BOR Papers in Press. Published on September 14, 2016 as DOI:10.1095/biolreprod.116.142190

SpermBase – A database for sperm-borne RNA contents[1]

Running title: A collection of sperm RNAs in various species.

Andrew Schuster,[3,4] Chong Tang,[3,4] Yeming Xie,[4] Nicole Ortogero,[4] Shuiqiao Yuan,[4] and Wei Yan[2,4,5]
[4]Department of Physiology and Cell Biology, University of Nevada School of Medicine, Reno, Nevada
[5]Department of Biology, University of Nevada, Reno, Reno, Nevada

[1]This work was supported, in part, by grants from the NIH (HD060858, HD071736 and HD085506 to W. Y.) and the Templeton Foundation (PID: 50183 to WY). Sequencing was conducted in the Single Cell Genomics Core supported, in part, by a NIH grant (1P30GM110767). RNA-Seq datasets have been deposited into the NCBI GEO database with the accession number of GSE81216.
[2]Correspondence: Wei Yan, University of Nevada, Reno School of Medicine, Center for Molecular Medicine, Room 207B, 1664 North Virginia Street, MS/0575, Reno, NV 89557.
E-mail: wyan@medicine.nevada.edu
[3]These authors contributed equally to this work.

ABSTRACT
Since their discovery ~three decades ago, sperm-borne RNAs, both large/small and coding/noncoding, have been reported in multiple organisms, and some have been implicated in spermatogenesis, early development, and epigenetic inheritance. Despite these advances, isolation, quantification and annotation of sperm-borne RNAs remain nontrivial. The yields and subspecies of sperm-borne RNAs isolated from sperm can vary drastically depending on the methods used, and no cross-species analyses of sperm RNA contents have ever been conducted using a standardized sperm RNA isolation protocol. To address these issues, we developed a simple RNA isolation method that is applicable to sperm of various species, thus allowing for reliable interspecies comparisons. Based on RNA-Seq analyses, we established SpermBase (www.spermbase.org), a database dedicated to sperm-borne RNA profiling of multiple species. Currently, SpermBase contains large and small RNA expression data for mouse, rat, rabbit and human total sperm and sperm heads. By analyzing large and small RNAs for conserved features, we found that many sperm-borne RNA species were conserved across all four species analyzed, and among the conserved small RNAs, sperm-borne tsRNAs and miRNAs can target a large number of genes known to be critical for early development.

INTRODUCTION
Despite the cessation of transcription and shedding of the cytoplasm at the final stages of spermiogenesis, mature spermatozoa have been shown to possess diverse populations of both small and large RNAs [1-5]. These RNA populations, along with the paternal genome, can be delivered into the oocyte during fertilization, where they can persist beyond zygotic genome activation [5, 6]. Since their discovery in 1989, sperm-borne RNAs have been shown to function in spermatogenesis and during fertilization and early embryonic development, suggesting that they are not merely the remnants of sperm development [4, 7-10]. Sperm with aberrant miRNA contents display compromised fertilization rate and preimplantation embryonic development

when injected into wild type oocytes [5]. Additionally, sperm small RNAs have been found to play a role in epigenetic inheritance [11-17].

The study of sperm RNAs has been challenging, in part, due to the difficulties associated with sperm RNA isolation. Compared to other cell types, sperm contain fewer RNAs, and some RNAs appear to be localized to the nucleus and associated protamine-packed chromatin highly enriched in disulfide bonds resistant to lysis by detergents [4, 18, 19]. Therefore, it is no surprise that different RNA extraction procedures have been shown to lead to highly variable sperm-borne RNAs contents– this issue is compounded by the inherent heterogeneity of RNA populations among individual sperm samples [20-24]. Additionally, a sperm RNA isolation protocol that works for one species does not necessarily work for another, due to inter-species differences in sperm morphology and chromatin condensation [25-27]. The aforementioned issues surrounding the study of sperm-borne RNAs render the comparative analyses of data obtained using different experimental approaches unreliable. Therefore, it is necessary that a broadly applicable methodology be established for studying the sperm RNA contents across species. To this end, we developed a simple and effective protocol for sperm RNA isolation, which can be applied to multiple species with only minor modifications. Based on sperm RNA-Seq data, we established SpermBase (www.spermbase.org), a database dedicated to sperm RNA expression profiling for various species. Currently, SpermBase contains expression data of both large and small RNAs in sperm of four mammalian species (rat, mouse, human, and rabbit), with plans to expand to more species in the future. To demonstrate the utility of SpermBase, we compared the sperm RNA-Seq data and identified highly conserved mammalian sperm-borne RNAs among the four mammalian species.

MATERIALS AND METHODS

Sperm collection

The Institutional Animal Care and Use Committee (IACUC) of the University of Nevada, Reno approved the use of mice and rats (Protocol#00494), as well as rabbits (Protocol#00536) in this study. A summary of the sperm processing procedures is shown in Table 1.

Mouse (C57BL/6J from JAX) total sperm and sperm head collection were conducted as follows: Epididymides from a mouse were placed in 37°C HEPES-buffered human tubule fluid (HEPES-HTF) and dissected into smaller fragments. The sperm were allowed to escape from the epididymis during a 30 minute incubation at 37°C. The sperm-containing supernatants were collected and centrifuged for 5 minutes at 700xg. After removing the supernatants, fresh HEPES-HTF was added gently on top of the sperm pellet. During a 30 minute incubation at 37°C, sperm were allowed to "swim-up" from the pellet. The "swim-up" sperm were collected with the supernatant. After pelleting the sperm at 700xg for 5 minutes, the sperm were washed three times with 1XDPBS and centrifugation at the same speed. During the third wash, a small aliquot of sperm suspension was examined under a phase contrast microscope to determine the purity. The purity was >98% in all samples used for this study. For total sperm isolation, we performed the procedure individually for each mouse, and the total sperm of three mice were pooled for total RNA isolation. Two biological replicates (each with pooled total sperm RNA from three mice) were used for large or small RNA sequencing. For sperm head isolation, total sperm from 5 mice were sonicated (Diagenode Bioruptor UCD-200) in PBS for 3 minutes and subsequently pelleted at 700xg for 5 minutes. The sonicated sperm (consisting now of separated sperm heads and tails) was then added to a 4.5 mL 83.5% sucrose cushion and centrifuged at

100,000g for one hour (Beckman SW41Ti rotor); afterwards, the sperm head pellet was collected. Since sonication broke all contaminating somatic cells, the sperm head purity was close to 100%. We used sperm heads purified from 5 mice for RNA isolation and the subsequent sequencing with two biological replicates.

Rat (Sprague Dawley from Charles River) total sperm were isolated as follows: Rat epididymides were dissected and minced in F12 culture medium containing 0.1 % bovine serum albumin followed by a 30 min incubation at 37°C. The sperm-containing supernatants were collected and washed with PBS by centrifugation (800-1,000xg for 5min). Sperm pellets were re-suspended in 200µL NIM medium (121.6 mM KCl, 7.8 mM Na_2HPO_4, 1.4 mM KH_2PO_4, 0.1% poluvinyl alchol, and 10 mM EDTA), 100 µL collagenase (200U/ml, Sigma), and 100 µL hyaluronidase (100U/ml, Sigma) followed by an incubation at 37°C for 1h, with occasional mixing. Rat total sperm were then washed three times using 500µL NIM though centrifugation (4,000xg for 3 min) and the purity was >95%, as determined by phase contrast microscopic observation. The rat sperm pellets were snap frozen in liquid nitrogen followed by storage at -80°C until RNA isolation. For total sperm isolation, we performed the procedure individually for each rat, and the total sperm of three rats were pooled for total RNA isolation and sequencing. Three biological replicates (each with pooled total sperm RNA form three rats) were used for large or small RNA sequencing.

Rabbit (New Zealand White from Charles River) ejaculates were collected using the artificial vagina method, as described [28]. Rabbit sperm were washed three times with HEPES-HTF medium and the purity was >98% based on phase contrast microscopy. The rabbit sperm pellets were snap frozen in liquid nitrogen followed by storage at -80°C until RNA isolation. Three rabbits were used for collecting ejaculates and sperm from the three were individually processed for RNA isolation, library construction and sequencing.

Human sperm samples used in this study were de-identified human donor sperm samples purchased from California CryoBank Inc. Three donors were all healthy adult men (19-38 years old) who underwent rigorous health screening and met all the criteria for a qualified sperm donor, as described in the company's website (https://cryobank.com/uploadedFiles/Cryobankcom/_forms/pdf/brochures/DonorPyramid.pdf). The use of purchased, de-identified human sperm for RNA isolation required no IRB approval, as determined by the University of Nevada, Reno (documents available upon request). The cryopreserved human sperm were thawed by incubation in a waterbath at 37°C followed by three washes with HEPES-HTF and the purity of the human sperm was >0.01% (fewer than 1 round cell per 10,000 sperm) based on phase contrast microscopic observation. The sperm pellets were subjected to RNA isolation using the method described below. Three human sperm samples were processed and sequenced individually.

Sperm RNA isolation

Total RNA was isolated using the mirVana miRNA Isolation Kit (Life Technologies) following the manufacturer's instructions with modifications at the lysis step. The same procedures were used for total sperm and sperm heads of each species. Differences in the lysis step of the sperm RNA isolation procedures for the four species are summarized in Table 1. For mouse sperm, after the addition of the lysis buffer containing guanidine thiocyanate and β-mercaptoethanol (a reducing agent required for a complete lysis of the sperm heads), the frozen sperm pellets were homogenized using a homogenizer (BenchMark) at low settings for 1 minute on ice. For human and rabbit samples, after the addition of the lysis buffer, the frozen sperm pellets were pipetted

up and down using a hand-held pipette (Eppendorf) on ice, until the pellet dissolved. For rat samples, after the addition of the lysis buffer, the frozen sperm pellets were homogenized using a homogenizer (BenchMark) at low settings for 90 seconds, followed by a five minute incubation at 65°C. After lysis, a small aliquot (~5-10 µl) was placed on a glass slides followed by covering with a coverslip for microscopic examination to ensure sperm nuclei were completely lysed. The completely lysed sperm samples were subjected to the remaining default protocol for RNA isolation. For RNA quality control, sperm RNA samples were analyzed using the RNA 6000 Nano chips on an Agilent 2100 Bioanalyzer (Agilent). The RNA integrity number (RIN) was measured to estimate the sperm RNA quality.

RNA-Seq

Library preparation was performed using the Ion Total RNA-Seq Kit v2 (Life Technologies). Large RNA libraries were prepared using the whole transcriptome protocol provided with the Kit, while small RNA libraries were prepared using the small RNA protocol. Large and small RNA libraries were barcoded using Ion Xpress Barcode Adapters (Life Technologies). Quality control was performed using Agilent High Sensitivity chips (Agilent) and Experion DNA 1K kits (Bio-Rad). Libraries were loaded onto Ion PI chips via the Ion PI Template OT2 200 v3 (for small RNAs) or v2 (for large RNAs), and Ion PI Sequencing 200 v3 (for small RNAs) or v2 (for large RNAs) Kits, and sequenced on an Ion Proton Sequencer (Life Technologies).

Bioinformatics

RNA-Seq datasets have been deposited into the NCBI GEO database with the accession number of GSE81216. The large RNA-Seq data was annotated as follows: Reads were trimmed using fastx_trimmer and fastq_quality_trimmer (t = 30), and the resulting trimmed reads were mapped to the genome of each respective species (mouse, mm10; rat, rn5; human, hg19; rabbit, oryCun2) with TopHat (v2.09; default settings plus --b2-very-sensitive -r 200 and --mate-std-dev to 100) [29]. Illumina iGenome references (ENSEMBL) were used for rat, mouse, and human, and generated via TopHat for rabbit (ENSEMBL, release 76) [29, 30]. The aligned reads were then assembled using Cufflinks (v2.1.1; default settings plus --frag-bias-correct, --max-bundle-length 1e7, and --multi-read-correct) using the same genome reference versions and mask GTF files containing all known RNA for each species (ENSEMBL) [30]. Expressed genes in the data with the 'protein_coding' biotype (ENSEMBL) were extracted for further study and are housed on SpermBase. The percentage of exon coverage for each coding transcript identified in our RNA-Seq data, for each species and sample type, was calculated using Bedtools (bedtools coverage –s –a –b –sorted) and the results for each replicate were averaged [31]. GFF references (mm10, mouse; rn5, rat; oryCun2, rabbit; hg19, human) were obtained from UCSC [32].

The sncRNA-Seq data was annotated as follows: Reads <15nt and >50nt were discarded. The remaining reads were matched to known sncRNA, consisting of, when available for each species, mature miRNA (miRBase release 21), tRNA (Genomic tRNA Database; mm10, mouse; rn5, rat; oryCun2, rabbit; hg19, human), piRNA (piRNABank), rRNA (ENSEMBL, release 76), snoRNA (ENSEMBL, release 76), snRNA (ENSEMBL, release 76), and mitochondrial RNA (ENSEMBL, release 76), using Sequery (0 – 2 mismatches allowed) [30, 33-36]. Unmatched reads were matched to mouse testis [37] and sperm endo-siRNA [5] with Sequery (0 mismatches allowed). The remaining unmatched reads were aligned to the genome of each respective species (human, hg19; mouse, mm10; rat, rn5; rabbit, oryCun2) via Bowtie (settings -n 2 -k 3 --best -S --al -q) [38]. Aligned reads were matched to the genomic coordinates of known mature

miRNA, tRNA, rRNA, snRNA, snoRNA, and mitochondrial RNA, when available for each species (same databases used previously). Additional annotation of the rabbit sncRNA-Seq data was performed, due to the incomplete annotation of the rabbit sncRNA transcriptome. Unmatched, aligned reads 18 – 25 nt in length were matched to known rat, human, and mouse mature miRNA (miRBase release 21) via Sequery (0 – 2 mismatches allowed) [33, 36]. Unmatched, aligned reads 26 – 32 nt in length were matched to known rat, human, and mouse piRNA (piRNABank) via Sequery (0 – 2 mismatches allowed) [35, 36]. Remaining unmatched, aligned reads 26 – 32nt in length were analyzed by piRNA Predictor, to detect novel rabbit piRNA [39]. The remaining unmatched, aligned reads 26 – 32 nt in length were then matched to the novel rabbit piRNA via Sequery (0 – 2 mismatches) [36]. Read counts were obtained by in-house Python scripts. Reads were normalized as reads-per-hundred thousand aligned reads (RPK). Genes with fewer than 1 RPK were not included in the expression tables of SpermBase; when a gene had greater than 1 RPK in one sample type (e.g., total sperm), but not the other (e.g., sperm head), the expression of the gene was included for both sample types. The sncRNA-Seq reads that initially matched to tRNA genes were extracted and matched to the 5' and 3' halves of the full length tRNA (split at the 3' end of the anticodon), via Sequery (0 – 2 mismatches allowed) [36]. Reads ≥ 27 nt and ≤ 26 nt were named halves and tRFs, respectively, and further classified as 5' halves, tRF-5's, 3' halves, and tRF-3's based on whether the read aligned to the 5' or 3' half of the tRNA.

