Stable intronic sequence RNAs (sisRNAs) are selected regions in introns with distinct properties

Jing Jin
Huazhong University of Science and Technology Tongji Medical College
Ximiao He
Huazhong University of Science and Technology Tongji Medical College
Elena M Silva (emc26@georgetown.edu)
Georgetown University https://orcid.org/0000-0002-1071-9081

Research article

Keywords: sisRNA, intron, noncoding RNA, oocyte, transcription, Xenopus

Posted Date: January 20th, 2020

DOI: https://doi.org/10.21203/rs.2.14880/v2

Version of Record: A version of this preprint was published on April 7th, 2020. See the published version at https://doi.org/10.1186/s12864-020-6687-9.

Abstract

Background: Stable introns and intronic fragments make up the largest population of RNA in the oocyte nucleus of the frog Xenopus tropicalis. These stable intronic sequence RNAs (sisRNAs) persist through the onset of zygotic transcription when synchronous cell division has ended and the developing embryo consists of approximately 8000 cells. Despite their abundance, the sequence properties and biological function of sisRNAs are just beginning to be understood.

Results: To characterize this population of noncoding RNA, we identified all of the sisRNAs in the X. tropicalis oocyte nucleus using published high-throughput RNA sequencing data. Our analysis revealed that sisRNAs, have an average length of ~360 bps, are widely expressed from genes with multiple introns, and are derived from specific regions of introns that are GC and TG rich, while CpG poor. They are enriched in introns at both ends of transcripts but preferentially at the 3' end. The consensus binding sites of specific transcription factors such as Stat3 are enriched in sisRNAs, suggesting an association between sisRNAs and transcription factors involved in early development. Evolutionary conservation analysis of sisRNA sequences in seven vertebrate genomes indicates that sisRNAs are as conserved as other parts of introns, but much less conserved than exons.

Conclusion: In total, our results indicate sisRNAs are selected intron regions with distinct properties, supporting a biological function in gene expression regulation.

Introduction

RNA is one of the three major macromolecules essential for living organisms and consists of three major types: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). Whereas mRNAs code for proteins [1], tRNAs and rRNAs are non-coding RNAs that play essential roles in protein translation ([2, 3]). In the past 15 years, several additional types of non-coding RNAs have been shown to play important roles in regulation of gene expression [4], including microRNAs (miRNA, 21-22 nt) [5], small interfering RNAs (siRNA, 20-25 nt) [6], and Piwi-interacting RNAs (piRNA, 29-30 nt) [7]. While these regulatory non-coding RNAs are usually very small (< 50 nt), there are also the long non-coding RNAs (lncRNA), which are longer than 200 nucleotides [8], and most recently identified, the stable intronic sequence RNAs (sisRNAs), the majority of which are several hundreds of nucleotides [9, 10].

In higher eukaryotes, the majority of protein-coding genes have one or more non-coding introns interspersed within the coding sequence that are spliced from the primary transcript [11]. The spliced introns are primarily in the form of a lariat in which the 5' end is linked to the 3'acceptor splice site. These lariats are debranched into a linear form and then degraded rapidly in most cases [12, 13]. The majority of intron fragments are believed to be unstable, with a few exceptions [14-17]. Recently, a number of sisRNAs were identified in the oocyte nucleus (germinal vesicle or GV) of *Xenopus tropicalis* [9] and later in the oocyte cytoplasm [10]. Although less stable than cytoplasmic mRNA, the sisRNAs are very stable. Transcription inhibition studies revealed that they are stable for at least 2 days and RNA sequencing

analysis demonstrated that they are transferred to the egg upon GV breakdown and persist until at least the blastula stage [9]. These sisRNAs are usually only a portion of the full intron and while nuclear sisRNAs are present as either linear or lariat molecules, most cytoplasmic sisRNAs are in lariat form [9, 10]. Thus far, sisRNAs have also been found in human, mouse, chicken, zebrafish the Epstein-Barr virus and *Drosophila melanogaster* [9, 18-22]. Little is known about the sequence properties and biological function of these abundant sisRNAs although recently, sisR-1 and sisR-4, have been shown to participate in a feedback loop to modulate their parental gene expression in *Drosophila* [20, 22-24].

To further characterize sisRNAs, we identified all sisRNAs in the *X. tropicalis* genome using the published high-throughput sequencing data of RNA from the GV [9]. We then determined the average length of sisRNAs, their distribution in genes, sequence composition, transcription factor binding site (TFBS) enrichment, gene ontology and evolutionary conservation. Here we show that sisRNAs are most widely expressed from genes with multiple introns, enriched in GC and TpG. They are enriched in introns at both ends of transcripts but preferentially at the 3' end. They also contain specific transcription factor binding sites (TFBS), which supports recent findings that sisRNAs play a role in the regulation of gene expression.

Results

Genome-wide identification of sisRNAs in Xenopus tropicalis

To study the sequence characteristics of sisRNAs, the high-throughput sequencing data of RNA (RNA-seq) from the GV of the frog *Xenopus tropicalis* [9] was used to detect sisRNA peaks with the model based analysis for ChIP-seq (MACS) algorithm [25]. We identified a total of 63,410 RNA peaks by a more stringent criterion (FDR=0.01) (Figure 1A, Table 1) and compared the location of each RNA peak to exons and introns to identify the true sisRNA peaks (i.e. those from spliced introns that did not cross intron–exon boundaries.) Two widely used and primarily manually annotated gene sets, RefSeq (9,448 protein-coding genes) and Ensembl (28,967 protein-coding genes), were used as references to identify the sisRNA peaks. 24,901 sisRNAs were identified by refSeq genes (termed as refSeq sisRNAs), with a total length of ~9 Mbps, which accounts for 0.6% of the *X. tropicalis* genome (Figure 1A, Table 1). Similarly, 34,169 sisRNAs were identified by Ensembl genes (termed as Ensembl sisRNAs), with a total length of ~12 Mbps, which accounts for 0.8% of the genome (Figure 1A, Table 1). Together, a total of 20,020 peaks were identified by both gene sets (Figure 1A), which represents ~80% of refSeq sisRNAs and ~60% of Ensembl sisRNAs respectively. The sisRNA length ranges from 160 bps to 2,908 bps with the majority 200-500 bps long and an average of 363.6 bps and 356.6 bps for refSeq and Ensembl sisRNAs, respectively (Figure 1B-D, Table 1). Taken together, two high quality datasets of genome-wide sisRNAs were generated for further investigation of sisRNA properties.

sisRNAs are widely expressed and preferentially located in genes with multiple introns

To determine the number of genes with sisRNAs, we first divided the genes of each dataset into 10 groups according to the number of introns in a gene. Genes without introns were excluded from further analysis. From the 9,448 RefSeq genes, 93.8% (8,864) have multiple exons while only 6.2% (584) of the genes have a single exon (Table 2A). Our data show that the more introns a gene has, the more likely the gene generates sisRNAs. While 24.9% of genes with a single intron have sisRNAs located in their intron, the possibility that a gene with 5 introns has a sisRNA increased to 66.6%. When a gene has 10 or more introns the possibility increased to 81.0% (Figure 2A). On average, 67.3% of Refseq genes (5,965/8,864) with introns produce sisRNAs (Figure 2A, Table 2A).

We also calculated the average number of different sisRNAs per gene. We observed that the average sisRNA number increases with intron number (the average number of introns per gene is 7.8 in *Xenopus tropicalis*). For example, genes with a single intron have an average of 0.60 sisRNAs, genes with 5 introns have an average of 2.06 sisRNAs, and genes with 10 or more introns have an average of 5.06 sisRNAs (Figure 2B, Table 2A). On average, there are 2.97 sisRNAs per gene and 0.36 sisRNAs per intron (Table 2A). We also compared the intron length to the number of sisRNAs for each intron, and it indicates that longer introns tender to have more sisRNAs (R=0.72, Supplementary Fig. 1A-B). We observed a similar result for Ensembl genes. On average 49.6% of genes (13,044/26,313) with introns encode sisRNAs. On average there are 1.94 sisRNAs encoded in an Ensemble gene and each intron encodes 0.22 sisRNAs (Figure 2, Table 2B).

The sisRNA sequences are GC and TpG rich

To investigate the base-pair composition in sisRNAs, we first calculated the prevalence of all 10 unique dinucleotides in sisRNAs and in the *X. tropicalis* genome. We observed that sisRNAs as compared to the *X. tropicalis* genome are rich in AC, CC, AG, CA and GC, while poor in CG, AT, AA, TA and GA, for both RefSeq- and Ensembl- identified sisRNAs (Figure 3A). We then extended the calculation to trinucleotides. As we expected, sisRNAs are rich in CCA, GCC, CAC, GCA, CAG, AGC, which are all trinucleotides containing GC or CA|TG, the dinucleotides shown to be enriched in sisRNAs (Figure 3B). sisRNAs are very poor in CGN, which are trinucleotides with CpG, and also poor in ATA, AAT, AAA and TAA (Figure 3B). Generally speaking, CpG rich regions (e.g. CpG islands) are G+C rich and CpG poor regions are A+T rich. Interestingly, we observed that sisRNAs are CpG poor while GC rich.

