorhabditis elegans*: WBcel235, release 104, *Drosophila melanogaster*: BDGP6.32, release 104, *Danio rerio*: GRCz11, release 104, *Xenopus tropicalis*: v9.1, release 104, *Gallus gallus*: GRCg6a, release 104, *Mus musculus*: GRCh38, release 104, *Homo sapiens*: GRCh38.p13, release 104). From the gtf files, only the gene information listed in the feature column was extracted and was converted to a csv file using Python (ver 3.8.8) and one of its modules, gtfparse (ver 1.2.1). To avoid the duplication of the genes to be analyzed, only ena and PomBase were used from *N. crassa* and *S. pombe*, respectively, as their gene_source. Concerning other species, every gene_source was used for the CCIVR analysis because no duplication was observed. The number of genes was counted from gene_id but not GeneSymbol because we found that gene_id was unique to every gene while this was not the case for a few genes in GeneSymbol. The phylogenetic tree was generated by phyloT-v2⁵³ (<https://phyloT.biobyte.de>).

RNA-seq analysis of pESCs and aESCs. The SRA files used for pESC and aESC analysis were downloaded from NCBI (<https://www.ncbi.nlm.nih.gov>) and are listed in Supplemental Table S1. Low-quality RNA-seq reads and adaptor sequences were removed using Trim Galore! (version 0.6.7) with the default condition. Sequence reads were aligned to the human reference genome (GRCh38/hg38) using STAR⁵¹ (version 2.7.9a) by the default condition with an option that allows up to three mismatches (-outFilterMismatchNmax 3), as previously described³⁵. For each gene, transcripts per kilobase million (TPM) was calculated by RSEM³² (version 1.3.3) using the “rsem-calculate-expression” command with the default condition. Differential expression



◀ **Figure 4.** Example of CCIVR analysis II: identification of cis-NATs with TGFβ stimulation. **(A)** Schematic representation for TGFβ stimulation of Huh-7 cells. For details, see also Figure S3. **(B)** CCIVR analysis with genes that demonstrate differential expression with TGFβ stimulation [upgenes: fold change ≥ 1.5, padj < 0.05; downgenes: fold change ≤ 0.67, padj < 0.05]. The numbers of four kinds of cis-NAT pairs with the different combinations of expression pattern including down-down, down-up, up-down, and up-up (sense-antisense) are shown. **(C)** GO analysis of biological processes with genes from differentially expressed cis-NAT pairs. The top 15 biological processes are shown. Processes related to TGFβ signaling in up-regulated genes are in red and, among them, those related to EMT are indicated with an asterisk (*). Processes related to cell growth in down-regulated genes are in green. **(D)** Heatmap analysis of cis-NAT pairs with inverse-correlated gene expression (down-up). Three kinds of cis-NATs, embedded, fully-overlapped, and head-to-head are shown. Fold change (f.c.) with an asterisk (*) indicates significant differential expression after TGFβ stimulation for 12 h and the other values are for 48 h after TGFβ stimulation. **(E–G)** Confirmation of SETD2 expression by RT-qPCR **(E)** and western blotting **(F)** and the level of H3K36me3 by western blotting **(G)** in SETD2 knockdown samples. Images for western blotting were cropped for improving the clarity and conciseness. Original blots are presented in Supplementary Figure S5. **(H)** Maps of cis-NAT pairs, *BCAR3-AS1/BCAR3* and *GPX2/CHURC1*. *: *BCAR3-AS1* replaced *ALI09613.1* by revisiting the latest version of Ensembl, Human GRCh38p13. **(I)** Expression dynamics of *BCAR3-AS1/BCAR3* and *GPX2/CHURC1* cis-NAT pairs upon TGFβ stimulation for 48 h. **(J, K)** RT-qPCR analysis for sense transcripts **(J)** and their cis-NATs **(K)** upon TGFβ stimulation with or without SETD2 knockdown. **(L)** ChIP-qPCR analysis of H3K36me3 accumulation at the promoter of sense transcripts. For comparison, blue and red dashed lines representing the level of H3K36me3 accumulation in “TGFβ 0 h siCtrl” and “TGFβ 48 h siCtrl” are shown. **(E, I–L)** Means ± SD across biological replicates are shown (**E, I–K**: n = 3; **L**: n = 5). *p < 0.05, **p < 0.01, ***p < 0.001, comparisons as indicated (**E, I–K**) and by dashed lines (**L**); Student’s t-test. Normalized to *GAPDH* (**E, J–K**).

analysis (Wald test), PCA plot analysis, and volcano plot analysis were performed using DESeq2⁵² with the default condition (Bioconductor version: Release 3.13). Heatmaps were generated using the R “*gplots*” function.