The conserved sperm miRNA gene targets were predicted using RNAhybrid (settings -n 50 -m 50000 -c -d xi,θ -p 0.05 -e -20) and miRanda (default settings; score ≥ 140 and energy ≤ -20) against the 3' UTR of their respective species (ENSEMBL, release 76), discarding predictions not made by both programs [30, 40-42]. Human miRNA were used to match against rabbit 3' UTR, due to the lack of confirmed mature miRNA sequences for the conserved miRNA identified. The 5' halve gene targets were predicted using RNAhybrid (settings -n 50 -m 50000 -c -d xi,θ -p 0.01 -e -20) against the 5' UTR, CDS, and 3' UTR sequences of their respective species' (ENSEMBL, release 76). The xi and θ values for each analysis were determined by RNAcalibrate (settings -n 50 -m 50000 -s), using randomly selected mature miRNA (miRBase, release 21) and known 5' halve sequences (tRFDB, human and mouse) (Supplemental Table S1; Supplemental Data are available online at www.biolreprod.org) [33, 40, 43].

Gene ontology term enrichment analysis was performed using the g:Profiler suite [44, 45]. Input gene lists were ordered by the number of miRNA or 5' halves that targeted each gene (highest to lowest), then by the average p-value (lowest to highest). The top 2000 and 4000 genes of the miRNA and 5' halve ordered lists, respectively, were analyzed in g:GOST (settings: significant only, ordered query, no sorting or hierarchical sorting - moderate, functional categories with 3 – 1000 terms only) [44, 45]. Word clouds were generated using Genes2WordCloud (http://www.maayanlab.net/G2W/) using the biological process terms found to be significantly enriched ($p \leq 0.05$) using the above settings (no sorting), and the aforementioned ordered lists (miRNA and 5' halve putative gene targets) or the total sperm coding genes (RPKM ≥ 3) found to be conserved across species [46].

The significance of the number of observed matches between the predicted conserved miRNA and 5' tRNA halve gene targets and genes present in early development (≥ 5 FPKM) was determined using a Chi-square test. Early development gene expression was obtained from the Database of Transcriptome in Mouse Early Embryos (DBTMEE) [47]. The observed number of matches were compared to the number of expected random matches, which was defined as the

number of early development genes for each stage multiplied by the percentage of all known mouse coding genes that were targeted by the conserved miRNA or 5' tRNA halves.

RESULTS

Complete lysis of sperm nuclei is key to maximal sperm RNA yields

Trizol, one of the most commonly used lysis solutions for RNA isolation, was used in several sperm RNA profiling studies [26, 27, 48-51]. However, Trizol is optimal for RNA isolation from somatic cells, but does not work efficiently for sperm RNA isolation, as revealed in our tests (Figure 1A). After incubation of the mouse sperm (at a concentration of 5,000-10,000 sperm/μl) at 37°C for 20 minutes, the number of sperm decreased by ~15-20%. In contrast, mouse sperm in the mirVana lysis buffer were almost completely dissolved after 20 minutes of incubation at room temperature (Figure 1A). This was not surprising given that Trizol solution contains neither detergents nor reducing agents, whereas the sperm chromatin is well known for the enriched disulfide bonds, which can only be broken down by both detergents (sodium dodecyl sulfate-SDS or sodium lauryl sulfate-SLS) and reducing agents (e.g., DTT or β-mercaptoethanol). Given that a large amount of sperm RNAs are embedded inside sperm nuclei [4, 18, 19], it is critical to lyse sperm heads completely so that sperm RNAs can be fully released. To this end, we attempted to establish a 'modular' RNA isolation protocol so that it is applicable to RNA isolation from sperm of any species. We adopted the mirVana Total RNA Isolation Kit because it uses a lysis buffer containing both 0.5% N-lauroyl sarcosine (a detergent) and 0.1M β-mercaptoethanol (a reducing agent), which are required for the complete lysis of the sperm heads. It should be noted that other similar kits with both detergents and reducing agents in their lysis solutions (e.g., the RNeasy Kit by Qiagen, as used previously [18]) can be as efficacious as the mirVana Kit for sperm RNA extraction. As outlined in Table 1, human and rabbit sperm heads could be easily lysed by pipetting up and down in the lysis buffer, whereas mouse sperm required mechanical homogenization and rat sperm even needed heating in addition to homogenization.

Electropherograms of the extracted sperm-borne RNAs were used to assess sperm RNA quality (Figure 1). A lack of intact 18S and 28S rRNAs in sperm renders the traditional RNA integrity number (RIN) conventions inappropriate for assessing sperm RNA quality [23, 52]. Instead of a RIN of ~10 being ideal for other cell types, we observed that a RIN of ~3 is indicative of high quality sperm RNA, consistent with a previous report [53]. Lower RINs (<2) typically represent excessive RNA degradation, while higher RINs (>4) may indicate somatic cell RNA contamination. We also observed that, in addition to a RIN of 2 – 4, sperm RNA electrophoretic profiles across species are characterized by a large population of RNA < 200nt in length (Figure 1), consistent with previous observation in mouse [50], human [52], horse [26], bull [54], and domestic swine [55].

Epididymal sperm collected from one adult male mouse routinely yielded ~350-500ng total RNA using our method (Table 1), which are sufficient for RNA-Seq analyse of both large and small RNAs. Similarly, a comparable amount of total RNA was obtained from sperm collected from one adult male rat, one ejaculate from one male rabbit or a man (Table 1). Consistent with previous reports showing sperm nuclei/heads contain higher concentration of RNAs [4, 7, 18, 19, 56], our data suggest that sperm heads contain ~60-80% of all sperm-borne RNAs at least in mice and rabbits (Table 1). Therefore, it is critical to completely lyse the sperm heads in order to obtain the maximal RNA yields. This may also explain why ~70-80 of mice

were used to obtain enough RNA for constructing one library for RNA-Seq in an earlier study using Trizol [50].

The number of small RNA sequencing reads in two biological replicates of the mouse and human total sperm as well as mouse sperm heads were analyzed, and excellent correlations between replicates (R^2 = 0.88-0.98) demonstrated a high degree of reproducibility of our sperm RNA isolation method (Figure 1C). By comparing our mouse total sperm small RNA dataset with that of a previous report [50] where Trizol was used for sperm RNA isolation, we found that the alignment rate (i.e., percentage of sequencing reads aligned to the genome) of our data is much greater than that of their data (~30% of ours vs. ~13% of theirs) (Supplemental Figure S1). Moreover, 24,303 unique small RNAs were identified from our dataset, whereas 18,915 were annotated from their dataset using exactly the same annotation pipeline (Supplemental Figure S1). Although 13,318 small RNAs were common in both datasets, our dataset yielded more unique small RNAs (10,985 vs. 5,597) compared to theirs (Supplemental Figure S1). These data further support that our sperm RNA isolation method can lead to better sequencing results, which allow for identification of more sperm-borne RNA species.

Sperm-borne mRNAs

RNA-Seq was performed using total RNA isolated from both sperm heads (for rabbit and mouse) and total sperm (for rat, rabbit, mouse and human) to determine the expression of coding genes. Previous investigations into the mRNA content of mammalian sperm have found diverse populations of coding genes within both mouse and human sperm [2, 57]. Similarly, we observed several thousand coding genes present in the total sperm and sperm head samples (Figures 2A and 2B). The sperm coding transcriptome was previously shown to consist of many fragmented mRNAs [2]. We investigated the transcript integrity in each of our samples and found that the degree of 'intactness' (i.e., the percentage of the mature transcript represented by sequencing reads in the data) varied considerably among the mRNA identified in our data (Supplemental Table S2). Next, we assessed the degree to which sperm-borne mRNAs are conserved across mammals. After removing any genes with expression levels below 3 RPKM from consideration, we identified 587 unique mRNAs present in the total sperm of all four species (Figure 2A, Supplemental Table S3A). Using the same cutoff, we observed 3,506 unique mRNAs that were commonly expressed in the mouse and rabbit sperm heads (Figure 2B, Supplemental Table S3B). Gene ontology (GO) term enrichment analyses were performed on the conserved total sperm mRNAs identified (Figure 2C, Supplemental Table S4). "Morphogenesis" and other development-related terms were the most prevalent amongst the significantly enriched biological process terms (Figure 2C, Supplemental Table S4A).

Sperm-borne small RNAs

In addition to mRNAs, we sequenced small noncoding RNA (sncRNA) in human, rat, rabbit, and mouse total sperm, as well as rabbit and mouse sperm heads. Consistent with previous reports, we observed diverse sncRNA populations in the sperm of each species surveyed [3, 50, 51, 58]. The most abundant classes of sncRNAs in sperm were miRNAs, tsRNAs, piRNAs, and mitosRNAs (Figure 3; Supplemental Table S5). Interestingly, significant differences were observed in RNA contents between total sperm and sperm heads (Figure 3). The tsRNAs were much more abundant in sperm heads than in total sperm, suggesting that tsRNAs are mainly localized to the sperm heads. Reduced proportions of mitosRNAs and miRNAs in sperm heads suggest that these two types of small RNAs, in addition to tsRNAs, are the major small RNA

constituents in other compartments of the total sperm. Similarly, in rabbit sperm heads, tsRNAs were the predominant small RNA species followed by mitosRNAs and miRNAs; mitosRNAs were much more abundant in the total sperm, compared to sperm heads, suggesting that mitosRNAs are mainly localized to the non-head regions of the total sperm. Considering their consistent abundance in both total sperm and sperm heads, we focused on four small RNA classes (piRNAs, mitosRNAs, miRNAs and tsRNAs) while investigating the sncRNA conservation in mammalian sperm.

piRNAs. Previous work has demonstrated that piRNAs are highly expressed in male germ cells, and play an essential role during spermatogenesis [59]. However, little is known about the purpose of their presence in mature sperm [1]. Unlike miRNAs, piRNAs are poorly conserved in their sequences across species [60]. Therefore, we investigated the general features of the piRNA populations in each species. In pachytene spermatocytes, there is a burst in piRNA expression – the piRNAs expressed in the meiotic phase of spermatogenesis are called 'pachytene piRNAs,' while the piRNAs expressed in spermatogonia are referred to as the 'pre-pachytene piRNAs' [59, 61]. Pre-pachytene piRNA are typically 26 – 28nt in length and have a strong preference towards uracil and adenine at their 1st and 10th nucleotide positions, respectively. In contrast, pachytene piRNAs are typically 30nt in length and only possess a preference for adenine at their 1st nucleotide [62]. Based on these characteristics, we assessed whether the piRNAs found in total sperm samples were pre-pachytene or pachytene in origin. The majority of piRNAs in total sperm samples were 29 – 32nt in length across species (Supplemental Figure S2A), and our analyses of nucleotide preferences showed a strong bias towards uracil for the 1st nucleotide (Supplemental Figure S2B) and only a slight preference towards adenine at the 10th nucleotide (Supplemental Figure S2C). These data suggest that the majority of the rabbit, human, rat, and mouse sperm-borne piRNAs are pachytene piRNAs.

mitosRNAs. The majority of the sncRNA-Seq reads aligned to mitochondrial genome were much shorter than their matching full-length transcripts. Many of the sequences aligned to a mitochondrial RNA were a consistent length, suggesting that they are not derived from the degradation of intact transcripts, but rather are mitochondrial genome-encoded small RNAs (mitosRNA) that were purposefully produced, as reported previously [63] (Supplemental Figures S3 - S7). The lack of a consensus sequence length for mitosRNAs has been observed previously [63]. We also found that sperm-borne mitosRNAs varied in length depending on their origin of the mitochondrial genes, and this feature was conserved across all four species surveyed (Supplemental Figures S4 - S7). mitosRNAs ranked second in terms of relative abundance among all four major sperm-borne small RNA species, i.e., tsRNAs > mitosRNAs > miRNAs > piRNAs (Figure 3).

miRNAs. Several groups have previously investigated the miRNA contents in sperm, for a variety of organisms such as mouse [50, 64], human [2, 4], bull [58, 65-67], and pig [68]. Despite this cache of sperm miRNA expression information, cross-species comparisons using this existing data would be, as discussed previously, intrinsically unreliable due to differences in methodology [20, 21, 23-27]. Using the standardized miRNA expression data available in SpermBase, we were able to identify 67 miRNAs that were present in the total sperm samples of all four species (Figure 4A; Supplemental Table S6A). These miRNAs accounted for the majority of all miRNAs expression in every species surveyed (Supplemental Table S6B). Many

of the conserved miRNAs were members of the same clusters (i.e., their genes were within 10kb of one another), for example, one of these clusters contains the miR-34b/c family, which, along with miR-449a/b/c, is known involved in the regulation of spermatogenesis and male fertility in mice [10]. Three (miR-34c-3p, miR-19b-3p, miR-148b-3p) of the seven miRNAs that were, in another study, found to be differentially expressed when comparing bulls of moderate and high fertility, were also members of the 67 conserved miRNAs [65].

To assess whether the conservation of these miRNAs across species held any functional significance, we predicted the gene targets for all 67 miRNAs *in silico* using both RNAhybrid and miRanda to compare the miRNAs to 3' UTR sequences, discarding the gene target predictions not made by both programs [40, 42]. The predicted gene targets of each conserved miRNA can be found in Supplemental Table S7. Several of these predicted targets, such as *Dnmt3a*, *Ezh2*, and *Gata4*, are known for their important role in early development (Figure 4B) [69-76]. Many miRNAs are functionally redundant, a phenomenon that is commonly observed in single miRNA knockout mouse models, the majority of which lack any aberrant phenotype [9, 77]. Because of this, we used genes that were targeted by at least two of the 67 conserved miRNAs for our subsequent studies. In order to gauge whether the putative functions of these miRNAs are conserved, for every species, we performed a gene ontology (GO) term enrichment analysis on the top 2,000 redundantly targeted genes, ranked by the number of matching conserved miRNAs, then by the average p-value of each predicted gene target. The results of the GO term enrichment analysis are summarized in Supplemental Table S8. Interestingly, 'morphogenesis' and development-related terms were the most prevalent ones for every species surveyed (Figure 4C and Supplemental Table S8). Likewise, of the top 20 biological process (BP) terms identified in our interspecies comparison, six were related to development (Supplemental Table S8G). These data suggest that these 67 miRNAs night have a regulatory role during early development, and this role is conserved across mammalian species.