We then calculated the GC content, CpG density and CA|TG density for each sisRNA as well as each scaffold in the whole genome (Figure 3C-E). Compared to the *X. tropicalis* genome, sisRNAs have a higher GC content (Figure 3C), a lower CpG density (Figure 3D) and a higher CA|TG density (Figure 3E). As a result, in terms of base-pair compositions, sisRNAs are different from the genome, with a different prevalence of dinucleotides, trinucleotides, and a lower CpG density, higher CA|TG density and GC content. In other words, sisRNAs are unique regions of the genome with their own properties. The results for RefSeq sisRNAs and Ensembl sisRNAs are nearly identical (Figure 3), providing strong support for our results.

The sisRNAs are specific regions of the introns

Since sisRNAs are derived from regions of introns, and introns have a distinct base-pair composition as compared to the whole genome (Haddrill et al. 2005), we expect sisRNAs to also have a sequence composition different from the whole genome. However, it remains unclear whether sisRNAs are randomly distributed in introns or in specific regions of the intron. GC content (G+C%), CpG density and CA|TG density are widely used and important parameters to analyze sequences characteristics. Thus, we calculated these parameters for each intron with/without sisRNAs, and compared these results with sisRNAs alone. The total GC content of the introns with sisRNAs (40.59%) is higher ($p<0.001$, t-test) than the genome (40.07%). The GC content in sisRNAs alone (41.92% and 41.79% for RefSeq and Ensembl sisRNAs, respectively) is higher than that of the introns ($p<0.001$, t-test) (Figure 4A, Table 3). The CpG density of introns when compared to the whole genome is poor (Figure 4B). Interestingly, introns with sisRNAs have a higher CpG density than both introns without sisRNAs and sisRNAs alone (Figure 4B). While the CA|TG density in introns is very similar to the genome, sisRNAs have a higher CA|TG density than introns with sisRNAs (Figure 4C). Thus, the sisRNAs are in regions of the introns with higher CA|TG density and GC%. We further divided the introns without sisRNAs into two groups according to whether the host gene is with/without sisRNAs: Intron A (the host gene without sisRNAs) or Intron B (the host gene has sisRNAs but the intron itself is without sisRNAs). We found that intron A and intron B are very similar, thus sisRNAs are closely associated with the introns from which they originate.

Taken together, these results reveal that sisRNAs are specific regions of introns with distinct sequence compositions.

sisRNAs are enriched at both 5’ and 3’ end of transcripts, with a preference for the 3’ end

After we observed that sisRNAs are in specific regions of introns, with unique base-pair compositions, we analyzed whether they are derived from specific introns along the gene. To study where sisRNAs are enriched, we concatenated all the introns for each gene along the transcript. We divided each joined intron transcript into 100 bins, and identified the bins in which the sisRNAs are located. As shown in Figure 5A, the sisRNAs are mostly enriched at the beginning (5’ end) and the end (3’end) of the transcript with a preference for the 3’ end. An example is shown for the gene *nasp*: more sisRNAs were observed at the 3’ end (Figure 5B). These results further confirmed that sisRNAs are not random sequences from the introns: they have distinct sequence compositions and are preferentially driven from the 3’ end of a transcript.

We next asked whether sisRNAs are derived from the 5’ and 3’ end of introns because these end introns have unique properties. To investigate this possibility, we divided all the introns from genes with more than 3 introns into 3 categories: 1st -5’ end introns (S), middle introns (M) and last -3’ end introns (E). The results indicates that last -3’ end introns are very similar to the middle intron, while the first -5’ end introns

have different compositions and are slightly longer (Supplementary Figure 2). The properties of the first introns may be different because they sometimes overlap with the promoter regions, which have higher CpG, CA|TG density, and GC content. These results indicate that sisRNAs are not preferentially derived from 3' end introns because these introns have unique properties, but instead some other mechanism of sisRNA production is involved.

Another possibility is that sisRNAs produced from 3' end introns are remnants of gene transcription and splicing. To test this idea, we performed the analysis of correlation between the expression levels of host genes and sisRNAs. As shown in Figure 5C, the expression levels of host genes (FPKM) and sisRNAs peak signals are negatively relevant (R=-0.1724), which does not support sisRNAs as remnants of gene transcription and splicing that have not been cleaned from cells. This suggests that sisRNAs play an inhibitory role in host gene expression. Taken together, sisRNAs are most enriched in the 3' end of introns, and the mechanism for this enrichment remains to be investigated.

Specific TFBS related to transcription regulation are enriched in sisRNAs

Because of the specific sequence properties and preference for introns on the ends of a gene, we performed an enrichment calculation of transcription factor binding sites (TFBS) in sisRNAs to study the potential functions of sisRNAs. We searched for enriched DNA motifs among the 935 position weight matrices (PWMs) collected from the TRANSFAC databases [26] in sisRNAs and in introns without sisRNAs (Figure 6). We also calculated the motif enrichment in both RefSeq and Ensembl sisRNAs and the enrichments in two datasets are nearly identical (R=0.99), indicating the calculation is robust and the quality of identified sisRNAs is high (Figure 6B). Stat3, NF-κB, p50:p50, MYOG:NF1 and GAF consensus sites are enriched in sisRNAs but depleted in introns without sisRNAs (Figure 6A). Stat3 (signal transducer and activator of transcription 3) is a member of the STAT protein family, functions as a transcriptional activator [27], and is highly expressed in the *X. tropicalis* cytoplasm (Figure 6C). NF-κB (nuclear factor kappa-light-chain-enhancer of activated B cells) is a protein complex that controls transcription [28]. p50 is the mature NF-κB subunit, which has no intrinsic ability to activate transcription and has been proposed to act as a transcriptional repressor when binding with κB elements as homodimers (p50:p50) [29]. Myogenin (MYOG) is one of four muscle-specific basic helix-loop-helix regulatory factors involved in controlling myogenesis [30], and NF-1 (Neurofibromin 1) is a negative regulator of the Ras signal transduction pathway [31] and is required for skeletal muscle development [32]. The GAGA factor (GAF) is one of a few transcription factors that can regulate transcription at multiple levels: depending on its target genomic location, it can act as either activator or repressor [33]. We also observed that some TFBS, such as Ncx, Prop1 and Nkx3a, are depleted in sisRNAs while slightly enriched in introns without sisRNAs (Figure 6A). These results suggest that specific TFBS involved in transcription regulation, either activation or suppression, are enriched in sisRNAs, which imply that sisRNAs may play a functional role in transcriptional regulation.

The GO terms Nucleotide binding, RNA binding and ATP binding are enriched in genes with sisRNAs

To further investigate the role of sisRNAs, we asked whether the genes with sisRNAs share any function. GO (Gene Ontology) enrichment analysis [34] of the 5,419 RefSeq genes with sisRNAs indicated that 13.9% of the genes are involved in nucleotide binding, 3.2% of the genes are involved in RNA binding and 7.9% genes are involved in ATP binding (Table 4). The other enriched gene sets included RNA processing, mRNA metabolic process and translation (Table 4). These results suggested that sisRNAs might have a potential biological function involved in gene regulation and metabolism.

The sisRNAs are as evolutionary conserved as introns, and much less than exons

Our data indicate that sisRNAs are in specific regions of introns, contain TFBSs, and are in the introns of genes involved in nucleotide, ATP and RNA binding. To investigate whether these sisRNAs are conserved across species, we determined the PhastCons conservation score for sisRNAs and introns (Figure 7). The PhastCons scores [35] are calculated based on multiple alignments of 6 vertebrate genomes (zebrafish, chicken, opossum, rat, mouse and human) with *X. tropicalis*. As expected, the boundary of exon and intron has the highest conservation score (Figure 7A, C). Although sisRNAs are as conserved as other intron regions, they are still much less conserved than exons, suggesting they might not be conserved among species, which was also shown in a recent study [10].

An alternative approach to evaluate the conservation of sisRNAs is to compare the sisRNAs in different cell types from various species. We obtained the cytoplasm sisRNAs data recently published [21], which contains several species and cell types, including human red blood cells, Hela cells, mouse red blood cells and 3T3 cells, chicken DF1 cells, and *Xenopus laevis* XTC cells. We compared the GC content, CpG density, and CA|TG density among these cytoplasm sisRNAs to the refSeq and Ensembl sisRNAs that we identified (Figure 7D-F). We found that between different cell types in the same species, for example, human red blood cells and Hela cells, the GC content, CpG density and CA|TG density of these sisRNAs are quite different from each other (Figure 7D-F). While for the same cell types between different species, for example, human red blood cells and mouse red blood cells, some sequence composition similarities exist (Figure 7D-F). These results indicate that sisRNAs are cell type specific. Although sequences of sisRNAs may not be so conserved among species, the sequence compositions are similar.