RNA-seq analysis of Huh-7 cells. Total RNA was purified using an RNeasy Mini kit (Qiagen, Hilden, Germany). RNA quality was measured using NanoDrop spectrophotometry (Thermo Fisher Scientific, Waltham, MA, USA) and its quantity was measured using the TapeStation Automated Electrophoresis System (Agilent Technologies, Santa Clara, CA, USA). All RNA-seq procedures, including library construction, purification, library quality control and quantification, sequencing cluster generation, high-throughput sequencing, and result generation, which included PCA and volcano plotting, were performed by Genewiz Biotechnology Co. Ltd (<https://www.genewiz.com>). Gene expression levels were measured by reading density and FPKM (fragments per kilobases per million reads) was calculated based on the read counts from HT-seq (V 0.6.1).

GO analysis. Gene ontology (GO) term enrichment analyses (biological processes) were performed using the bioinformatics tool, DAVID (ver 6.8)^{54–56}.

Cell culture and reagents. Human hepatoma cell line Huh-7 (JCRB0403) was purchased from JCRB Cell Bank (National Institute of Biomedical Innovation, Osaka, Japan), grown in Dulbecco’s modified Eagle’s medium (DMEM) (Sigma-Aldrich, St. Louis, MO, USA) supplemented with 10% fetal bovine serum (FBS) (Sigma-Aldrich) and 1 × penicillin/streptomycin (Meiji Seika Pharma Co., Ltd., Tokyo, Japan) at 37 °C under an atmosphere containing 5% CO₂. Recombinant hTGFβ1 (240-B; R&D systems, Minneapolis, MN, USA) was added to a final concentration of 10 ng/ml for TGFβ stimulation. A BZ-8000 phase-contrast microscope (Keyence, Osaka, Japan) was used to monitor morphological changes upon TGFβ stimulation.

RNA interference. Huh-7 cells were transfected with *SETD2* siRNA (si*SETD2*) or control siRNA (siCtrl) using Lipofectamine RNAiMAX (Invitrogen, Waltham, MA, USA), in accordance with the manufacturer’s protocol. At 48 h post-transfection, the cells were again transfected with *SETD2* siRNA or control siRNA as per the first transfection. At 24 h after the second transfection, the cells were subjected to with or without TGFβ stimulation. *SETD2* siRNA and negative control siRNA (non-targeting pools) were purchased (si*SETD2*: L-012448-00-0005, siCtrl: D-001810-10-05; Horizon discovery, Cambridge, UK). The *SETD2* siRNA consisted of four different oligonucleotides with the following target sequences: 5′-UAA AGG AGG UAU AUC GAA U-3′ (J-012448-05); 5′-GAG AGG UAC UCG AUC AUA A-3′ (J-012448-06); 5′-GCU CAG AGU UAA CGU UUG A-3′ (J-012448-07); and 5′-CCA AAG AUU CAG ACA UAU A-3′ (J-012448-08). The nucleotide sequence of the control siRNA consisted of four different oligonucleotides with the following non-targeting sequences: 5′-UGG UUU ACA UGU CGA CUA A-3′; 5′-UGG UUU ACA UGU UGU GUG A-3′; 5′-UGG UUU ACA UGU UUU CUG A-3′; and 5′-UGG UUU ACA UGU UUU CCU A-3′.

RT-qPCR. Total RNA was purified using an RNeasy Mini Kit (Qiagen). For RT-qPCR, cDNA was prepared using SuperScript II reverse transcriptase (Invitrogen) with random primers (Invitrogen). RT-qPCRs were performed in duplicate using Thunderbird SYBR qPCR mix (Toyobo, Osaka, Japan) with the primers listed in Supplemental Table S2 on a StepOnePlus Real-Time PCR system (Life Technologies, Carlsbad, CA, USA). The standard curve method was used for quantification and expression levels were normalized against *GADPH*.