To further assess whether the conserved sperm miRNAs are active during early development, we compared the list of redundantly targeted murine genes (i.e., targeted by at least two miRNAs) to genes known to be expressed in the very early stages of development (e.g., oocyte to four-cell). We found that the numbers of predicted gene targets that matched to early development genes were significantly higher than the expected number of random matches (Table 2), supporting the putative role of the 67 conserved sperm miRNAs as regulators of gene expression during early development.

tsRNAs. While annotating the sncRNA-Seq data for SpermBase, we observed that tRNAs accounted for more reads than any other sncRNAs, with the exception of human total sperm, in which mitosRNAs were more abundant (Figure 3; Supplemental Table S5). A closer look at these tRNA-aligned reads revealed that the majority of them were ~30nt, indicating that they actually represented tsRNAs and not intact mature tRNA species (Supplemental Figure S8). These tsRNAs have previously been identified in a myriad of organisms and cell types including sperm, where they were found to be highly abundant [51, 78]. Similar to piRNAs, tsRNAs are divided into further sub-groups based on their length as well as their origins. The tsRNA species of 27nt or longer were classified as 5' or 3' tRNA halves depending on whether they were derived from the 5' or 3' half of the mature tRNA (split at the anticodon), while those 19 – 26nt in length were classified as tRF-5's or tRF-3's depending on their half preference [79]. We found that 5' tRNA halves were the most abundant tsRNAs in every species and sample type surveyed, especially sperm heads, consistent with previous findings in mouse sperm (Figure 5A)

[51]. The predominance of 5' halves in sperm-borne small RNA populations, in addition to speculation that they may play an important role in epigenetic inheritance, led us to focus our attention on this particular class of tsRNAs [16, 17, 51, 80-82].

To assess the similarity of the 5' tRNA halves between the species on SpermBase, we sorted the total sperm 5' tRNA halves by the amino acid of their precursor tRNAs. In every organism, $tRNA^{Gly}$ species accounted for the most 5' tRNA halves, followed by $tRNA^{Glu}$, $tRNA^{Val}$, $tRNA^{Met}$, and $tRNA^{Lys}$ (Figure 5B). This finding is in agreement with previous studies on murine sperm-borne 5' tRNA halves, which found that 5' tRNA halves from $tRNA^{Gly}$ and $tRNA^{Glu}$ were the most abundant tsRNAs in sperm [51]. The observation that the majority of the 5' tRNA halves in each species surveyed are derived from the same group of tRNAs also indicates that the production and retention of these sncRNAs is conserved across mammalian sperm, similar to miRNAs (Figure 4A).

Several studies have found that 5' tRNA halves may possess the ability to act as post-transcriptional gene regulators, similar to miRNAs [83-86]. In order to evaluate whether the 5' tRNA halves that we found in the SpermBase data were capable of complementary sequence-based gene regulation, we performed unbiased gene target prediction analyses, matching the 5' halves against the 3' UTR, as well as the 5' UTR and CDS sequences available for each species. Instead of using both miRanda and RNAhybrid as we did for the conserved sperm miRNAs, we solely used RNAhybrid. This is because RNAhybrid accounts for the length of the target sequence, ensuring that the longer 3' UTR and CDS sequences would not receive a higher number of random matches than the 5' UTR [40]. The predicted gene targets for each species are provided in Supplemental Table S9. We observed that, on average, the 5' UTR (Figure 5C) was targeted the most, compared to the 3' UTR and CDS sequences, and that this was true for all four species (Figure 5D; Supplemental Table S9). The proportion of genes that were targeted by more than one 5' tRNA half was also consistently higher on average for the 5' UTR-based analyses (~88% targeted by at least two 5' tRNA halves) when compared to the CDS- and 3' UTR-based analyses (~68% and ~78%, respectively), suggesting that there were fewer random target predictions when the 5' UTR sequences were analyzed. Based on these findings, we utilized the 5' UTR-based putative gene targets for our subsequent analyses.

From the 5' UTR-based gene target predictions, we selected the top 4,000 redundantly targeted genes, which were ranked according to the number of 5' tRNA halves that targeted each gene, then by the average p-value for each prediction, to perform a GO term enrichment analyses for each species. The results are summarized in Supplemental Table S10. Terms relevant to development and morphogenesis were common in the results for every species, similar to what was observed for the conserved sperm miRNAs (Figures 4C and 5E, Supplemental Tables S7G and S9G). Of the top 20 BP terms for the 5' tRNA halves, three were directly related to development (Supplemental Table S10G). Interestingly, the (5' tRNA halves) BP term ranked second on Supplemental Table S10G was "WNT signaling pathway," a pathway vital for proper early embryonic development across the animal kingdom [87]. Additionally, 'catabolic' and 'metabolic' were prevalent in the enriched terms (Figure 5E). This is not surprising, as sperm-borne 5' tRNA halves have previously been linked to altered expression of metabolic genes in the early embryo [16]. Overall, it seems that, like the conserved sperm miRNAs, sperm-borne 5' tRNA halves might also play a role in regulating gene expression during early development. To investigate this further, we compared the redundantly targeted murine genes (i.e., targeted by at least two 5' tRNA halves) identified in the 5' UTR-based analyses to genes known to be expressed throughout early development. As also seen with the putative miRNA targets, the

number of predicted targets for 5' tRNA halves that matched these early development genes was significantly higher than the number of anticipated random matches, providing additional evidence that 5' tRNA halves may play a functional role after fertilization (Table 2). The 5' tRNA halves were predicted to target a much larger percentage of the early development genes (~88%) relative to the miRNAs (~25%) (Table 2). This observation is consistent with another recent study, which determined that sperm-borne 5' tRNA halves targeted ~80% of the 8-cell embryo transcriptome [16].

SpermBase is easy to use and will be expanded to cover more species

To share sperm RNA profiling data, we established SpermBase, a publically accessible database, and can be found at www.spermbase.org. The website is separated into five main pages – Home, Search, Method, Species, and FAQ & Contact. The homepage (Supplemental Figure S9) provides an introduction to the database itself, touching on the topics discussed above in the introduction of this article. Links can be found throughout the introduction to other parts of the SpermBase website.

At the Search page, users can search SpermBase for expression data for their gene(s) of interest. Each RNA that was found in our data is categorized by its original or, if it was identified by the authors, its given name ('Gene'). As discussed below, the reads aligned to tRNA genes were classified as different subclasses (e.g., 5' tRNA halves, tRF-5) of tRNA-derived sncRNAs (tsRNAs). The naming convention for these tsRNAs was to add the subclass of tsRNAs after the original tRNA name, e.g., the 5' halves of "trna1000-PheGAA" would be named "trna1000-PheGAA-5halves." Each RNA is also organized by the species it was identified in ('Species'), the 'Sample' type (i.e., total sperm or sperm head), the class of large or small RNA, e.g., mRNA, piRNA ('Class'), and its expression ('Expression'). The expression values shown in SpermBase are the normalized read counts, given as RPK (reads per hundred thousand mapped reads) for sncRNA genes and RPKM (reads per kilobase of transcript per million mapped reads) for mRNAs. The sequences (5' ⟶ 3') of the tsRNAs and novel rabbit piRNAs we identified in this study are also provided under the 'Sequence' column.

The Method page describes the methodology (discussed previously) used for sperm RNA isolation followed RNA-Seq analyses for building SpermBase. As SpermBase expands to include additional species, we will update this page with the lysis stage parameters used for the new species. Users who employ our modular RNA isolation protocol on sperm from animals not described in SpermBase are encouraged to send us information on the lysis parameters that they found to be most optimal for that species.

On the Species page, a list of each species currently available on SpermBase is provided, along with planned future additions to SpermBase. Users can obtain the expression data for each individual species on the Download page; this data is available in multiple formats (.txt, .csv, and Excel .xls). The files can be downloaded by either left clicking the file or right clicking to "Save link as…" if the user wishes to rename the file prior to downloading.

The FAQ & Contact page provides answers to frequently asked questions related to SpermBase and detailed contact information. Users who have questions not addressed in the FAQ are encouraged to email SpermBase. Troubleshooting queries and comments about the design and functionality of SpermBase are also welcome.

DISCUSSION

Currently, SpermBase contains sperm RNA-Seq data of four mammalian species (human,

mouse, rat, and rabbit). To ensure the comparability of multi-species sperm RNA data in SpermBase, we conducted RNA-Seq on sperm RNA samples isolated using our standard sperm RNA isolation protocol. The key to high yield of high quality sperm RNAs is to dissolve the sperm heads completely, which requires reducing agents, mechanical disruptions, and increased temperature during lysis, depending on the extent to which the sperm chromatins are compacted. In general, the more compact the sperm heads are, the more vigorous treatment is needed in the lysis step. In our protocol, the rat sperm requires longer homogenization (90 seconds) plus heating (65°C for 5min) for a complete lysis of the sperm chromatin, whereas human and rabbit sperm only need up and down pipetting. Interestingly, rat sperm appear to be much more compact than human and rabbit sperm.

Another interesting observation is the difference in RNA contents between total sperm and sperm heads. The fact that RNAs in sperm heads account for 60-80% of the total sperm-borne RNA contents suggests that numerous RNAs are localized to the sperm nucleus and may be associated with sperm chromatin, as suggested in previous reports [51]. Because of the considerable difference in RNA contents between sperm heads and total sperm, it is critical to verify that all sperm heads are completely lysed through microscopic observation after the lysis step. It is our hope that, in addition to using the data on SpermBase, other groups will utilize our RNA extraction method and quality control metric (i.e., RIN of 2 – 4 as an indicator of ideal sperm RNA quality) for their own sperm RNA profiling studies. While we are confident in the efficacy of our method, we do caution other groups to make sure that they initially test, and if necessary, optimize the lysis step for their own samples, to ensure the complete lysis of the entire sperm cells. In the future, we hope to develop an RNA extraction method that can be applied to sperm from any species, without making any alterations to the protocol. Other future improvements to SpermBase will include the addition of more sperm head fraction data, as well as the inclusion of other species, expanding the utility of SpermBase for the scientific community. Currently, we are working on adding large and small RNA expression data for the sperm of zebrafish, horse, monkey, and bull.

Using SpermBase data, we investigated the conserved general features of mammalian sperm mRNA and small RNA populations. We limited our analyses of the conservation of sperm-borne small RNAs to four classes – piRNAs, mitosRNAs, miRNAs, and tsRNAs, because they represent the most abundant small RNA species in both total sperm and sperm heads. The piRNA populations in the total sperm of all four mammals appeared to be pachytene in origin. As the sequences of individual piRNAs are poorly conserved across species, it is difficult at this time to say whether the piRNA present in sperm serve some biological purpose or are just the random remnants of the pachytene piRNA expression burst [59]. In every species (with the exception of human), mitosRNAs appear to be the second most abundant small RNA species in both total sperm and sperm heads; however, their physiological role remains unknown. A total of 67 conserved miRNAs account for the majority of all miRNAs expressed in each of the four species examined. The highly conserved sperm-borne miRNAs may suggest a potential role in fertilization and early development.

Our analyses indicate that tsRNAs are the most abundant sperm small RNAs in every species (with the exception of human), with 5’ tRNA halves as the most dominant subclass, consistent with a previous report [51]. In every species surveyed, the bulk of these 5’ tRNA halves originated from the same precursor mature tRNAs, namely $tRNA^{Gly}$, $tRNA^{Glu}$, $tRNA^{Val}$, $tRNA^{Met}$, and $tRNA^{Lys}$. Unlike miRNAs, which preferentially bind to the 3’ UTR of their target genes [88], sperm-borne 5’ tRNA halves preferentially target the 5’ UTR. Interestingly, the

genes that are targeted by these 5' tRNA halves are mostly related to early development, suggesting the sperm-borne tsRNAs may also have a role in regulating early development. Previous studies have demonstrated that 5' tRNA halves have the potential to act as post-transcriptional regulators by binding to the 5' UTR in a reverse complimentary manner in cell lines [83-86]. In another study, 5' tRNA halves inhibited translation *via* the displacement of eIF4G/eIF4A from the 5' ends of mRNA [84]. In a recent report, Chen *et al* determined that putative 5' tRNA halves target sites in the CDS were less frequent than the number observed in promoter regions (i.e., 2kb upstream of the transcriptional start site) [16]. The data housed at SpermBase will undoubtedly be a boon to future investigations into the mechanism behind the proposed regulatory functions of sperm-borne 5' halves.

Sperm-borne RNAs have been implicated as potential mediators of epigenetic inheritance [11-17, 89]. Of particular interest are the sperm-borne tsRNAs, as a methylatransferase involved with tsRNA production, Dnmt2, was found to be necessary for the transmission of two epigenetically inherited phenotypes [81]. While the effects of *Dnmt2* deficiency on the tsRNA populations in sperm have yet to be determined, in other cell types, the production of 5' halve species was found to increase in the absence of Dnmt2 [90]. It is therefore possible that changes in sperm-borne 5' halve levels can have an effect on the early embryo; this is in agreement with the results of our analyses of the predicted gene targets of the 5' tRNA halves, suggesting that 5' tRNA halves interact with genes related to development. Recently, in three separate epigenetic inheritance models, expression of sperm-borne 5' tRNA halves was observed to be affected by an altered diet (i.e., high fat or low protein) or vinclozolin (a common agricultural fungicide) exposure [16, 17, 82]. Future investigations into other examples of epigenetic inheritance could potentially benefit from the tsRNA expression data and consensus sequences housed in SpermBase.