The *X. laevis* XTC sisRNAs are very similar to *X. tropicalis* GV sisRNAs in terms of sequence composition. We next asked how many common host genes with sisRNAs exist between the two species. We compared 5,563 unique genes host sisRNAs in GV of *X. tropicalis* to 2,029 unique genes which have sisRNAs in *X. laevis* XTS cells, there are only 602 genes overlapped and this chance is even lower than random selection (Figure 7G). Among these 602 genes, the number of sisRNAs from the same intron is

less than 50 (data not shown). Overall, our calculation suggested there may not be positional or sequence conservation between sisRNAs, which is also supported by other recent work [10].

We also performed GO (Gene Ontology) enrichment analysis for the genes host sisRNAs in each cell type (Supplementary Figure 3A-F). In chicken DF1 cells, the sisRNAs are not enriched in any gene sets (Supplementary Figure 3E). Even though sisRNAs are enriched in some gene sets in the other 5 cell types, there are no common enriched gene sets present in all 5. The most common enriched gene set is related to sister chromatid segregation or nuclear division, which is observed in human RBC, Hela, mouse RBC, and *X. laevis* XTC cells (Supplementary Figure 3A-C, F). Another common enriched biological process is covalent chromatin modification, which is observed in human RBC, mouse RBC and 3T3 cells (Supplementary Figure 3A, C-D). These results imply that sisRNA may play functional roles in cell division or chromatin organization.

Discussion

The function of sisRNAs in Xenopus tropicalis

The biological function of sisRNAs in the *Xenopus* oocyte nucleus is unclear but of particular interest especially in light of their stability and abundance. The majority of zygotic transcription in the *Xenopus* embryo begins at the midblastula transition (MBT), after the 12^{th} cell division. The transcription rate must be extremely high in the oocyte and the transcripts must be very stable to allow a sufficient amount of RNA to be deposited in the cytoplasm such that the RNA:DNA ratio is not significantly depleted as the cells divide before MBT [36, 37]. In this case, the sisRNAs could simply be the byproducts of universal stable RNA transcripts [9]. However, we found that sisRNAs are not randomly distributed in introns, but rather in specific regions of the intron, with a unique sequence composition, indicating that these sisRNAs are selected to be stable and abundant and thus are very likely to have a relevant biological function in the *Xenopus* oocyte and early embryonic development.

Recently, 9000 sisRNAs have been found in the cytoplasm with about half of these confirmed as lariat molecules [10]. These sisRNAs are only derived from a relatively small number of specific introns [10], which further confirmed our observations that sisRNAs are not random sequences but specific regions of introns. Besides *Xenopus*, numerous circular intronic sequences have been identified in cultured human cell lines (Hela and H9) [19] implying that they are widely expressed in many different species and may have a significant biological role[22].

The association of sisRNAs with Stat3

We showed that consensus TFBS of Stat3 is the most enriched motif in the introns from which sisRNAs are generated, and observed that Stat3 is highly expressed in *Xenopus tropicalis* oocytes [9]. A recent microarray study showed that Stat3 is expressed at stage 2 and peaks at stage 8 and is still detectable as late as stage 33 in both *X. tropicalis* and *X. laevis* [38]. It has been widely reported that Stat3 can bind to

intronic regions to regulate gene expression. For example, the expression of BCL3 is induced by IL-6 via Stat3 binding to intronic enhancer HS4 [39]. The signaling factor, WNT5A is an evolutionarily conserved target of the Stat3 signaling cascade based on 11-bp-spaced tandem Stat3-binding sites within intron 4 of human, chimpanzee, cow, mouse and rat WNT5A orthologs [40]. Stat3 binding to the introns of Foxp3, RORα, RORγt, and IL-6Rα have also been reported [41-43]. Analysis of ~75,000 Stat3 binding sites identified by chromatin immunoprecipitation (ChIP)-seq in a transformed human breast cell line revealed that most Stat3 binding sites are located within introns [44]. In total, these data indicate a role for Stat3 in the regulation of expression of the genes from which sisRNAs arise. It is also possible that specific TF interactions are involved in the biogenesis of sisRNAs. Both of these seem more likely considering STAT3 is a TF that binds DNA and RNA [47]. A recent study showed that lncRNA directly binds STAT3 in the cytoplasm of human dendritic cells (DC), thereby preventing dephosphorylation of STAT3 by SHP1, and controlling the differentiation of DC [48]. Another study demonstrated an interaction between STAT3 and circular RNAs [49], which may be a type of sisRNA. Taken together, future experiments can be designed to test if sisRNAs interact with STAT3 and if their formation is dependent on STAT3.

The evolutionary origin of sisRNAs

DNA methylation at the 5' position of cytosine (5mC), primarily in CpG context, is observed in nearly every vertebrate examined, including *Xenopus tropicalis* [50]. 5mC can deaminate to thymine 10-50 times faster than the mutation rate of other nucleotides [51]. Deamination of 5mC caused the high depletion of the CpG dinucleotide in mammalian genomes [52], and as a result, TG (the deamination product of 5mCG) is the most abundant dinucleotide in vertebrates [53]. The unmethylated CGs [54] tend to be clustered together into CG islands (CGI) [55]. As a result, CpG rich regions (CG islands) have a higher GC content, whereas CpG poor regions have a lower GC content. We observed that CpG is depleted in sisRNAs as compared to the *X. tropicalis* genome, and as expected, TG is enriched. Interestingly, we observed higher GC content in sisRNAs, which suggested that the base composition of sisRNAs is not merely the consequence of deamination of 5mC, there must be other mechanisms playing a role in the evolutionary origin of sisRNAs. A recent study showed that exons of lncRNA loci also have a high GC content due to purifying selection [56], thus it is possible that sisRNAs share some evolutionary properties with lncRNAs. Surprisingly, the CpG density in sisRNAs alone is very similar to introns without sisRNAs, and lower than that of introns with sisRNAs. This suggests that sisRNAs are surrounded by regions with a high CpG level. In other words, although sisRNAs themselves are CpG poor, they are derived from introns with high CpG density, and higher CpG density in the introns may be indicative of producing sisRNAs.

We also assessed the evolutionary conservation of the *X. tropicalis* sisRNA sequences in six other vertebrate genomes, including zebrafish, chicken, opossum, rat, mouse and human. Our results indicated that sisRNAs are not conserved: sisRNAs are as conserved as other parts of introns, but much less conserved than the exons. Cytoplasmic sisRNAs have been identified recently in *X. tropicalis* oocytes and these sisRNAs are only derived from a relatively small number of specific introns [10], it is worth noting that these sisRNAs are also not conserved.

Thus far sisRNAs are identified in human, mouse, chicken, zebrafish, *Xenopus* and *Drosophila*, implying that sisRNAs might be widely expressed in many other species[21, 22]. Considering that sisRNAs are not conserved, if sisRNAs do have some biological functions such as gene regulation, it might be cell type and species-specific.

In conclusion, our results suggest that sisRNAs are not transcribed from random part of introns but specific regions with distinct properties. The sisRNAs are GC rich while CpG poor, and preferentially enriched in the 3' end of the mRNA transcript. Specific TFBSs involved in gene regulation are enriched in the regions from which sisRNAs arise, suggesting an association with specific proteins, such as Stat3, and further experiments are required to investigate this association. With more and more sisRNA data available in different species, the potential biological functions of sisRNAs would hopefully to be revealed soon.

Methods

Dataset generation

The GV and cytoplasmic RNA-seq data were obtained from the Gall lab[9]. sisRNAs in human red blood cells, Hela cells, mouse red blood cells and 3T3 cells, chicken DF1 cells, *X. laevis* XTC cells were obtained from [21]. The dataset of refSeq and Ensembl genes, PhastCons conservation scores, as well as the genome sequences of *Xenopus tropicalis* were downloaded from the University of California Santa Cruz Genome Bioinformatics website (http://genome.ucsc.edu/) [57]. The reference *X. tropicalis* genome assembly was xenTro2 (assembly version 4.1). Exons, introns, 5'UTRs, and 3'UTRs for refSeq genes and Ensembl genes were determined using UCSC annotations. UCSC genome browser screen shots were generated using custom tracks of the UCSC web site (https://genome.ucsc.edu/).

Identification of sisRNAs in germinal vesicles

The Model-Based Analysis of ChIP-seq algorithm (MACS) [25] was used for detecting sisRNAs peaks by analyzing germinal vesicle (GV) RNA-seq data [9]. First, we identified all 63,410 peaks of the GV RNA-seq data with default parameters, but a more stringent FDR=0.01 (default is 0.05). We then determined the location of each peak relative to exon and intron annotation for refSeq and Ensembl genes, respectively. A peak would be identified as sisRNA if it was located within an intron and did not overlap with an exon. In this way we identified 24,091 and 34,169 sisRNAs for refSeq and Ensembl annotated genes, respectively.