Western blot analysis. For SETD2, western blot analysis was performed as previously described⁵⁷ with minor modification. In brief, cells were lysed with RIPA buffer in the presence of a protease inhibitor cocktail (cOmplete™; Roche, Basel, Switzerland). Lysed cells were rotated at 4 °C for 20 min and sonicated using a UCS-250 Bioruptor (Cosmobio, Tokyo, Japan). The sonication conditions were as follows: high, on 30 s/off 30 s, eight cycles. After collection of the supernatant by centrifugation the protein concentration was measured by a DC protein assay (Bio-Rad, Hercules, CA, USA). After denaturation of the cell lysate by diluting to final 1 × using 4 × SDS sample buffer [255 mM Tris–HCl (pH 6.8), 12% SDS, 40% glycerol, 20% β-mercaptoethanol, and 0.01% bromophenol blue] and incubation at 95 °C for 8 min, the cell lysate was separated by SDS-PAGE electrophoresis and transferred onto a polyvinylidene difluoride (PVDF) membrane (Immobilon-P IPVH00010; Millipore, Burlington, MA, USA), followed by immunoblotting with a primary α-SETD2 antibody (#EB08118; Everest Biotech, Oxford, UK) and a secondary α-Goat IgG, HRP Conjugate antibody (V805A; Promega, Madison, WI, USA), or an α-β-actin mAb, HRP conjugated antibody (289-99361; FUJIFILM Wako Pure Chemical Corp., Osaka, Japan). The signals were visualized using Clarity™ Western ECL Substrate (Bio-Rad) and the ChemiDoc Touch imaging system (Bio-Rad). H3K36me3 was detected as previously described⁸. In brief, cells were lysed with Triton extraction buffer (0.5% Triton X-100 and 0.02% NaN₃ in PBS) in the presence of a protease inhibitor cocktail (cOmplete™; Roche) on ice for 10 min. After centrifugation, the pellet was washed once with Triton extraction buffer. The pellet was resuspended with 0.2 N HCl, and the histone protein was extracted by rotation at 4 °C overnight. After collection of the supernatant by centrifugation, the protein concentration was measured by a Bradford assay (Bio-Rad). After denaturation of the cell lysate by diluting to final 1 × using 4 × SDS sample buffer and incubation at 95 °C for 8 min, the cell lysate was separated by SDS-PAGE electrophoresis and transferred onto a 0.2 μm pore nitrocellulose membrane (Whatman PROTRAN; Merck, Darmstadt, Germany), followed by immunoblotting with a primary α-H3K36me3 antibody (ab9050; Abcam, Cambridge, UK) and a secondary α-Rabbit IgG (H + L), HRP Conjugate antibody (W401B, Promega) or a primary α-H3 antibody (#39,763; Active motif, Carlsbad, CA, USA) and a secondary α-Mouse IgG (H + L), HRP Conjugate antibody (W402B, Promega). The signals were visualized as described above.

ChIP-qPCR. ChIP was performed with a commercial kit (SimpleChIP Enzymatic Chromatin IP Kit; Cell Signaling Technology, Danvers, MA, USA) in accordance with the manufacturer's procedure. After de-crosslinking and proteinase K treatment, DNA was purified using phenol–chloroform extraction and ethanol precipitation with the co-precipitation reagent, Pellet Paint (Merck). For qPCR, see the RT-qPCR section. The primer sequences used in ChIP-qPCR assays are listed in Supplemental Table S2. The following antibody was used: α-H3K36me3 (CMA333; a gift from Dr. Naohito Nozaki, MAB Institute, Inc.).

Data availability

DNA sequencing data have been deposited in the DDBJ Sequence Read Archive (DRA) of the DNA Data Bank of Japan (DDBJ) with accession number DRA013542. CCIVR is available from github at <https://github.com/CCIVR/ccivr>.