Since their discovery, research on sperm-borne RNAs has revealed that they are not simply the remnants of spermatogenesis – to the contrary, the likelihood that sperm RNAs play a functional role, in many different contexts (e.g., post-transcriptional regulation during early development and epigenetic inheritance), has only increased in recent years. The difficulties associated with sperm RNA isolation and sperm RNA contents in various species have no doubt slowed the advance of efforts to elucidate additional functions and to assess the conservation of these RNAs across evolution. These issues will be alleviated, at least partially, by SpermBase, which provides public expression data as well as a simple and effective RNA extraction methodology, thereby making the study of sperm-borne RNA more accessible to other labs. We hope that, with this contribution, we will see an acceleration of our understanding of the physiological roles of sperm-borne RNAs.

REFERENCES

1. Casas E, Vavouri T. Sperm epigenomics: challenges and opportunities. Front Genet 2014; 5:330.
2. Sendler E, Johnson GD, Mao S, Goodrich RJ, Diamond MP, Hauser R, Krawetz SA. Stability, delivery and functions of human sperm RNAs at fertilization. Nucleic Acids Res 2013; 41:4104-4117.
3. Krawetz SA, Kruger A, Lalancette C, Tagett R, Anton E, Draghici S, Diamond MP. A survey of small RNAs in human sperm. Hum Reprod 2011; 26:3401-3412.
4. Jodar M, Selvaraju S, Sendler E, Diamond MP, Krawetz SA. The presence, role and clinical use of spermatozoal RNAs. Hum Reprod Update 2013; 19:604-624.

5. Yuan S, Schuster A, Tang C, Yu T, Ortogero N, Bao J, Zheng H, Yan W. Sperm-borne miRNAs and endo-siRNAs are important for fertilization and preimplantation embryonic development. Development 2016; 143:635-647.
6. Ostermeier GC, Miller D, Huntriss JD, Diamond MP, Krawetz SA. Reproductive biology: delivering spermatozoan RNA to the oocyte. Nature 2004; 429:154.
7. Pessot CA, Brito M, Figueroa J, Concha, II, Yanez A, Burzio LO. Presence of RNA in the sperm nucleus. Biochem Biophys Res Commun 1989; 158:272-278.
8. Boerke A, Dieleman SJ, Gadella BM. A possible role for sperm RNA in early embryo development. Theriogenology 2007; 68 Suppl 1:S147-155.
9. Wu J, Bao J, Kim M, Yuan S, Tang C, Zheng H, Mastick GS, Xu C, Yan W. Two miRNA clusters, miR-34b/c and miR-449, are essential for normal brain development, motile ciliogenesis, and spermatogenesis. Proc Natl Acad Sci U S A 2014; 111:E2851-2857.
10. Yuan S, Tang C, Zhang Y, Wu J, Bao J, Zheng H, Xu C, Yan W. mir-34b/c and mir-449a/b/c are required for spermatogenesis, but not for the first cleavage division in mice. Biol Open 2015; 4:212-223.
11. Wagner KD, Wagner N, Ghanbarian H, Grandjean V, Gounon P, Cuzin F, Rassoulzadegan M. RNA induction and inheritance of epigenetic cardiac hypertrophy in the mouse. Dev Cell 2008; 14:962-969.
12. Rassoulzadegan M, Grandjean V, Gounon P, Vincent S, Gillot I, Cuzin F. RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. Nature 2006; 441:469-474.
13. Yuan S, Oliver D, Schuster A, Zheng H, Yan W. Breeding scheme and maternal small RNAs affect the efficiency of transgenerational inheritance of a paramutation in mice. Sci Rep 2015; 5:9266.
14. Gapp K, Jawaid A, Sarkies P, Bohacek J, Pelczar P, Prados J, Farinelli L, Miska E, Mansuy IM. Implication of sperm RNAs in transgenerational inheritance of the effects of early trauma in mice. Nat Neurosci 2014; 17:667-669.
15. Rodgers AB, Morgan CP, Leu NA, Bale TL. Transgenerational epigenetic programming via sperm microRNA recapitulates effects of paternal stress. Proc Natl Acad Sci U S A 2015; 112:13699-13704.
16. Chen Q, Yan M, Cao Z, Li X, Zhang Y, Shi J, Feng GH, Peng H, Zhang X, Zhang Y, Qian J, Duan E, et al. Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. Science 2016; 351:397-400.
17. Sharma U, Conine CC, Shea JM, Boskovic A, Derr AG, Bing XY, Belleannee C, Kucukural A, Serra RW, Sun F, Song L, Carone BR, et al. Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. Science 2016; 351:391-396.
18. Goodrich RJ, Anton E, Krawetz SA. Isolating mRNA and small noncoding RNAs from human sperm. Methods Mol Biol 2013; 927:385-396.
19. Goodrich R, Johnson G, Krawetz SA. The preparation of human spermatozoal RNA for clinical analysis. Arch Androl 2007; 53:161-167.
20. Mao S, Goodrich RJ, Hauser R, Schrader SM, Chen Z, Krawetz SA. Evaluation of the effectiveness of semen storage and sperm purification methods for spermatozoa transcript profiling. Syst Biol Reprod Med 2013; 59:287-295.
21. Mao S, Sendler E, Goodrich RJ, Hauser R, Krawetz SA. A comparison of sperm RNA-seq methods. Syst Biol Reprod Med 2014; 60:308-315.

22. Barragan M, Martinez A, Llonch S, Pujol A, Vernaeve V, Vassena R. Effect of ribonucleic acid (RNA) isolation methods on putative reference genes messenger RNA abundance in human spermatozoa. Andrology 2015; 3:797-804.
23. Cappallo-Obermann H, Schulze W, Jastrow H, Baukloh V, Spiess AN. Highly purified spermatozoal RNA obtained by a novel method indicates an unusual 28S/18S rRNA ratio and suggests impaired ribosome assembly. Mol Hum Reprod 2011; 17:669-678.
24. Lalancette C, Platts AE, Johnson GD, Emery BR, Carrell DT, Krawetz SA. Identification of human sperm transcripts as candidate markers of male fertility. J Mol Med (Berl) 2009; 87:735-748.
25. Johnson V. From a sperm's eye view - revisiting our perception of this intriguing cell. In: Proc. 53rd Annual Conv. Am. Assoc. Equine. Practitioners. Orlando, FL, USA; 2007.
26. Das PJ, Paria N, Gustafson-Seabury A, Vishnoi M, Chaki SP, Love CC, Varner DD, Chowdhary BP, Raudsepp T. Total RNA isolation from stallion sperm and testis biopsies. Theriogenology 2010; 74:1099-1106, 1106e1091-1092.
27. Shafeeque CM, Singh RP, Sharma SK, Mohan J, Sastry KV, Kolluri G, Saxena VK, Tyagi JS, Kataria JM, Azeez PA. Development of a new method for sperm RNA purification in the chicken. Anim Reprod Sci 2014; 149:259-265.
28. Bredderman PJ, Foote RH, Yassen AM. AN IMPROVED ARTIFICIAL VAGINA FOR COLLECTING RABBIT SEMEN. J Reprod Fertil 1964; 7:401-403.
29. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009; 25:1105-1111.
30. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, et al. Ensembl 2014. Nucleic Acids Res 2014; 42:D749-755.
31. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010; 26:841-842.
32. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 2004; 32:D493-496.
33. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 2014; 42:D68-73.
34. Chan PP, Lowe TM. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res 2009; 37:D93-97.
35. Sai Lakshmi S, Agrawal S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Res 2008; 36:D173-177.
36. Ortogero N, Hennig GW, Langille C, Ro S, McCarrey JR, Yan W. Computer-assisted annotation of murine Sertoli cell small RNA transcriptome. Biol Reprod 2013; 88:3.
37. Song R, Hennig GW, Wu Q, Jose C, Zheng H, Yan W. Male germ cells express abundant endogenous siRNAs. Proc Natl Acad Sci U S A 2011; 108:13159-13164.
38. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009; 10:R25.
39. Zhang Y, Wang X, Kang L. A k-mer scheme to predict piRNAs and characterize locust piRNAs. Bioinformatics 2011; 27:771-776.
40. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. Rna 2004; 10:1507-1517.
41. Kruger J, Rehmsmeier M. RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res 2006; 34:W451-454.

42. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. PLoS Biol 2004; 2:e363.
43. Kumar P, Mudunuri SB, Anaya J, Dutta A. tRFdb: a database for transfer RNA fragments. Nucleic Acids Res 2015; 43:D141-145.
44. Reimand J, Kull M, Peterson H, Hansen J, Vilo J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res 2007; 35:W193-200.
45. Reimand J, Arak T, Vilo J. g:Profiler--a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res 2011; 39:W307-315.
46. Baroukh C, Jenkins SL, Dannenfelser R, Ma'ayan A. Genes2WordCloud: a quick way to identify biological themes from gene lists and free text. Source Code Biol Med 2011; 6:15.
47. Park SJ, Shirahige K, Ohsugi M, Nakai K. DBTMEE: a database of transcriptome in mouse early embryos. Nucleic Acids Res 2015; 43:D771-776.
48. Parthipan S, Selvaraju S, Somashekar L, Kolte AP, Arangasamy A, Ravindra JP. Spermatozoa input concentrations and RNA isolation methods on RNA yield and quality in bull (Bos taurus). Anal Biochem 2015; 482:32-39.
49. Bissonnette N, Levesque-Sergerie JP, Thibault C, Boissonneault G. Spermatozoal transcriptome profiling for bull sperm motility: a potential tool to evaluate semen quality. Reproduction 2009; 138:65-80.
50. Kawano M, Kawaji H, Grandjean V, Kiani J, Rassoulzadegan M. Novel small noncoding RNAs in mouse spermatozoa, zygotes and early embryos. PLoS One 2012; 7:e44542.
51. Peng H, Shi J, Zhang Y, Zhang H, Liao S, Li W, Lei L, Han C, Ning L, Cao Y, Zhou Q, Chen Q, et al. A novel class of tRNA-derived small RNAs extremely enriched in mature mouse sperm. Cell Res 2012; 22:1609-1612.
52. Johnson GD, Sendler E, Lalancette C, Hauser R, Diamond MP, Krawetz SA. Cleavage of rRNA ensures translational cessation in sperm at fertilization. Mol Hum Reprod 2011; 17:721-726.
53. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol 2006; 7:3.
54. Gilbert I, Bissonnette N, Boissonneault G, Vallee M, Robert C. A molecular analysis of the population of mRNA in bovine spermatozoa. Reproduction 2007; 133:1073-1086.
55. Yang CC, Lin YS, Hsu CC, Wu SC, Lin EC, Cheng WT. Identification and sequencing of remnant messenger RNAs found in domestic swine (Sus scrofa) fresh ejaculated spermatozoa. Anim Reprod Sci 2009; 113:143-155.
56. Johnson GD, Lalancette C, Linnemann AK, Leduc F, Boissonneault G, Krawetz SA. The sperm nucleus: chromatin, RNA, and the nuclear matrix. Reproduction 2011; 141:21-36.
57. Fang P, Zeng P, Wang Z, Liu M, Xu W, Dai J, Zhao X, Zhang D, Liang D, Chen X, Shi S, Zhang M, et al. Estimated diversity of messenger RNAs in each murine spermatozoa and their potential function during early zygotic development. Biol Reprod 2014; 90:94.
58. Stowe HM, Calcatera SM, Dimmick MA, Andrae JG, Duckett SK, Pratt SL. The bull sperm microRNAome and the effect of fescue toxicosis on sperm microRNA expression. PLoS One 2014; 9:e113163.
59. Gou LT, Dai P, Yang JH, Xue Y, Hu YP, Zhou Y, Kang JY, Wang X, Li H, Hua MM, Zhao S, Hu SD, et al. Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. Cell Res 2014; 24:680-700.

60. Mani SR, Juliano CE. Untangling the web: the diverse functions of the PIWI/piRNA pathway. Mol Reprod Dev 2013; 80:632-664.
61. Meikar O, Da Ros M, Korhonen H, Kotaja N. Chromatoid body and small RNAs in male germ cells. Reproduction 2011; 142:195-209.
62. Ortogero N, Schuster AS, Oliver DK, Riordan CR, Hong AS, Hennig GW, Luong D, Bao J, Bhetwal BP, Ro S, McCarrey JR, Yan W. A novel class of somatic small RNAs similar to germ cell pachytene PIWI-interacting small RNAs. J Biol Chem 2014; 289:32824-32834.
63. Ro S, Ma HY, Park C, Ortogero N, Song R, Hennig GW, Zheng H, Lin YM, Moro L, Hsieh JT, Yan W. The mitochondrial genome encodes abundant small noncoding RNAs. Cell Res 2013; 23:759-774.
64. Garcia-Lopez J, Alonso L, Cardenas DB, Artaza-Alvarez H, Hourcade Jde D, Martinez S, Brieno-Enriquez MA, Del Mazo J. Diversity and functional convergence of small noncoding RNAs in male germ cell differentiation and fertilization. Rna 2015; 21:946-962.
65. Fagerlind M, Stalhammar H, Olsson B, Klinga-Levan K. Expression of miRNAs in Bull Spermatozoa Correlates with Fertility Rates. Reprod Domest Anim 2015; 50:587-594.
66. Du Y, Wang X, Wang B, Chen W, He R, Zhang L, Xing X, Su J, Wang Y, Zhang Y. Deep sequencing analysis of microRNAs in bovine sperm. Mol Reprod Dev 2014; 81:1042-1052.
67. Govindaraju A, Uzun A, Robertson L, Atli MO, Kaya A, Topper E, Crate EA, Padbury J, Perkins A, Memili E. Dynamics of microRNAs in bull spermatozoa. Reprod Biol Endocrinol 2012; 10:82.
68. Curry E, Safranski TJ, Pratt SL. Differential expression of porcine sperm microRNAs and their association with sperm morphology and motility. Theriogenology 2011; 76:1532-1539.
69. Molkentin JD, Lin Q, Duncan SA, Olson EN. Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis. Genes Dev 1997; 11:1061-1072.
70. Hu YC, Okumura LM, Page DC. Gata4 is required for formation of the genital ridge in mice. PLoS Genet 2013; 9:e1003629.
71. Pilon N, Raiwet D, Viger RS, Silversides DW. Novel pre- and post-gastrulation expression of Gata4 within cells of the inner cell mass and migratory neural crest cells. Dev Dyn 2008; 237:1133-1143.
72. O'Carroll D, Erhardt S, Pagani M, Barton SC, Surani MA, Jenuwein T. The polycomb-group gene Ezh2 is required for early mouse development. Mol Cell Biol 2001; 21:4330-4336.
73. Huang XJ, Wang X, Ma X, Sun SC, Zhou X, Zhu C, Liu H. EZH2 is essential for development of mouse preimplantation embryos. Reprod Fertil Dev 2014; 26:1166-1175.
74. Amouroux R, Nashun B, Shirane K, Nakagawa S, Hill PW, D'Souza Z, Nakayama M, Matsuda M, Turp A, Ndjetehe E, Encheva V, Kudo NR, et al. De novo DNA methylation drives 5hmC accumulation in mouse zygotes. Nat Cell Biol 2016; 18:225-233.
75. Auclair G, Guibert S, Bender A, Weber M. Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. Genome Biol 2014; 15:545.
76. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 1999; 99:247-257.
77. Olive V, Minella AC, He L. Outside the coding genome, mammalian microRNAs confer structural and functional complexity. Sci Signal 2015; 8:re2.
78. Megel C, Morelle G, Lalande S, Duchene AM, Small I, Marechal-Drouard L. Surveillance and cleavage of eukaryotic tRNAs. Int J Mol Sci 2015; 16:1873-1893.