Calculation of sisRNA density along introns

[See supplementary files]

Calculation of motif enrichment in sisRNAs and Introns

[See supplementary files]

Gene Ontology analysis

Gene Ontology (GO) analysis was performed using DAVID (The Database for Annotation, Visualization and Integrated Discovery, http://david.abcc.ncifcrf.gov/) [59, 60]. Go terms with P-values < 0.01 were considered as significantly enriched.

Evolutionary conservation analysis

Base by base PhastCons conservation scores based on an alignment and a model of neutral evolution among the seven vertebrate genomes [35] were downloaded from UCSC database (http://genome.ucsc.edu/). The seven genomes and assemblies are: zebrafish (danRer4), *X. tropicalis* (xenTro2), chicken (galGal2), opossum (monDom4), rat (rn4), mouse (mm8) and human (hg18). PhastCons scores in each sisRNA or intron were extracted for each nucleotide for ±150 bps relative to the 5' end, midpoint, and 3' end respectively. Values were averaged for all sisRNAs or introns.

Declarations

Ethics approval and consent to participate

Not applicable

Consent to publish

Not applicable

Availability of data and materials

The GV and cytoplasmic RNA-seq data were published by the Gall lab, which can be found at their lab website (https://emb.carnegiescience.edu/grace/datasets). All of the other datasets were download from UCSC web site (https://genome.ucsc.edu/). The datasets supporting the conclusions of this article are included within the article and its additional files.

Competing interests

The authors have no competing interests to declare.

Funding

There was no outside funding source for completion of this work. Both JJ and ES were supported by Georgetown University during completion of this work. XH was supported by Tongji Medical College during preparation of the manuscript and revisions.

Authors' contributions

JJ, XH and ES conceived the study. ES supervised the study. JJ and XH performed the bioinformatical analyses. JJ drafted the manuscript. XH and ES revised the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgments

We thank Dr. Joseph G. Gall for sharing the high-throughput sequencing data of RNA from GV in *Xenopus tropicalis*.

References

1. Jacob F, Monod J: **Genetic regulatory mechanisms in the synthesis of proteins**. *J Mol Biol* 1961, **3**:318-356.
2. Sharp SJ, Schaack J, Cooley L, Burke DJ, Soll D: **Structure and transcription of eukaryotic tRNA genes**. *CRC Crit Rev Biochem* 1985, **19**(2):107-144.
3. van Nues RW, Venema J, Rientjes JM, Dirks-Mulder A, Raue HA: **Processing of eukaryotic pre-rRNA: the role of the transcribed spacers**. *Biochem Cell Biol* 1995, **73**(11-12):789-801.
4. Huttenhofer A, Schattner P, Polacek N: **Non-coding RNAs: hope or hype?** *Trends Genet* 2005, **21**(5):289-297.
5. Ambros V: **The functions of animal microRNAs**. *Nature* 2004, **431**(7006):350-355.
6. Hamilton AJ, Baulcombe DC: **A species of small antisense RNA in posttranscriptional gene silencing in plants**. *Science* 1999, **286**(5441):950-952.
7. Girard A, Sachidanandam R, Hannon GJ, Carmell MA: **A germline-specific class of small RNAs binds mammalian Piwi proteins**. *Nature* 2006, **442**(7099):199-202.
8. Perkel JM: **Visiting "noncodarnia"**. *Biotechniques* 2013, **54**(6):301, 303-304.

9. Gardner EJ, Nizami ZF, Talbot CC, Jr., Gall JG: **Stable intronic sequence RNA (sisRNA), a new class of noncoding RNA from the oocyte nucleus of Xenopus tropicalis**. *Genes Dev* 2012, **26**(22):2550-2559.
10. Talhouarne GJ, Gall JG: **Lariat intronic RNAs in the cytoplasm of Xenopus tropicalis oocytes**. *RNA* 2014, **20**(9):1476-1487.
11. Wahl MC, Will CL, Luhrmann R: **The spliceosome: design principles of a dynamic RNP machine**. *Cell* 2009, **136**(4):701-718.
12. Domdey H, Apostol B, Lin RJ, Newman A, Brody E, Abelson J: **Lariat structures are in vivo intermediates in yeast pre-mRNA splicing**. *Cell* 1984, **39**(3 Pt 2):611-621.
13. Chapman KB, Boeke JD: **Isolation and characterization of the gene encoding yeast debranching enzyme**. *Cell* 1991, **65**(3):483-492.
14. Michaeli T, Pan ZQ, Prives C: **An excised SV40 intron accumulates and is stable in Xenopus laevis oocytes**. *Genes Dev* 1988, **2**(8):1012-1020.
15. Kopczynski CC, Muskavitch MA: **Introns excised from the Delta primary transcript are localized near sites of Delta transcription**. *J Cell Biol* 1992, **119**(3):503-512.
16. Qian L, Vu MN, Carter M, Wilkinson MF: **A spliced intron accumulates as a lariat in the nucleus of T cells**. *Nucleic Acids Res* 1992, **20**(20):5345-5350.
17. Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL: **Genomewide characterization of non-polyadenylated RNAs**. *Genome Biol* 2011, **12**(2):R16.
18. Moss WN, Steitz JA: **Genome-wide analyses of Epstein-Barr virus reveal conserved RNA structures and a novel stable intronic sequence RNA**. *BMC Genomics* 2013, **14**:543.
19. Zhang Y, Zhang XO, Chen T, Xiang JF, Yin QF, Xing YH, Zhu S, Yang L, Chen LL: **Circular intronic long noncoding RNAs**. *Mol Cell* 2013, **51**(6):792-806.
20. Pek JW, Osman I, Tay ML, Zheng RT: **Stable intronic sequence RNAs have possible regulatory roles in Drosophila melanogaster**. *J Cell Biol* 2015, **211**(2):243-251.
21. Talhouarne GJS, Gall JG: **Lariat intronic RNAs in the cytoplasm of vertebrate cells**. *Proc Natl Acad Sci U S A* 2018, **115**(34):E7970-E7977.
22. Chan SN, Pek JW: **Stable Intronic Sequence RNAs (sisRNAs): An Expanding Universe**. *Trends Biochem Sci* 2019, **44**(3):258-272.
23. Tay ML, Pek JW: **Maternally Inherited Stable Intronic Sequence RNA Triggers a Self-Reinforcing Feedback Loop during Development**. *Curr Biol* 2017, **27**(7):1062-1067.
24. Wong JT, Akhbar F, Ng AYE, Tay ML, Loi GJE, Pek JW: **DIP1 modulates stem cell homeostasis in Drosophila through regulation of sisR-1**. *Nat Commun* 2017, **8**(1):759.
25. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al*: **Model-based analysis of ChIP-Seq (MACS)**. *Genome Biol* 2008, **9**(9):R137.
26. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K *et al*: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes**. *Nucleic Acids Res* 2006, **34**(Database issue):D108-110.

27. Akira S, Nishio Y, Inoue M, Wang XJ, Wei S, Matsusaka T, Yoshida K, Sudo T, Naruto M, Kishimoto T: **Molecular cloning of APRF, a novel IFN-stimulated gene factor 3 p91-related transcription factor involved in the gp130-mediated signaling pathway**. *Cell* 1994, **77**(1):63-71.
28. Sen R, Baltimore D: **Multiple nuclear factors interact with the immunoglobulin enhancer sequences**. *Cell* 1986, **46**(5):705-716.
29. Plaksin D, Baeuerle PA, Eisenbach L: **KBF1 (p50 NF-kappa B homodimer) acts as a repressor of H-2Kb gene expression in metastatic tumor cells**. *J Exp Med* 1993, **177**(6):1651-1662.
30. Funk WD, Wright WE: **Cyclic amplification and selection of targets for multicomponent complexes: myogenin interacts with factors recognizing binding sites for basic helix-loop-helix, nuclear factor 1, myocyte-specific enhancer-binding factor 2, and COMP1 factor**. *Proc Natl Acad Sci U S A* 1992, **89**(20):9484-9488.
31. Trovo-Marqui AB, Tajara EH: **Neurofibromin: a general outlook**. *Clin Genet* 2006, **70**(1):1-13.
32. Kossler N, Stricker S, Rodelsperger C, Robinson PN, Kim J, Dietrich C, Osswald M, Kuhnisch J, Stevenson DA, Braun T *et al*: **Neurofibromin (Nf1) is required for skeletal muscle development**. *Hum Mol Genet* 2011, **20**(14):2697-2709.
33. Adkins NL, Hagerman TA, Georgel P: **GAGA protein: a multi-faceted transcription factor**. *Biochem Cell Biol* 2006, **84**(4):559-567.
34. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
35. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al*: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res* 2005, **15**(8):1034-1050.
36. Davidson EH: **Gene activity in early development**, 3rd edn. Orlando: Academic Press; 1986.
37. Callan HG: **Lampbrush chromosomes**. *Mol Biol Biochem Biophys* 1986, **36**:1-252.
38. Yanai I, Peshkin L, Jorgensen P, Kirschner MW: **Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility**. *Dev Cell* 2011, **20**(4):483-496.
39. Brocke-Heidrich K, Ge B, Cvijic H, Pfeifer G, Loffler D, Henze C, McKeithan TW, Horn F: **BCL3 is induced by IL-6 via Stat3 binding to intronic enhancer HS4 and represses its own transcription**. *Oncogene* 2006, **25**(55):7297-7304.
40. Katoh M, Katoh M: **STAT3-induced WNT5A signaling loop in embryonic stem cells, adult normal tissues, chronic persistent inflammation, rheumatoid arthritis and cancer (Review)**. *Int J Mol Med* 2007, **19**(2):273-278.
41. Zorn E, Nelson EA, Mohseni M, Porcheray F, Kim H, Litsa D, Bellucci R, Raderschall E, Canning C, Soiffer RJ *et al*: **IL-2 regulates FOXP3 expression in human CD4+CD25+ regulatory T cells through a STAT-dependent mechanism and induces the expansion of these cells in vivo**. *Blood* 2006, **108**(5):1571-1579.