Received: 14 June 2022; Accepted: 5 September 2022

Published online: 15 September 2022

References

- Lacatena, R. M. & Cesareni, G. Base pairing of RNA I with its complementary sequence in the primer precursor inhibits ColE1 replication. *Nature* **294**, 623–626 (1981).
- Wight, M. & Werner, A. The functions of natural antisense transcripts. *Essays Biochem.* **54**, 91–101 (2013).
- Khorkova, O., Myers, A. J., Hsiao, J. & Wahlestedt, C. Natural antisense transcripts. *Hum. Mol. Genet.* **23**, R54–R63 (2014).
- Rosikiewicz, W. & Makalowska, I. Biological functions of natural antisense transcripts. *Acta Biochim. Pol.* **63**, 665–673 (2016).
- Staleno, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).
- Latos, P. A. *et al.* Airn transcriptional overlap, but not its lncRNA products, induces imprinted *Igf2r* silencing. *Science* **338**, 1469–1472 (2012).
- Xue, Z. *et al.* Transcriptional interference by antisense RNA is required for circadian clock function. *Nature* **514**, 650–653 (2014).
- Ohhata, T. *et al.* Histone H3 lysine 36 trimethylation is established over the *Xist* promoter by antisense *Tsix* transcription and contributes to repressing *Xist* expression. *Mol. Cell Biol.* **35**, 3909–3920 (2015).
- Ohhata, T. *et al.* Dynamics of transcription-mediated conversion from euchromatin to facultative heterochromatin at the *Xist* promoter by *Tsix*. *Cell Rep.* **34**, 108912 (2021).
- Pandey, R. R. *et al.* Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol. Cell* **32**, 232–246 (2008).
- Kotake, Y. *et al.* Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. *Oncogene* **30**, 1956–1962 (2011).
- Yap, K. L. *et al.* Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell* **38**, 662–674 (2010).
- Blank-Giwojna, A., Postepska-Igielska, A. & Grummt, I. lncRNA KHPS1 activates a poised enhancer by triplex-dependent recruitment of epigenomic regulators. *Cell Rep.* **26**, 2904–2915.e4 (2019).
- Arab, K. *et al.* GADD45A binds R-loops and recruits TET1 to CpG island promoters. *Nat. Genet.* **51**, 217–223 (2019).
- Faghihi, M. A. *et al.* Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat. Med.* **14**, 723–730 (2008).
- Wang, G.-Q. *et al.* Sirt1 AS lncRNA interacts with its mRNA to inhibit muscle formation by attenuating function of miR-34a. *Sci. Rep.* **6**, 21865 (2016).
- Simone, R. *et al.* MIR-NATs repress MAPT translation and aid proteostasis in neurodegeneration. *Nature* **594**, 117–123 (2021).
- Carrieri, C. *et al.* Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* **491**, 454–457 (2012).