79. Gebetsberger J, Polacek N. Slicing tRNAs to boost functional ncRNA diversity. RNA Biol 2013; 10:1798-1806.
80. Kiani J, Rassoulzadegan M. A load of small RNAs in the sperm - how many bits of hereditary information? Cell Res 2013; 23:18-19.
81. Kiani J, Grandjean V, Liebers R, Tuorto F, Ghanbarian H, Lyko F, Cuzin F, Rassoulzadegan M. RNA-mediated epigenetic heredity requires the cytosine methyltransferase Dnmt2. PLoS Genet 2013; 9:e1003498.
82. Schuster A, Skinner MK, Yan W. Ancestral vinclozolin exposure alters the epigenetic transgenerational inheritance of sperm small noncoding RNAs. Environmental Epigenetics 2016; 2.
83. Elbarbary RA, Takaku H, Uchiumi N, Tamiya H, Abe M, Takahashi M, Nishida H, Nashimoto M. Modulation of gene expression by human cytosolic tRNase Z(L) through 5'-half-tRNA. PLoS One 2009; 4:e5908.
84. Ivanov P, Emara MM, Villen J, Gygi SP, Anderson P. Angiogenin-induced tRNA fragments inhibit translation initiation. Mol Cell 2011; 43:613-623.
85. Wang Q, Lee I, Ren J, Ajay SS, Lee YS, Bao X. Identification and functional characterization of tRNA-derived RNA fragments (tRFs) in respiratory syncytial virus infection. Mol Ther 2013; 21:368-379.
86. Yamasaki S, Ivanov P, Hu GF, Anderson P. Angiogenin cleaves tRNA and promotes stress-induced translational repression. J Cell Biol 2009; 185:35-42.
87. Petersen CP, Reddien PW. Wnt signaling and the polarity of the primary body axis. Cell 2009; 139:1056-1068.
88. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005; 120:15-20.
89. Yan W. Potential roles of noncoding RNAs in environmental epigenetic transgenerational inheritance. Mol Cell Endocrinol 2014; 398:24-30.
90. Schaefer M, Pollex T, Hanna K, Tuorto F, Meusburger M, Helm M, Lyko F. RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. Genes Dev 2010; 24:1590-1595.

FIGURE LEGENDS

Figure 1 Quality and reproducibility of the sperm RNA isolation procedure reported. **A**) Comparison of the capability of Trizol and the mirVana lysis buffer to dissolve the mouse sperm heads. Bar =10μm. **B**) Representative electropherograms of total RNA isolated from mouse and rat epididymal sperm, and rabbit and human ejaculated sperm along with their respective RNA integrity number (RIN), as determined using an Agilent Bioanalyzer 2100. **C**) Scatter plots showing the distribution of the number of normalized small RNA sequencing reads (log2 values) in two biological replicates of mouse and human total sperm and mouse sperm heads. R^2 values represent Pearson correlation coefficients.

Figure 2 Sperm-borne mRNA fragments. **A**) Venn diagram showing the number of sperm-borne mRNA fragments conserved among four mammalian species (mouse, rat, rabbit and human). **B**) Venn diagram showing conserved mRNA fragments between rabbit and mouse sperm heads. Only genes with common names were considered to be conserved ones. **C**) Word cloud showing the biological process (BP) terms significantly enriched ($p \leq 0.05$) among the conserved coding

genes in total sperm of four species (mouse, rat, rabbit and human) analyzed in this study The size of each word is based on its frequency within the enriched BP terms and the original data can be found in Supplemental Table S4A (no filtering).
Supplemental Table S4 (no filtering).

Figure 3 Sperm small RNA contents. Pie charts showing the proportional distribution of each of the eight sncRNA populations in total sperm (TS) or sperm heads (SH) based on sncRNA datasets currently available in SpermBase. The original data can be found in Supplemental Table S5.

Figure 4 Sperm-borne miRNAs. **A**) Venn diagram showing sperm miRNAs conserved among four mammalian species (moue, rat, rabbit and human). **B**) Examples of critical early development genes predicted to be targeted by conserved sperm-borne miRNAs. **C**) Word cloud showing the biological process (BP) terms significantly enriched ($p \leq 0.05$) in the predicted targets of conserved sperm miRNAs among the four mammalian species analyzed in this study (mouse, rat, rabbit and human). The size of each word is based on its frequency within the enriched BP terms and the original data can be found in Supplemental Table S8 (no filtering).

Figure 5 Sperm-borne tsRNAs. **A**) Pie charts showing the proportional distribution of each of the five tsRNA subclasses in total sperm (TS) and sperm heads (SH) in the four mammalian species analyzed in this study. "Other" refers to tRNA-aligned reads that were not classified as tsRNAs. **B**) Histogram showing the origin of sperm-borne 5' tRNA halves in the four mammalian species. The tRNA genes from which the 5' tRNA halves were derived were grouped by amino acid (AA), and ranked. "Misc" refers to amino acids not shown on the table. **C**) Example of potential binding of the 5' UTR of *H1foo* mRNA by a 5' tRNA half. **D**) Scatter plots showing the relative number of targeting sites for the 5' tRNA halves in different regions of the target genes (5'UTR, CDS and 3'UTR). For each species, the number of predicted gene targets for each 5' tRNA half when matched against 5' UTR, CDS, and 3' UTR sequences were normalized to the number of 3' UTR-based gene targets observed. **E**) Word cloud showing the biological process (BP) terms significantly enriched ($p \leq 0.05$) in the predicted targets of conserved sperm-borne 5' tRNA halves among the four mammalian species analyzed in this study (mouse, rat, rabbit and human). The size of each word is based on its frequency within the enriched BP terms and the original data can be found in Supplemental Table S10 (no filtering).

TABLE 1 Summary of sperm RNA isolation procedures.

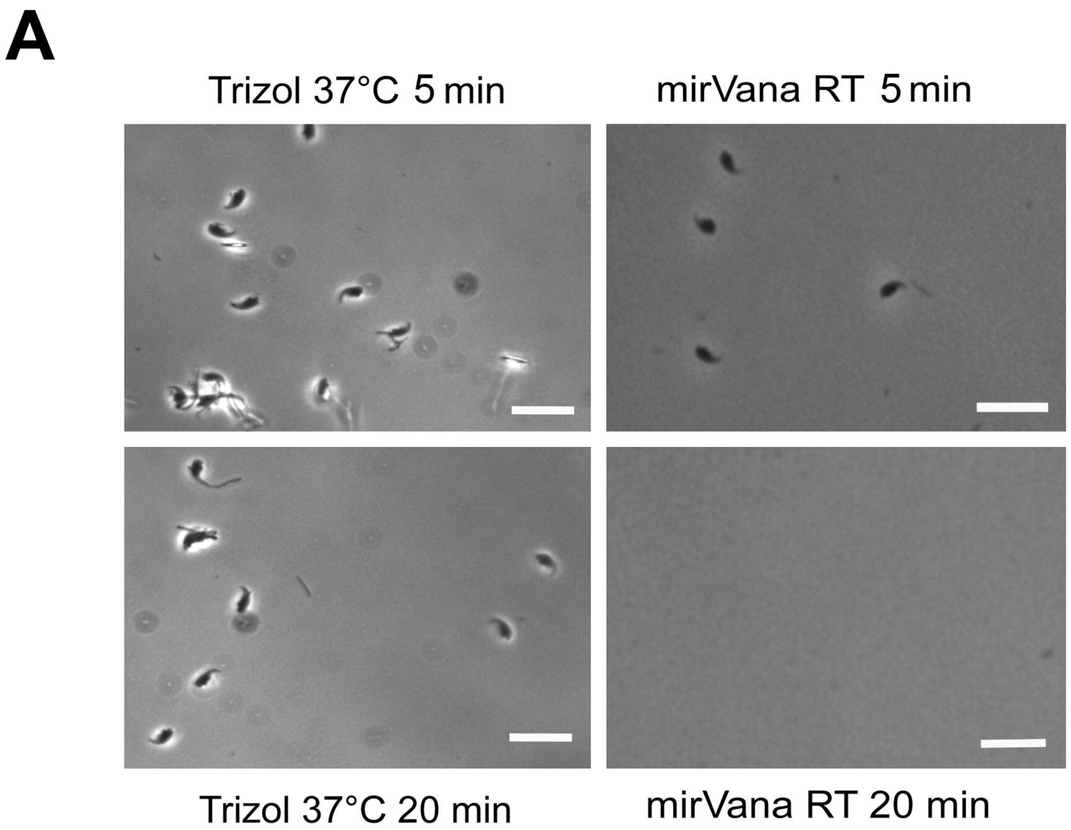
Procedure	Mouse	Rat	Rabbit	Human
Sperm collection	Whole epididymis	Whole epididymis	Ejaculate (artificial vagina method)	Ejaculate (masturbation)
Sperm processing	Sperm release from the epididymis→one wash followed by swim-up → three washes of the swim-up sperm → microscopic examination for purity → snap freezing the sperm pellet in LN2→ storage at -80°C until RNA isolation	Sperm release from the epididymis→ light digestion using collagenase and hyaluronidase →three washes → microscopic examination for purity → Snap freezing the sperm pellet in LN2 → storage at -80°C until RNA isolation	Three washes → microscopic examination for purity → Snap freezing the sperm pellet in LN2 → storage at -80°C until RNA isolation	Three washes → microscopic examination for purity → Snap freezing the sperm pellet in LN2 → storage at -80°C until RNA isolation
Sperm lysis using the buffer from the mirVana miRNA isolation kit, or a lysis buffer for RNA isolation containing detergent (SDS or SLS) and a reducing agent (DTT or β-mercaptoethanol)	Add the lysis buffer and homogenize on a low setting for one minute on ice → microscopic observation to ensure a complete lysis of sperm →Repeat, if necessary, until all sperm heads get dissolved completely.	Add the lysis buffer and homogenize on a low setting for 90 seconds on ice, followed by a five minute incubation at 65°C. → microscopic observation to ensure a complete lysis of sperm →Repeat, if necessary, until all sperm heads get dissolved completely.	Add the lysis buffer and pipet the sperm gently up and down on ice until sperm pellet dissolves → microscopic observation to ensure a complete lysis of sperm →Repeat, if necessary, until all sperm heads get dissolved completely.	Add the lysis buffer and pipet the sperm gently up and down on ice until sperm pellet dissolves → microscopic observation to ensure a complete lysis of sperm →Repeat, if necessary, until all sperm heads get dissolved completely.
Total RNA isolation using the mirVana miRNA isolation kit	Follow the default protocol.	Follow the default protocol.	Follow the default protocol.	Follow the default protocol.
RNA quality control using the RNA 6000 Nano chips on an Agilent 2100 Bioanalyzer	RIN = 2-4: good sperm RNA quality. RINs <2: suggestive of excessive RNA degradation. RINs >4: indicative of somatic cell contamination.	RIN = 2-4: good sperm RNA quality. RINs <2: suggestive of excessive RNA degradation. RINs >4: indicative of somatic cell contamination.	RIN = 2-4: good sperm RNA quality. RINs <2: suggestive of excessive RNA degradation. RINs >4: indicative of somatic cell contamination.	RIN = 2-4: good sperm RNA quality. RINs <2: suggestive of excessive RNA degradation. RINs >4: indicative of somatic cell contamination.
Yield	~35-50ng per million swim-up epididymal total sperm, and ~25-40ng per million sonication-resistant sperm heads. ~10 million swim-up sperm can be obtained from one male mouse (two epididymides), yielding~350-500ng total RNA.	~35-65ng per million epididymal total sperm. ~20 million sperm can be obtained from one adult male rat, which yield ~700-1,300ng total RNA.	~40-55ng per million ejaculated total sperm, and ~25-40ng per million sonication-resistant sperm heads. 20 million sperm can be obtained after washes, which yield 800-1,100ng total RNA.	~45-60ng per million ejaculated sperm. An average of ~10 million sperm can be obtained from one ml donor semen, leading to 450-600ng total RNA.