42. Durant L, Watford WT, Ramos HL, Laurence A, Vahedi G, Wei L, Takahashi H, Sun HW, Kanno Y, Powrie F *et al*: **Diverse targets of the transcription factor STAT3 contribute to T cell pathogenicity and homeostasis**. *Immunity* 2010, **32**(5):605-615.
43. Carpenter RL, Lo HW: **STAT3 Target Genes Relevant to Human Cancers**. *Cancers (Basel)* 2014, **6**(2):897-925.
44. Fleming J, Giresi P, Lindahl-Allen M, Krall E, Lieb J, Struhl K: **STAT3 acts through pre-existing nucleosome-depleted regions bound by FOS during an epigenetic switch linking inflammation to cancer**. *Epigenetics & Chromatin* 2015, **8**(1):7.
45. He G, Karin M: **NF-kB and STAT3 - key players in liver inflammation and cancer**. *Cell Res* 2011, **21**:159 - 168.
46. Fan Y, Mao R, Yang J: **NF-kB and STAT3 signaling pathways collaboratively link inflammation to cancer**. *Protein Cell* 2013, **4**:176 - 185.
47. Sigova AA, Abraham BJ, Ji X, Molinie B, Hannett NM, Guo YE, Jangi M, Giallourakis CC, Sharp PA, Young RA: **Transcription factor trapping by RNA in gene regulatory elements**. *Science* 2015, **350**(6263):978-981.
48. Wang P, Xue Y, Han Y, Lin L, Wu C, Xu S, Jiang Z, Xu J, Liu Q, Cao X: **The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation**. *Science* 2014, **344**(6181):310-313.
49. Yang ZG, Awan FM, Du WW, Zeng Y, Lyu J, Wu, Gupta S, Yang W, Yang BB: **The Circular RNA Interacts with STAT3, Increasing Its Nuclear Translocation and Wound Repair by Modulating Dnmt3a and miR-17 Function**. *Mol Ther* 2017, **25**(9):2062-2074.
50. Bogdanovic O, Long SW, van Heeringen SJ, Brinkman AB, Gomez-Skarmeta JL, Stunnenberg HG, Jones PL, Veenstra GJ: **Temporal uncoupling of the DNA methylome and transcriptional repression during embryogenesis**. *Genome Res* 2011, **21**(8):1313-1327.
51. Coulondre C, Miller JH, Farabaugh PJ, Gilbert W: **Molecular basis of base substitution hotspots in Escherichia coli**. *Nature* 1978, **274**(5673):775-780.
52. Bird A, Tate P, Nan X, Campoy J, Meehan R, Cross S, Tweedie S, Charlton J, Macleod D: **Studies of DNA methylation in animals**. *J Cell Sci Suppl* 1995, **19**:37-39.
53. Burge C, Campbell AM, Karlin S: **Over- and under-representation of short oligonucleotides in DNA sequences**. *Proc Natl Acad Sci U S A* 1992, **89**(4):1358-1362.
54. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM *et al*: **Human DNA methylomes at base resolution show widespread epigenomic differences**. *Nature* 2009, **462**(7271):315-322.
55. Gardiner-Garden M, Frommer M: **CpG islands in vertebrate genomes**. *J Mol Biol* 1987, **196**(2):261-282.
56. Haerty W, Ponting CP: **Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci**. *RNA* 2015.

57. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M *et al*: **The UCSC Genome Browser database: 2015 update**. *Nucleic Acids Res* 2015, **43**(Database issue):D670-681.
58. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching**. *Nucleic Acids Res* 2009, **37**(Web Server issue):W202-208.
59. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**(1):44-57.
60. Huang DW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Res* 2009, **37**(1):1-13.

Tables

Please see the supplementary files section to access the tables.

Figures

Figure 1. sisRNAs in Ensembl and RefSeq genes using GV RNA-seq reads

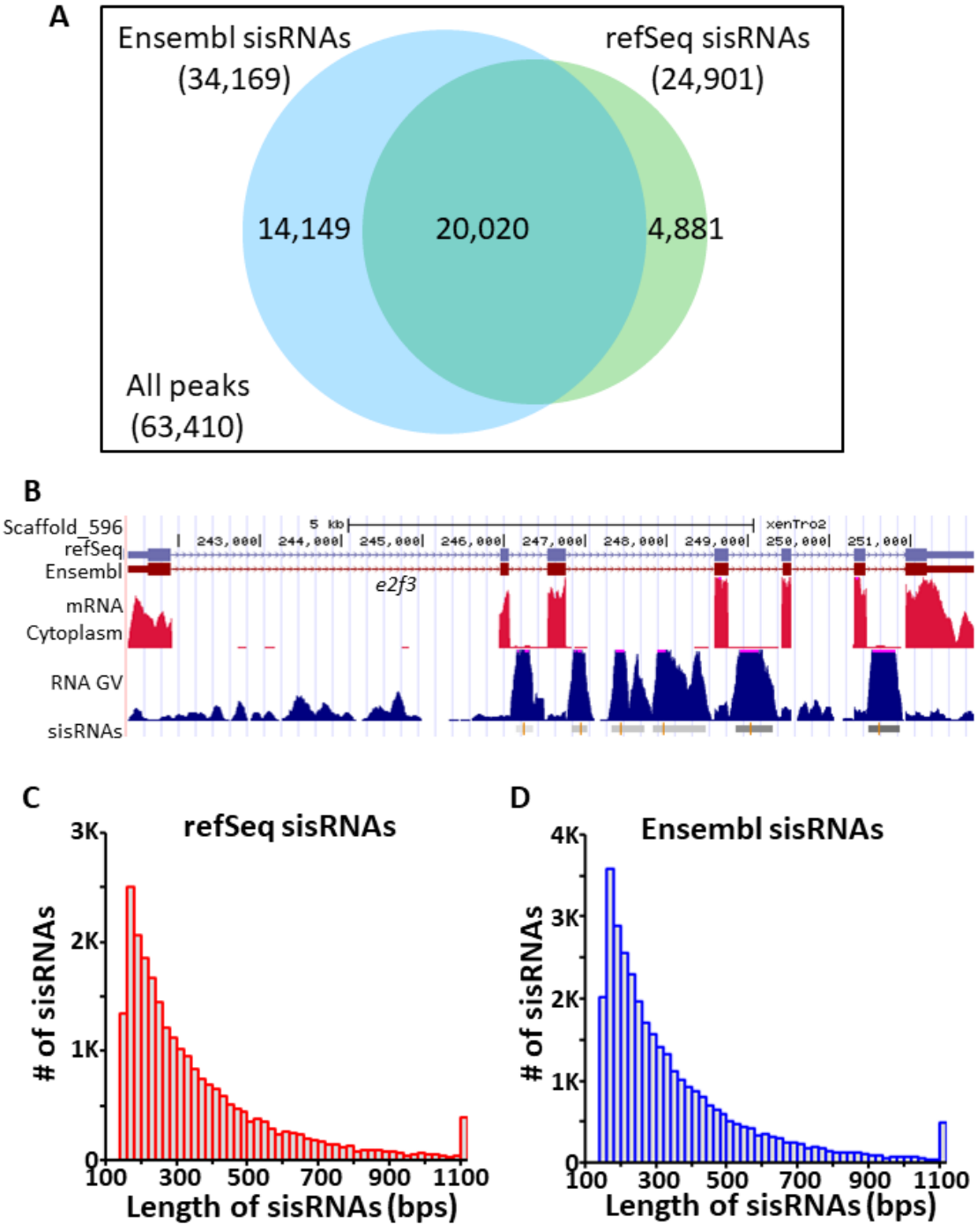

Figure 1

sisRNAs are identified from GV RNA-seq data according to Ensemble and RefSeq gene sets in X. tropicalis. A. Venn diagram of peaks called by MACS with FDR=0.01, and determined to be sisRNAs according to refSeq and Ensembl gene sets. B. UCSC screenshot of identified sisRNAs in the gene E2F3. Red and blue blocks indicate exons, red peaks indicate mRNA detected in the cytoplasm, blue peaks indicate RNA detected in the GV, grey blocks indicate sisRNAs identified, orange lines indicate the

summits of sisRNA peaks, and arrows highlight each sisRNA. C-D. Histograms of length distributions for (C) refSeq and (D) Ensembl sisRNAs.