19. Zucchelli, S. *et al.* SINEUPs are modular antisense long non-coding RNAs that increase synthesis of target proteins in cells. *Front. Cell Neurosci.* **9**, 174 (2015).
20. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
21. Ramilowski, J. A. *et al.* Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Res.* **30**, 1060–1072 (2020).
22. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
23. Bouchard, J., Oliver, C. & Harrison, P. M. The distribution and evolution of Arabidopsis thaliana cis natural antisense transcripts. *BMC Genomics* **16**, 444 (2015).
24. Lu, T. *et al.* Strand-specific RNA-seq reveals widespread occurrence of novel cis-natural antisense transcripts in rice. *BMC Genomics* **13**, 721 (2012).
25. Xu, J. *et al.* Natural antisense transcripts are significantly involved in regulation of drought stress in maize. *Nucleic Acids Res.* **45**, 5126–5141 (2017).
26. Lembke, C. G., Nishiyama, M. Y., Sato, P. M., de Andrade, R. F. & Souza, G. M. Identification of sense and antisense transcripts regulated by drought in sugarcane. *Plant Mol. Biol.* **79**, 461–477 (2012).
27. Ling, M. H. T., Ban, Y., Wen, H., Wang, S. M. & Ge, S. X. Conserved expression of natural antisense transcripts in mammals. *BMC Genomics* **14**, 243 (2013).
28. Zhang, Y., Liu, X. S., Liu, Q.-R. & Wei, L. Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res.* **34**, 3465–3475 (2006).
29. Li, S., Liberman, L. M., Mukherjee, N., Benfey, P. N. & Ohler, U. Integrated detection of natural antisense transcripts using strand-specific RNA sequencing data. *Genome Res.* **23**, 1730–1739 (2013).
30. Yu, D., Meng, Y., Zuo, Z., Xue, J. & Wang, H. NATpipe: An integrative pipeline for systematical discovery of natural antisense transcripts (NATs) and phase-distributed nat-siRNAs from de novo assembled transcriptomes. *Sci. Rep.* **6**, 21666 (2016).
31. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
32. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 323 (2011).
33. Kobayashi, H. Canonical and non-canonical genomic imprinting in rodents. *Front. Cell Dev. Biol.* **9**, 713878 (2021).
34. Weissbein, U., Schachter, M., Egli, D. & Benvenisty, N. Analysis of chromosomal aberrations and recombination by allelic bias in RNA-Seq. *Nat. Commun.* **7**, 12144 (2016).
35. Sagi, I. *et al.* Distinct imprinting signatures and biased differentiation of human androgenetic and parthenogenetic embryonic stem cells. *Cell Stem Cell* **25**, 419–432.e9 (2019).
36. Conley, A. B. & Jordan, I. K. Epigenetic regulation of human cis-natural antisense transcripts. *Nucleic Acids Res.* **40**, 1438–1445 (2012).
37. Lee, M. P. *et al.* Loss of imprinting of a paternally expressed transcript, with antisense orientation to KVLQT1, occurs frequently in Beckwith-Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 5203–5208 (1999).
38. Sakai, S. *et al.* Long Noncoding RNA ELIT-1 acts as a Smad3 cofactor to facilitate TGFβ/Smad signaling and promote epithelial-mesenchymal transition. *Cancer Res.* **79**, 2821–2838 (2019).
39. Lee, J. T. & Lu, N. Targeted mutagenesis of *Tsix* leads to nonrandom X inactivation. *Cell* **99**, 47–57 (1999).
40. Lee, J. T., Davidow, L. S. & Warshawsky, D. *Tsix*, a gene antisense to *Xist* at the X-inactivation centre. *Nat. Genet.* **21**, 400–404 (1999).
41. Lee, J. T. Disruption of imprinted X inactivation by parent-of-origin effects at *Tsix*. *Cell* **103**, 17–27 (2000).
42. Ohhata, T., Hoki, Y., Sasaki, H. & Sado, T. Crucial role of antisense transcription across the *Xist* promoter in *Tsix*-mediated *Xist* chromatin modification. *Development* **135**, 227–235 (2008).
43. van Aghoven, T. *et al.* Identification of BCAR3 by a random search for genes involved in antiestrogen resistance of human breast cancer cells. *EMBO J.* **17**, 2799–2808 (1998).
44. Near, R. I., Zhang, Y., Makinje, A., Vanden Borre, P. & Lerner, A. AND-34/BCAR3 differs from other NSP homologs in induction of anti-estrogen resistance, cyclin D1 promoter activation and altered breast cancer cell morphology. *J. Cell. Physiol.* **212**, 655–665 (2007).
45. Sheng, G., dos Reis, M. & Stern, C. D. Churchill, a zinc finger transcriptional activator, regulates the transition between gastrulation and neurulation. *Cell* **115**, 603–613 (2003).
46. Brigelius-Flohé, R. & Maiorino, M. Glutathione peroxidases. *Biochim. Biophys. Acta* **1830**, 3289–3303 (2013).
47. Wang, Y. *et al.* GPX2 suppression of H2O2 stress regulates cervical cancer metastasis and apoptosis via activation of the β-catenin-WNT pathway. *Oncol. Targets Ther.* **12**, 6639–6651 (2019).
48. Rozanova, S. *et al.* Quantitative mass spectrometry-based proteomics: An overview. *Methods Mol. Biol.* **2228**, 85–116 (2021).
49. Najafi, S. *et al.* Gene regulation by antisense transcription: A focus on neurological and cancer diseases. *Biomed. Pharmacother.* **145**, 112265 (2022).
50. Wanowska, E., Kubiak, M. R., Rosikiewicz, W., Makalowska, I. & Szcześniak, M. W. Natural antisense transcripts in diseases: From modes of action to targeted therapies. *Wiley Interdiscip. Rev. RNA* **9**, e1461 (2018).
51. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
52. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
53. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
54. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
55. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
56. Sherman, B. T. *et al.* DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, 216–221 (2022).
57. Tamura, Y. *et al.* Homologous recombination is reduced in female embryonic stem cells by two active X chromosomes. *EMBO Rep.* **22**, e52190 (2021).

Acknowledgements

We thank Mika Yoshida, Kumi Akita, Naoki Kurita, Momoka Iwase, Kohei Hashida, Yuma Yamamoto, Ryo Iwamatsu, and Nene Imai for technical support, Michio Kimura and Sanshiro Togo for technical advice concerning programming, Masaomi Kato for help with next-generation sequencing, Naohito Nozaki for providing an anti-H3K36me3 antibody, and Jeremy Allen, Ph.D., from Edanz (<https://jp.edanz.com/ac>) for editing a draft of this

manuscript. This work was supported by a Grant-in-Aid for Scientific Research (C) Grant Number JP20K06541 and HUSM Grant-in-Aid, years 2019 and 2020 to T.O.

Author contributions

Conceptualization, T.O., Methodology, T.O., M.S., K.O., H.Y., C.U., Validation, T.O., M.S., S.S., H.N., Formal Analysis, T.O., M.S., Investigation, T.O., M.S., H.Y., Resources, T.O., M.S., S.S., K.O., Writing—Original Draft, T.O., M.S., Supervision, T.O., M.K., Project Administration, T.O., M.K., Funding Acquisition, T.O., M.K.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-19782-5>.

Correspondence and requests for materials should be addressed to T.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022