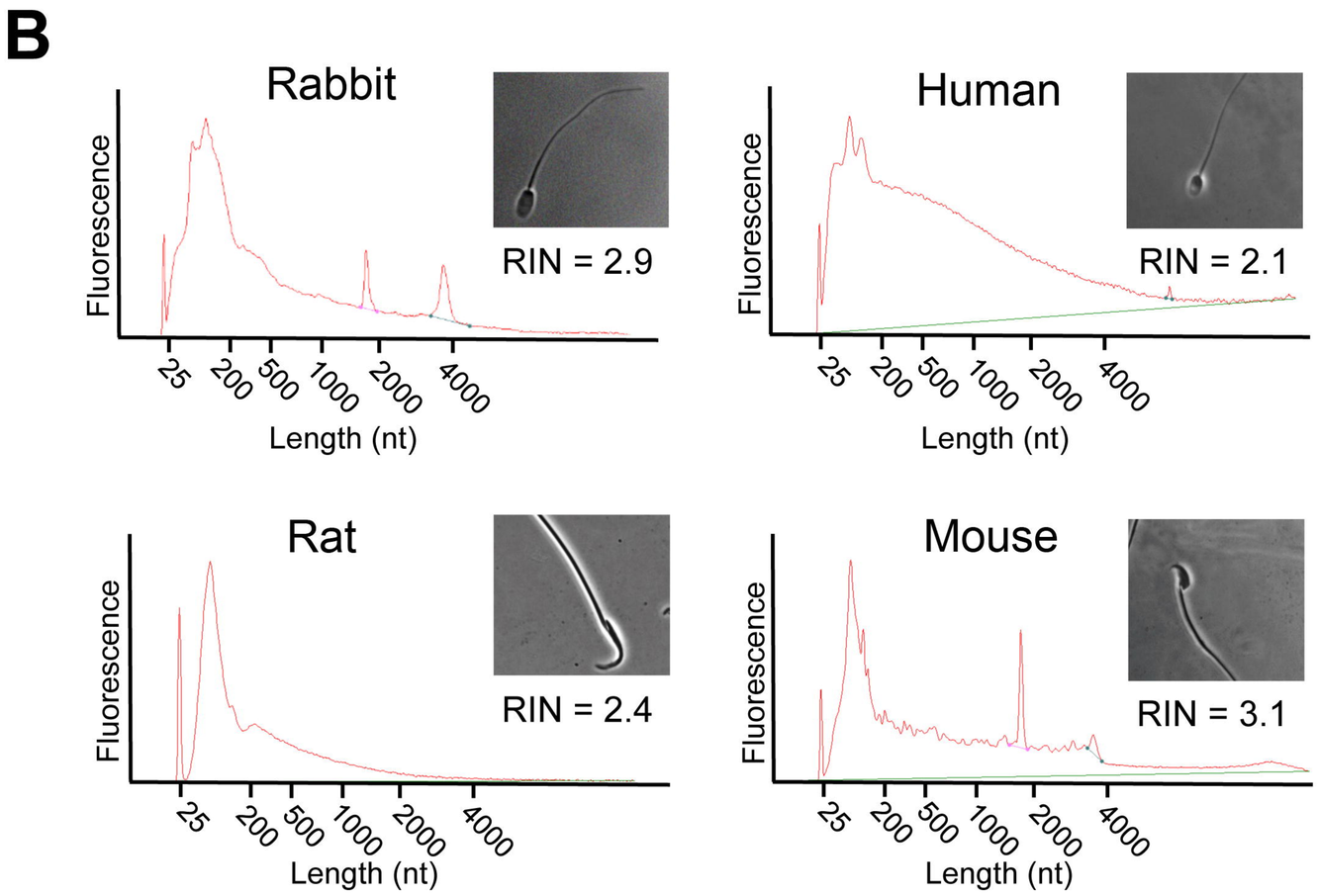
TABLE 2. Predicted targets of highly conserved sperm-borne miRNAs and 5' tRNA halves compared to early development genes.*

Stage	No. early development genes (≥ 5 FPKM)	miRNA target genes		Target genes of 5' tRNA halves	
		No. expected random matches	No. actual matches	No. expected random matches	No. actual matches
Oocyte	7,374	1,681	1,818	6,164	6,476
1-Cell	7,111	1,621	1,779	5,944	6,240
2-Cell	7,309	1,666	1,818	6,110	6,446
4-Cell	7,273	1,658	1,804	6,080	6,401

*Statistical analyses were conducted using χ^2 test; predicted targets for both the sperm-borne miRNAs and 5' tRNA halves significantly matched early development genes, with *P*-values of 1.6E-11 and 3.2E-14, respectively.

Figure 1


A
Trizol 37°C 5 min
mirVana RT 5 min
Trizol 37°C 20 min
mirVana RT 20 min


B
Rabbit
RIN = 2.9
Human
RIN = 2.1
Rat
RIN = 2.4
Mouse
RIN = 3.1
Fluorescence
25
200
500
1000
2000
4000
Length (nt)

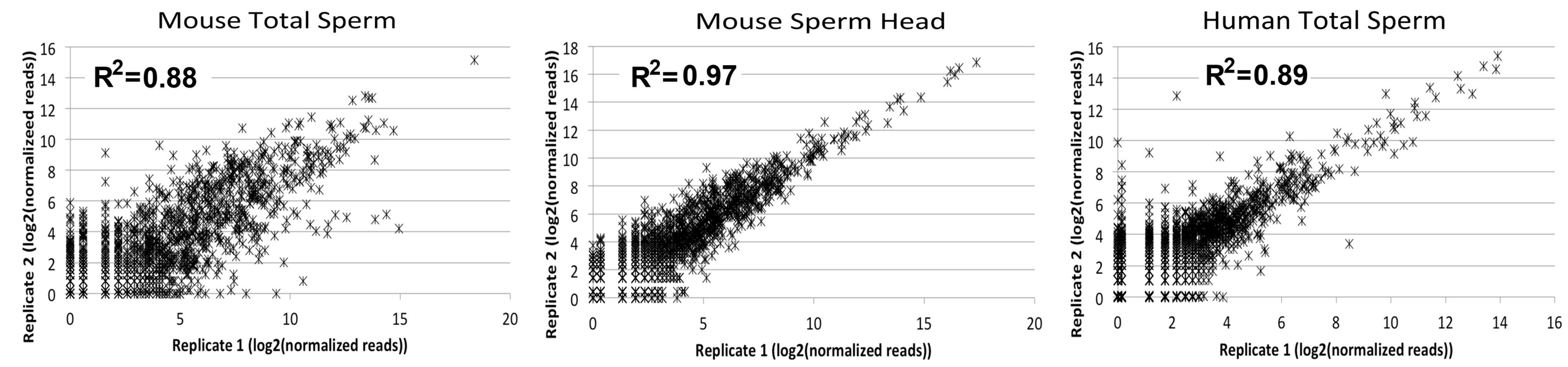
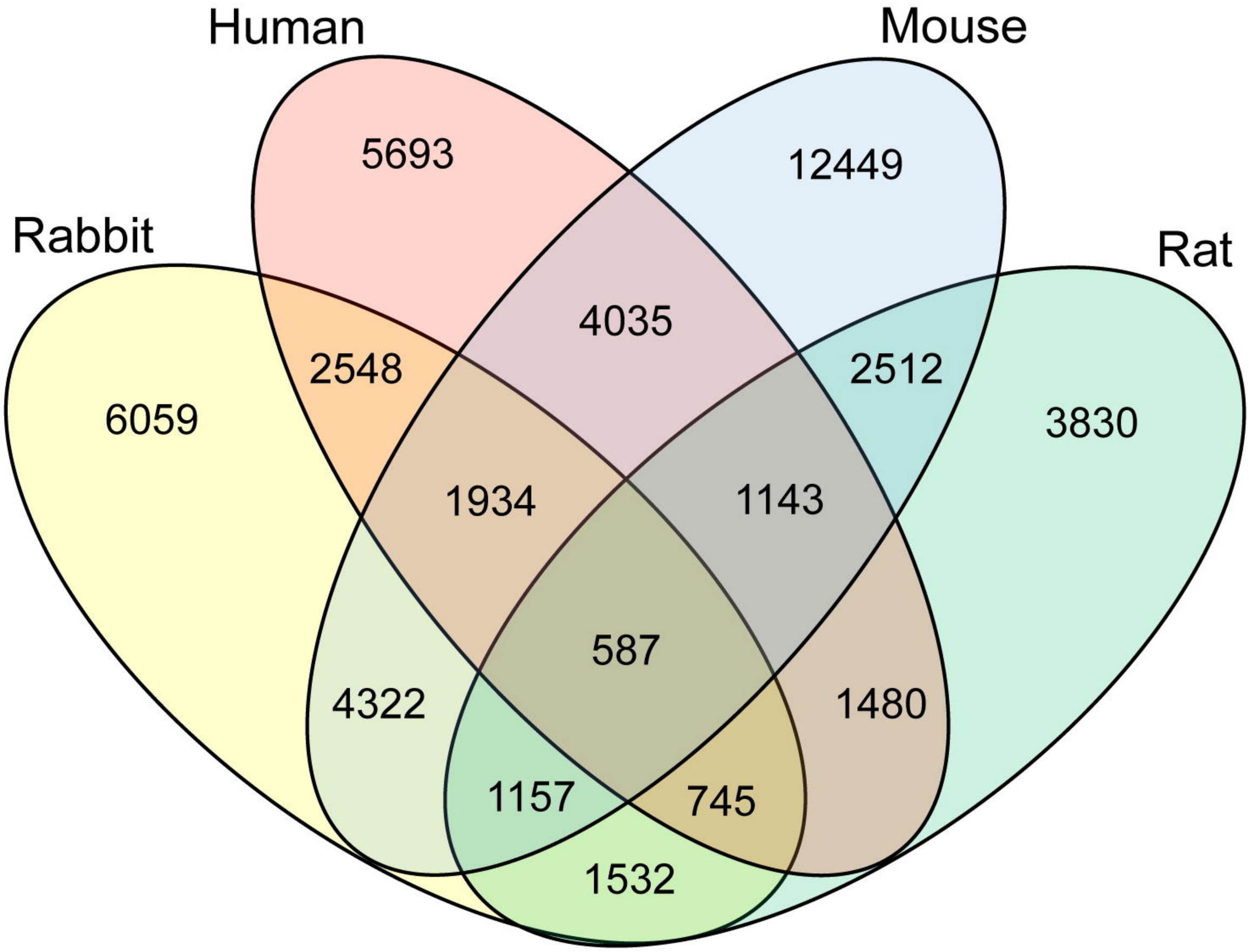
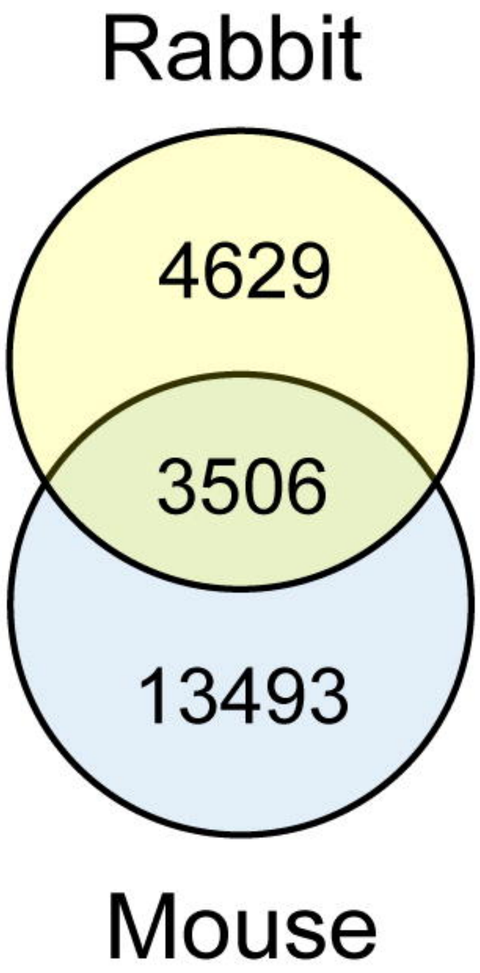

C
Mouse Total Sperm
R2=0.88
Mouse Sperm Head
R2=0.97
Human Total Sperm
R2=0.89
Replicate 2 (log2(normalizeed reads))
Replicate 2 (log2(normalized reads))
Replicate 1 (log2(normalized reads))

Figure 2

A


Human
Mouse
Rabbit
Rat
5693
12449
4035
2548
2512
6059
3830
1934
1143
587
4322
1480
1157
745
1532