Figure 2. sisRNAs are widely expressed in multiple intron genes

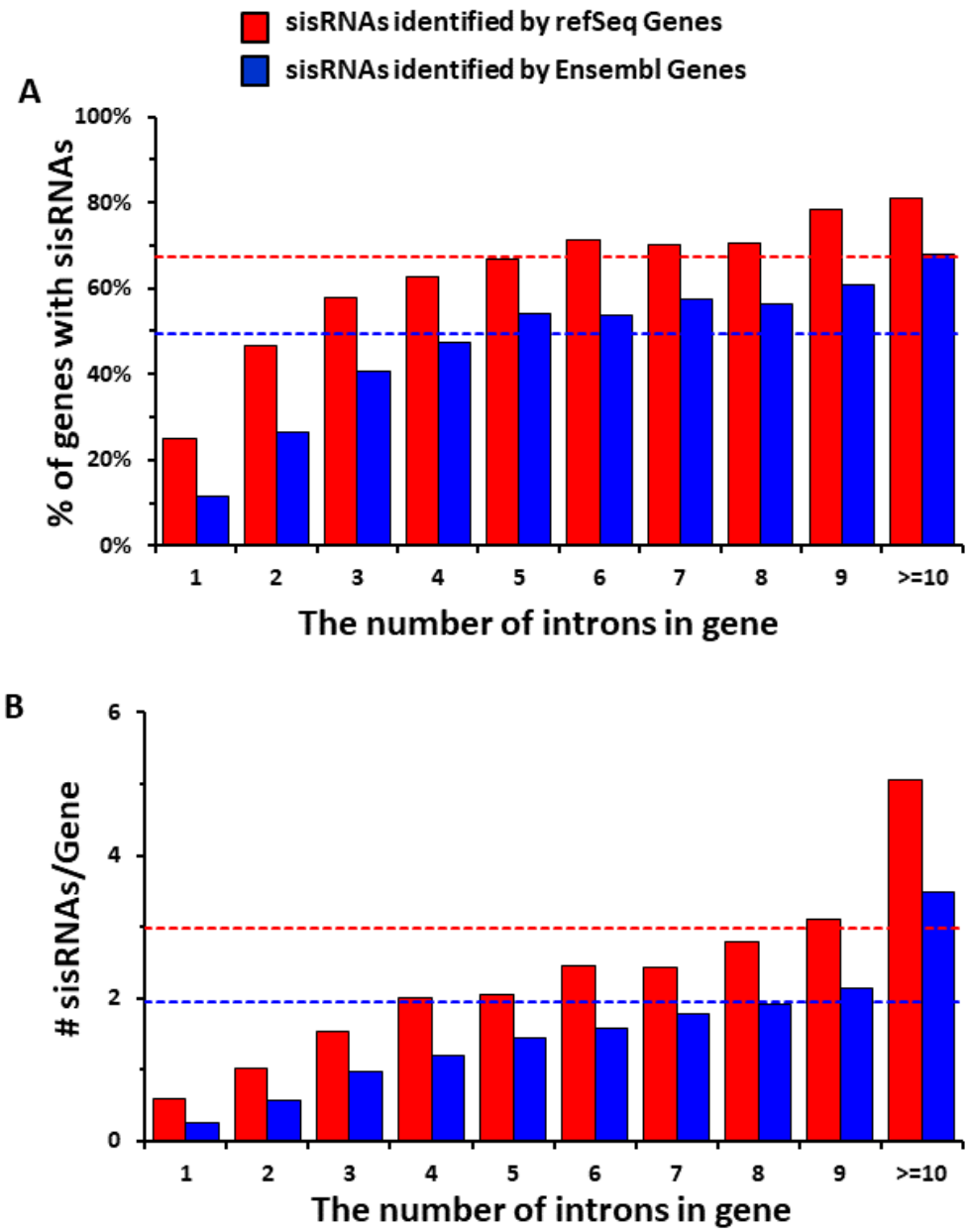

Figure 2

sisRNAs are highly expressed from genes with multiple introns. A-B. Histograms show (A) the percentage of genes with sisRNAs, and (B) the average number of sisRNAs per gene in refSeq (red) and Ensembl

(blue) genes. Genes are grouped by the intron number. The dashed line indicates the average value for all genes.

Figure 3. sisRNAs sequences are GC/CA rich and CpG poor

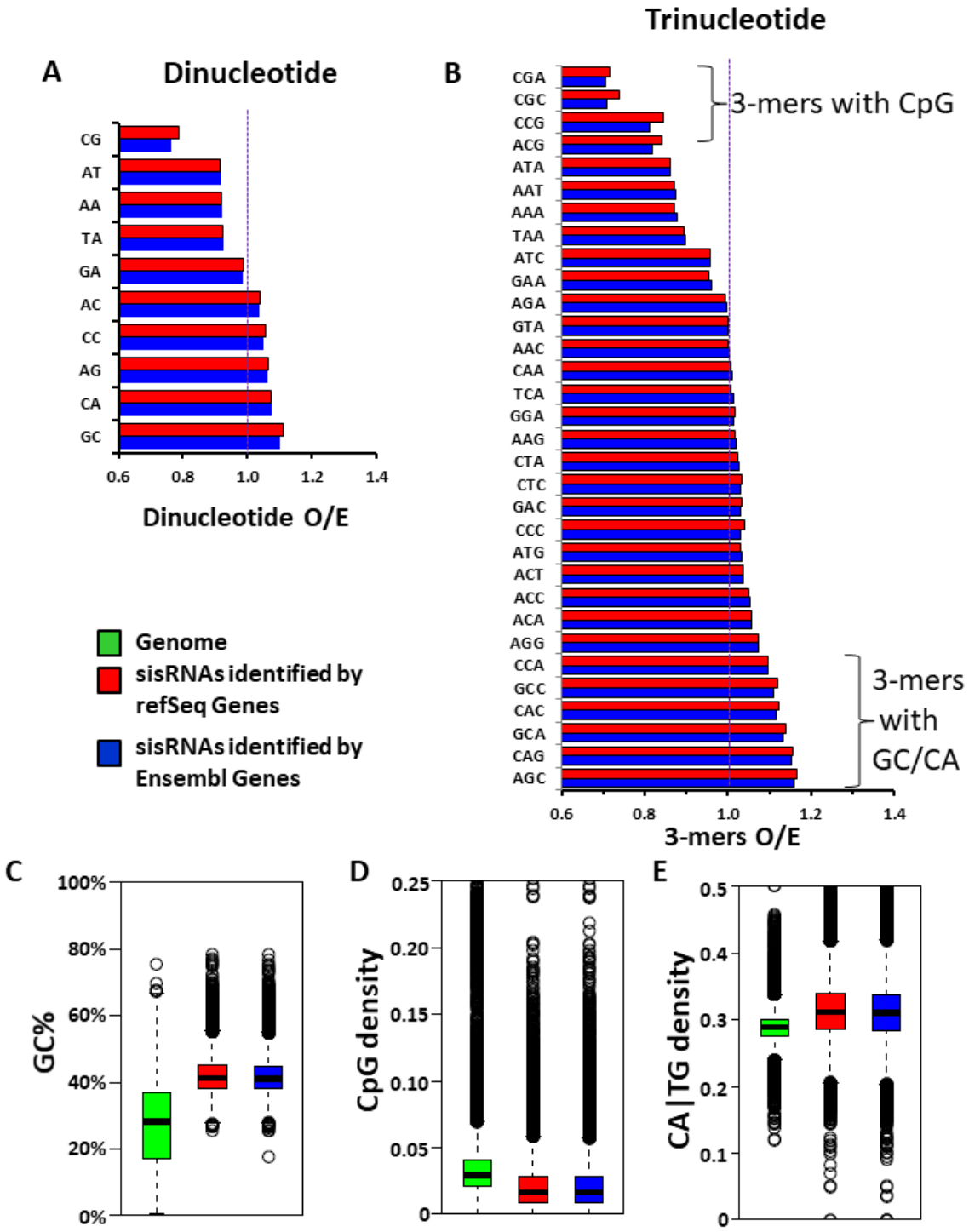

Figure 3

The sequences of sisRNAs have higher GC and TG content, while CpG poor compared to the genome. A-B. The ratio of observed/expected (O/E) for the occurrence of dinucleotide (A) and trinucleotide (B)combinations in sisRNAs identified by refSeq genes (red) and by Ensembl genes (blue). C-E. Boxplots

show the GC% (C), CpG density (D) and CA/TG density (E) of the genome (green) compared to sisRNAs identified by refSeq (red) and Ensembl (blue) genes.