B


Rabbit
4629
3506
13493
Mouse

C

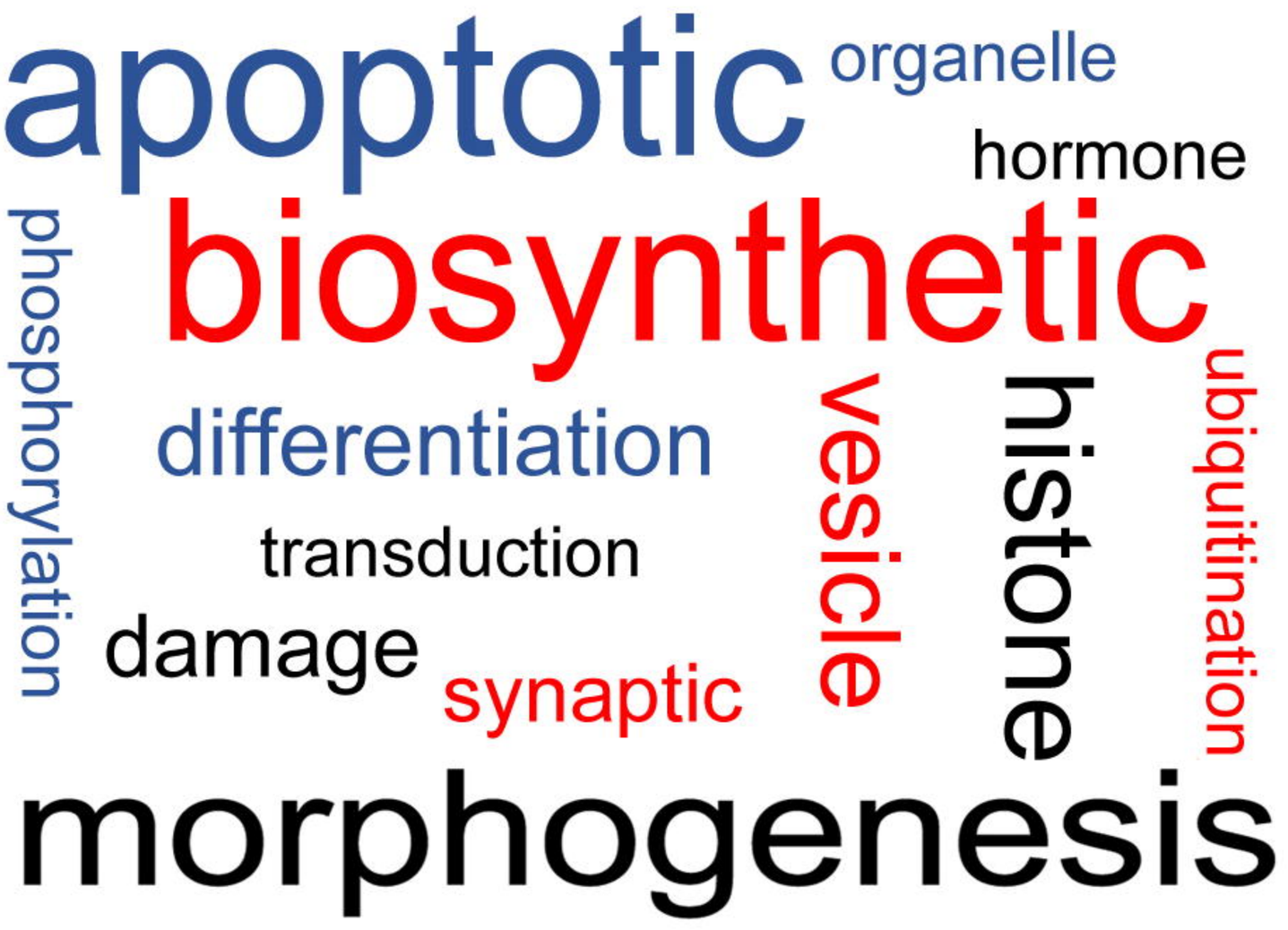
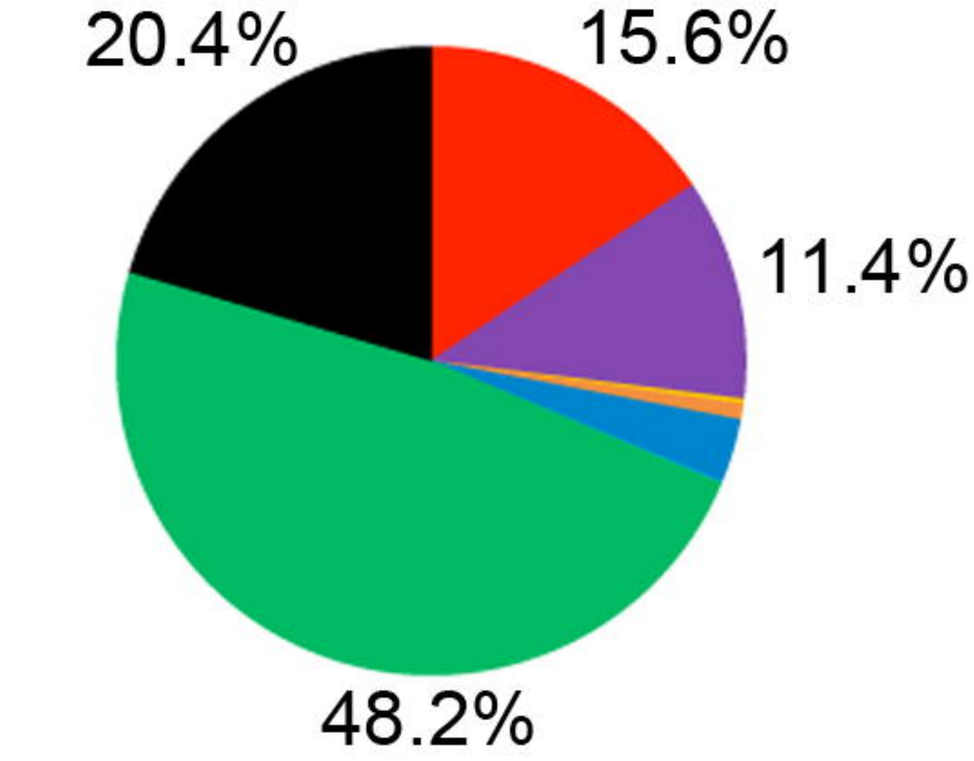
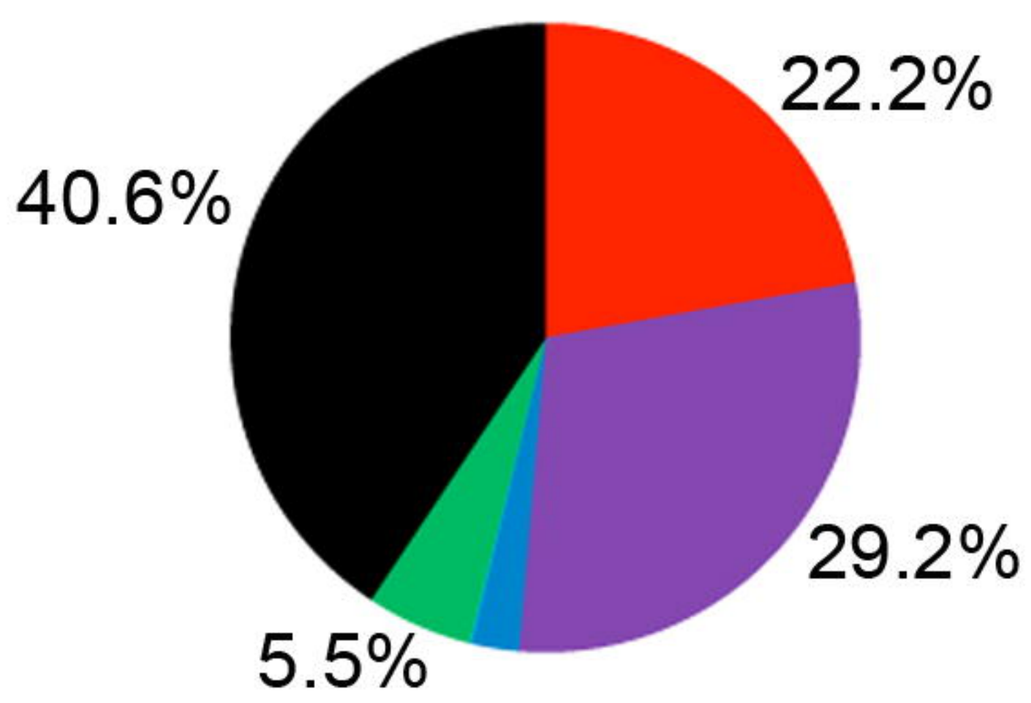
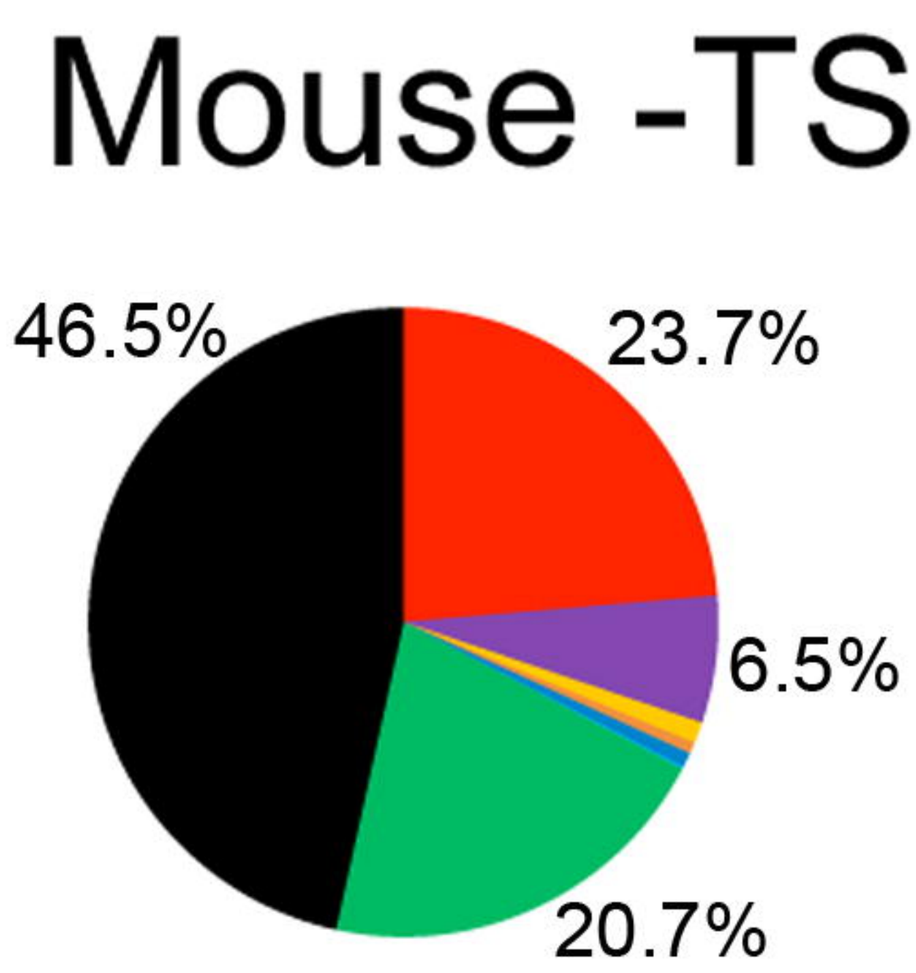
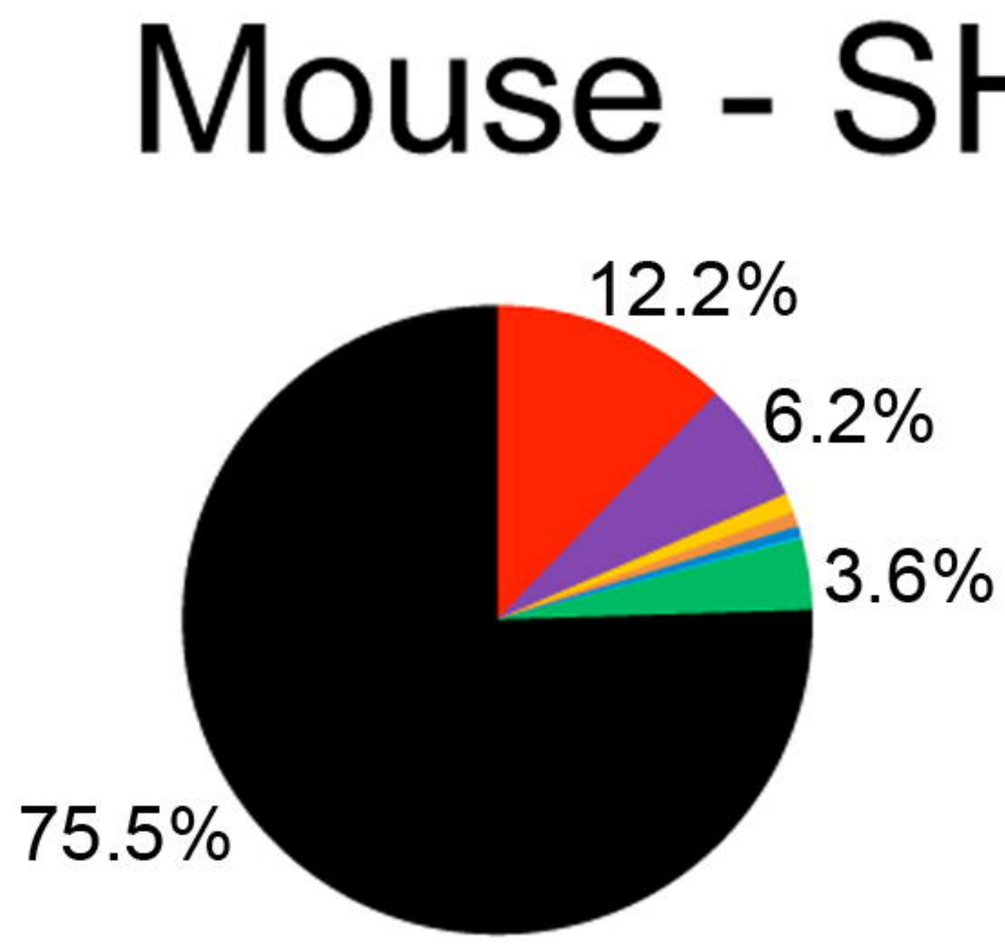

apoptotic
organelle
hormone
biosynthetic
phosphorylation
differentiation
vesicle
histone
ubiquitination
transduction
damage
synaptic
morphogenesis