Figure 4. sisRNAs are specific regions of the introns

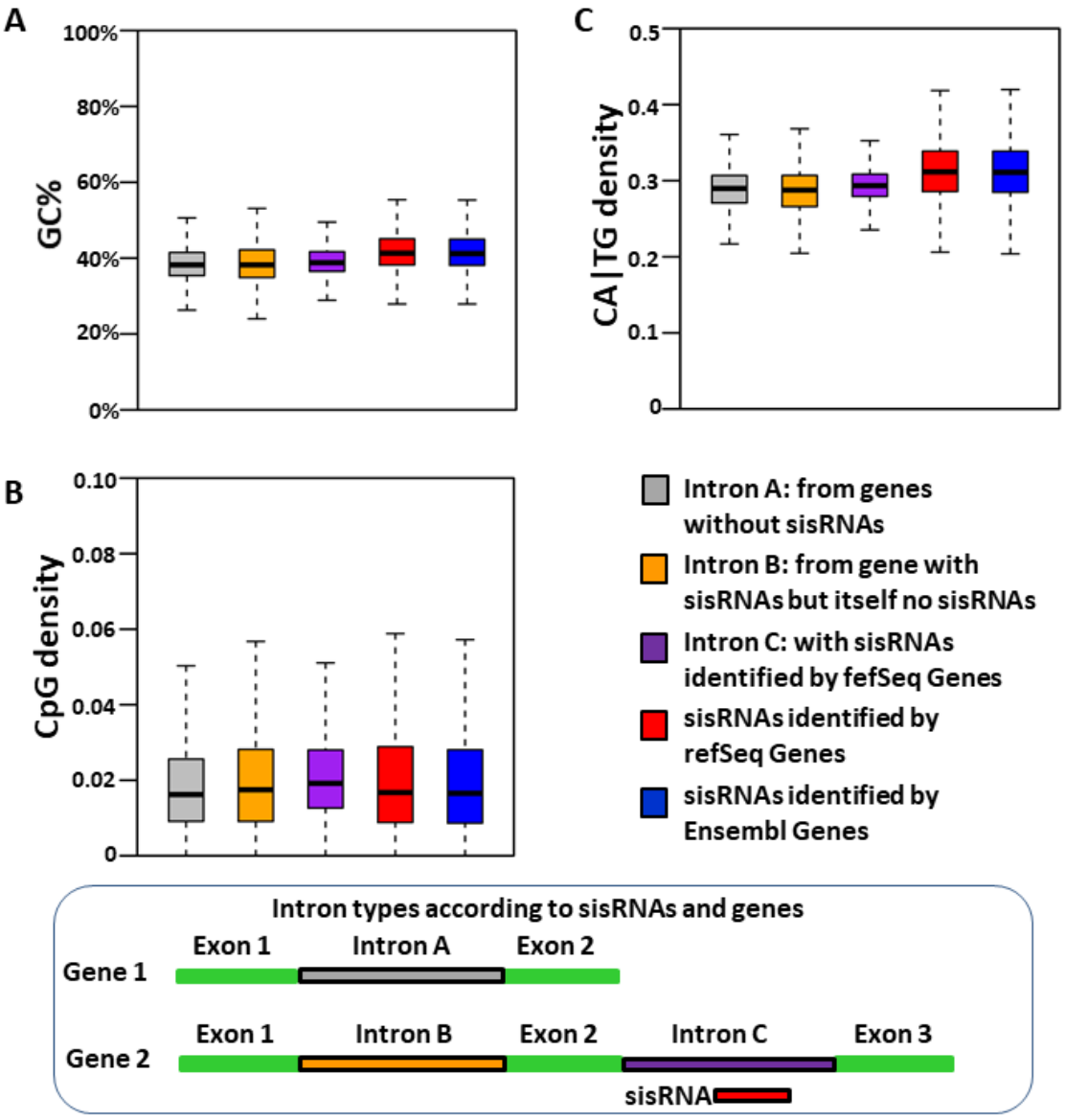

Figure 4

sisRNAs are specific regions of the introns. Boxplots of the comparisons of GC% (A), CpG density (B) and CA|TG density (C) of the introns from genes without any sisRNA (grey), introns without sisRNA from host

genes with sisRNAs (orange), and introns with sisRNAs (purple) to sisRNAs identified by refSeq (red) and Ensembl (blue) genes.

Figure 5. sisRNAs are enriched in introns at the 5’ and 3’ ends of transcripts

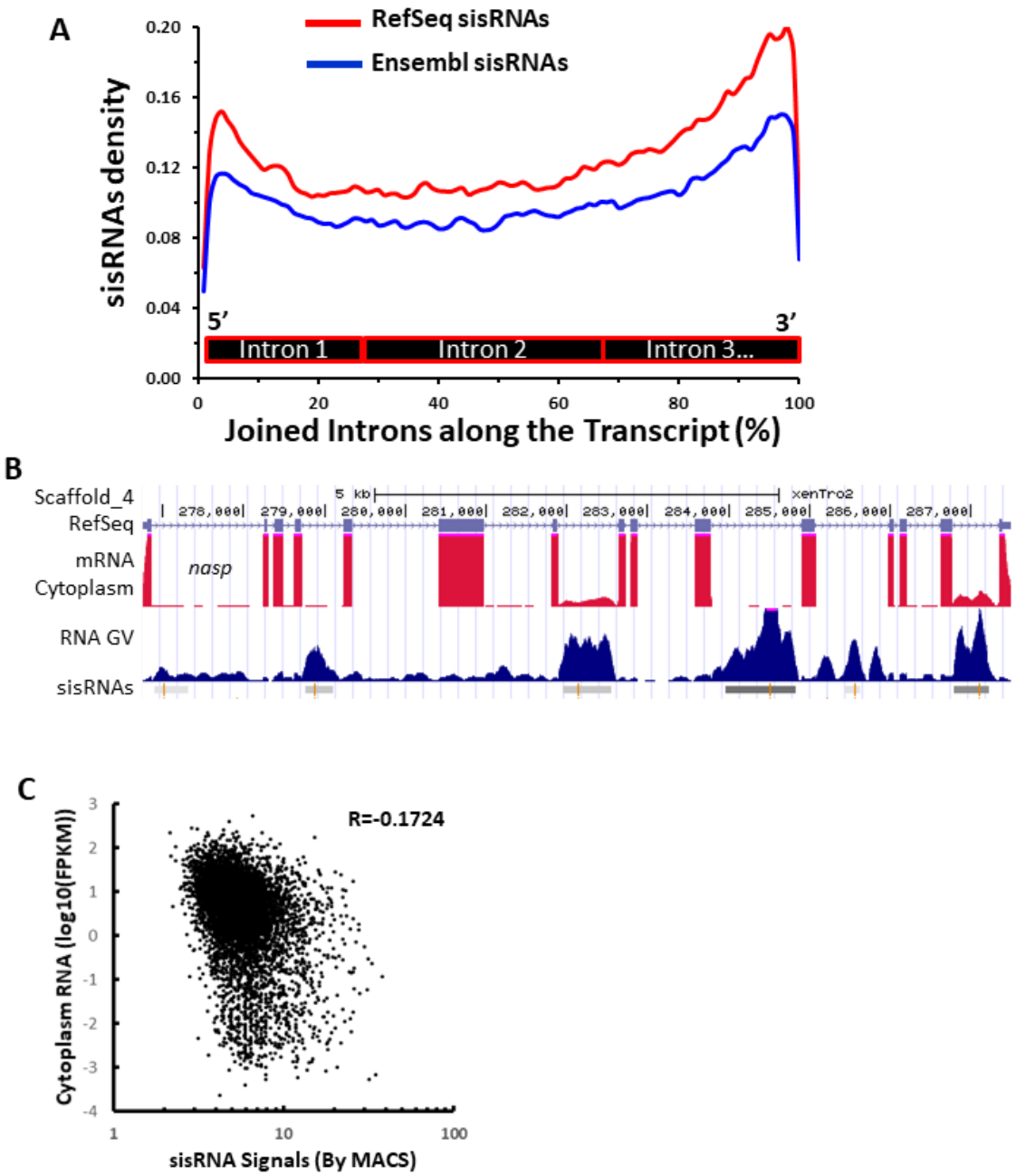

Figure 5

sisRNAs are enriched at both the start and the end of transcript, preferentially the 3’ end of transcript. A. The density of sisRNAs in introns. For each transcript, all the introns are concatenated from 5’ to 3’ end, and then divided into 100 bins. B. UCSC image of sisRNAs distribution along the introns of the gene nasp.

A higher number of sisRNAs are located at the 3’ end. C. Scatter plot of sisRNA peak signals versus host gene expression level (FPKM).

Figure 6. Specific transcription factor binding sites are enriched in sisRNAs

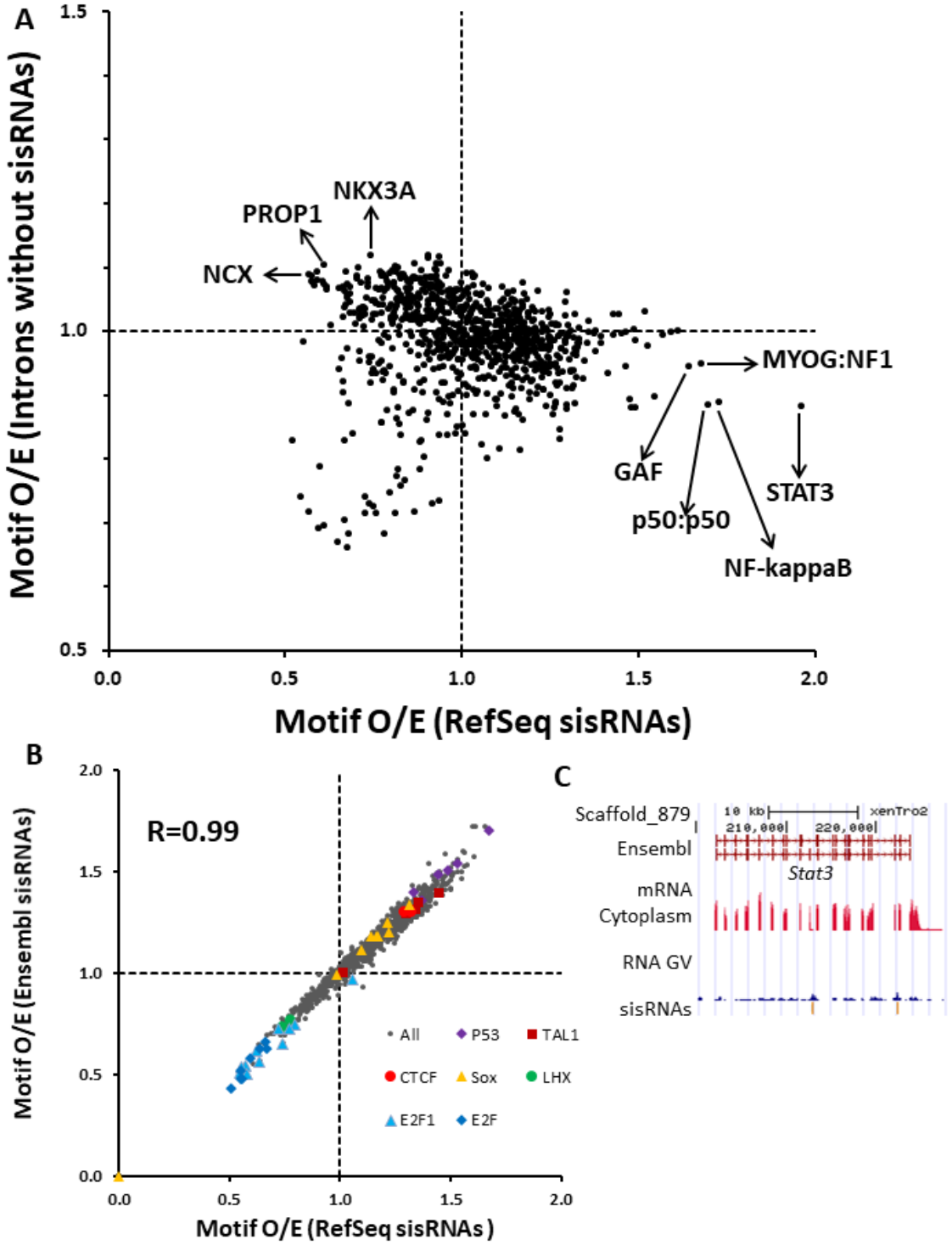

Figure 6

Specific TFBS are enriched in sisRNAs. The enrichment of TFBS motifs was plotted for (A) the introns without any sisRNAs versus the sisRNAs identified by RefSeq genes, and (B) the RefSeq sisRNAs versus the Ensembl sisRNAs. C. UCSC image of Stat3 shows it is highly expressed in the cytoplasm.

Figure 7. sisRNAs are evolutionary conserved as introns.

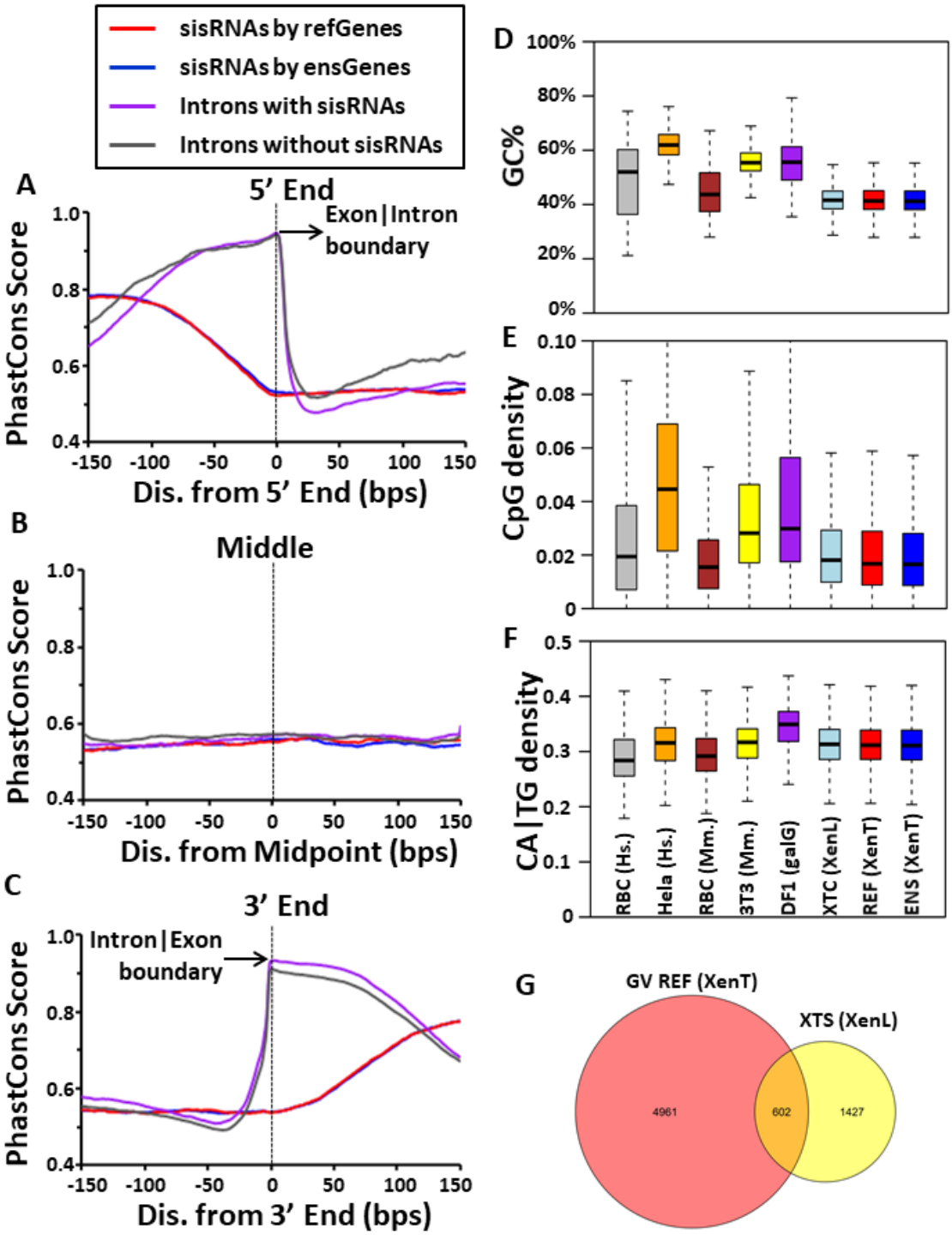

Figure 7

sisRNAs are as evolutionary conserved as introns, but much less than exons. Average PhastCons conservation scores of sisRNAs and introns for upstream and downstream (±150-bps) relative to (A) 5' end, (B) midpoint, and (C) 3' end. Boxplots of the comparisons of GC% (D), CpG density (E) and CA|TG density (F) of the human red blood cells (grey), human Hela cells (orange), mouse red blood cells (brown), mouse 3T3 cells (yellow), chicken DF1 cells (purple), and Xenopus laevis XTC cells (light blue)

cytoplasmic sisRNAs, to sisRNAs identified by refSeq (red) and Ensembl (blue) genes. (G) Venn diagram shows the overlap of host genes with sisRNAs identified by RefSeq genes in Xenopus tropicalis GV and host genes with cytoplasmic sisRNAs in Xenopus laevis XTC.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.pptx
- Tables.pptx
- Methodsformulas.docx