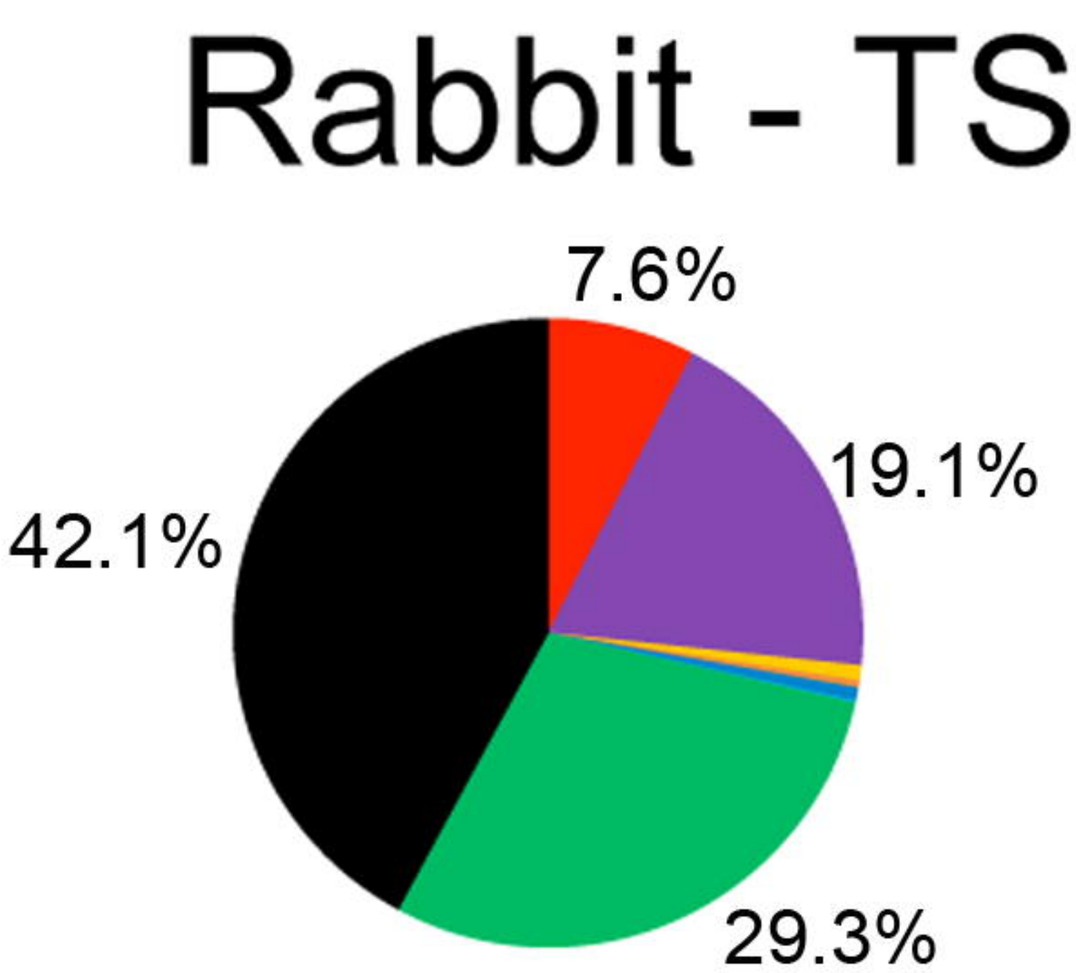
Figure 3

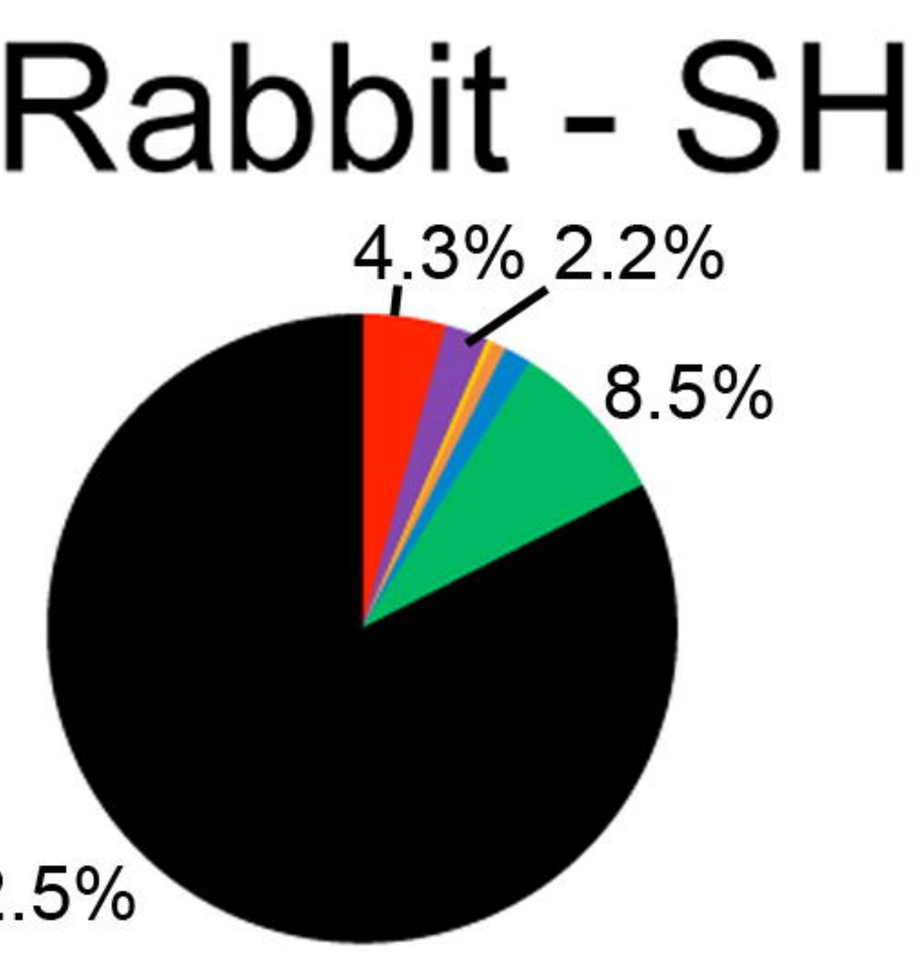

Human - TS
20.4%
15.6%
11.4%
48.2%


Rat - TS
40.6%
22.2%
29.2%
5.5%


Mouse -TS
46.5%
23.7%
6.5%
20.7%


Mouse - SH
12.2%
6.2%
3.6%
75.5%

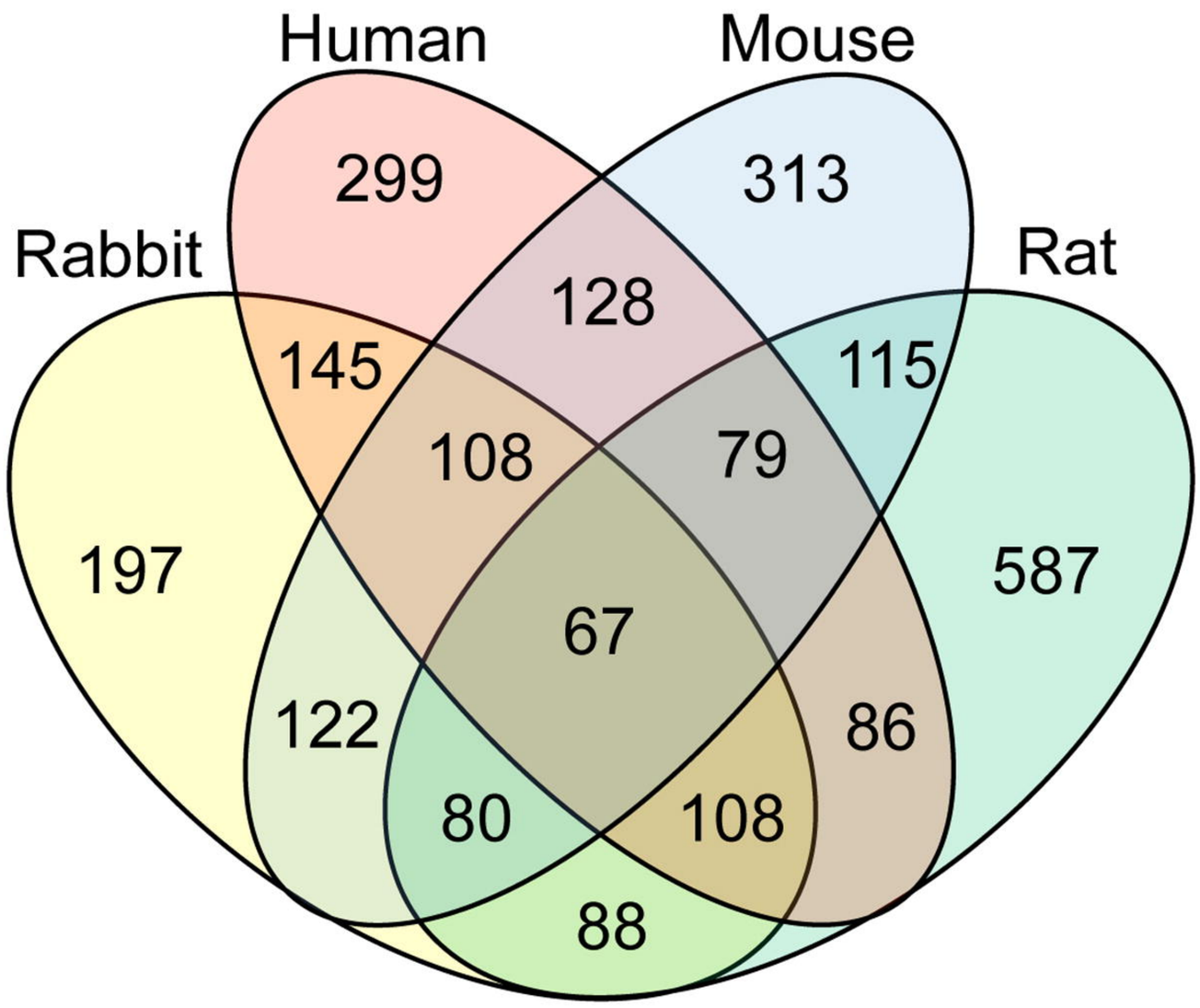

Rabbit - TS
7.6%
19.1%
42.1%
29.3%


Rabbit - SH
4.3%
2.2%
8.5%
82.5%

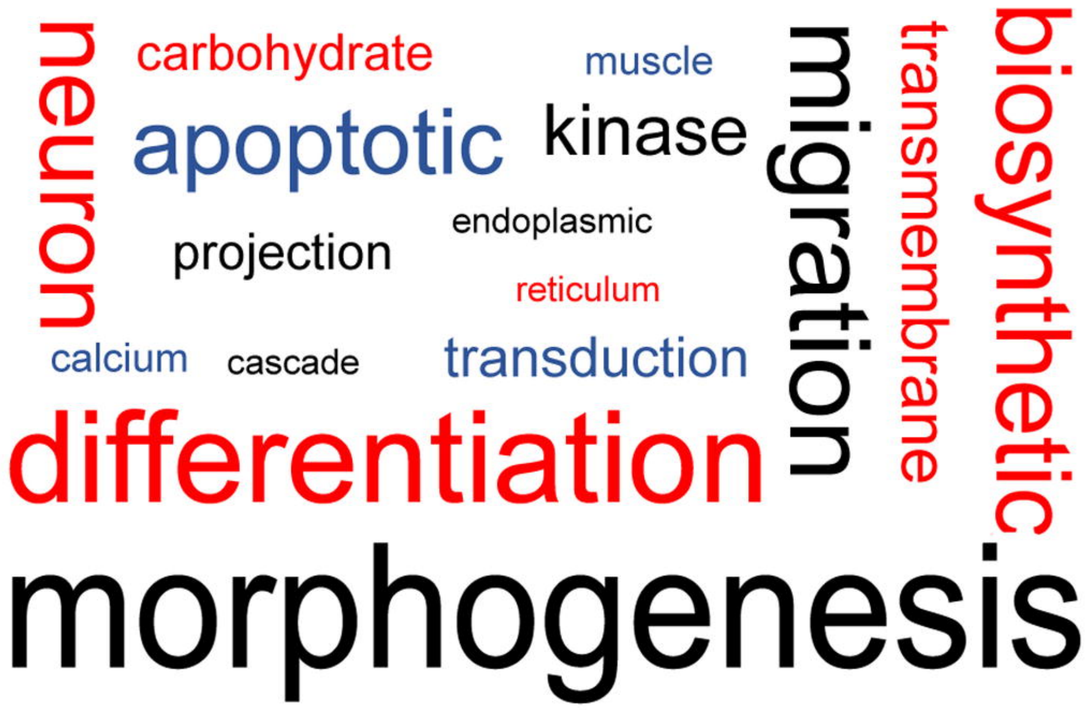

miRNA
piRNA
snoRNA
snRNA
rRNA
endo-siRNA
mt-RNA
tsRNA

Figure 4

A


Human
Mouse
Rabbit
Rat
299
313
128
145
115
108
79
197
587
67
122
86
80
108
88

C

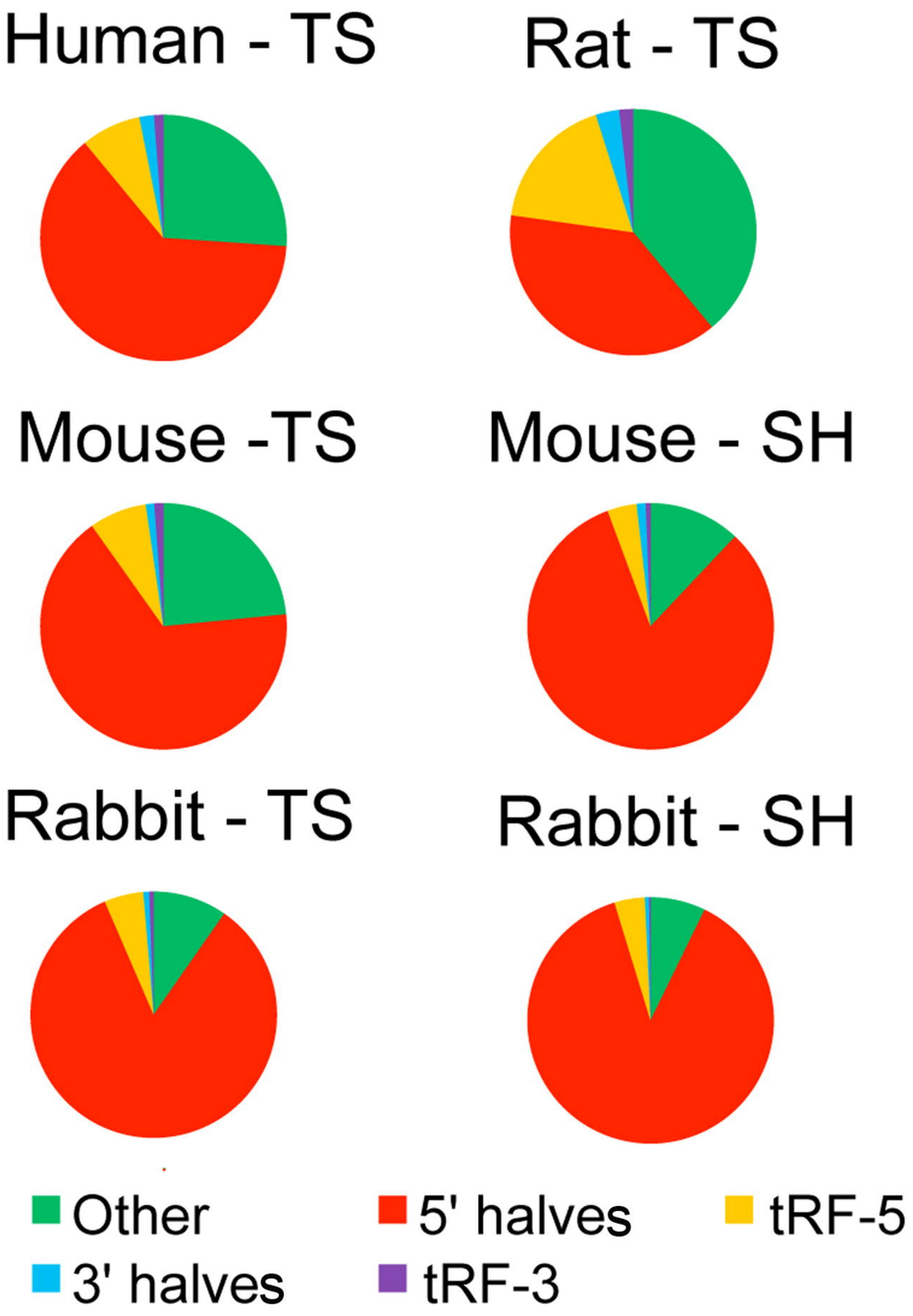

neuron
carbohydrate
muscle
apoptotic
kinase
migration
transmembrane
biosynthetic
endoplasmic
projection
reticulum
calcium
cascade
transduction
differentiation
morphogenesis

B

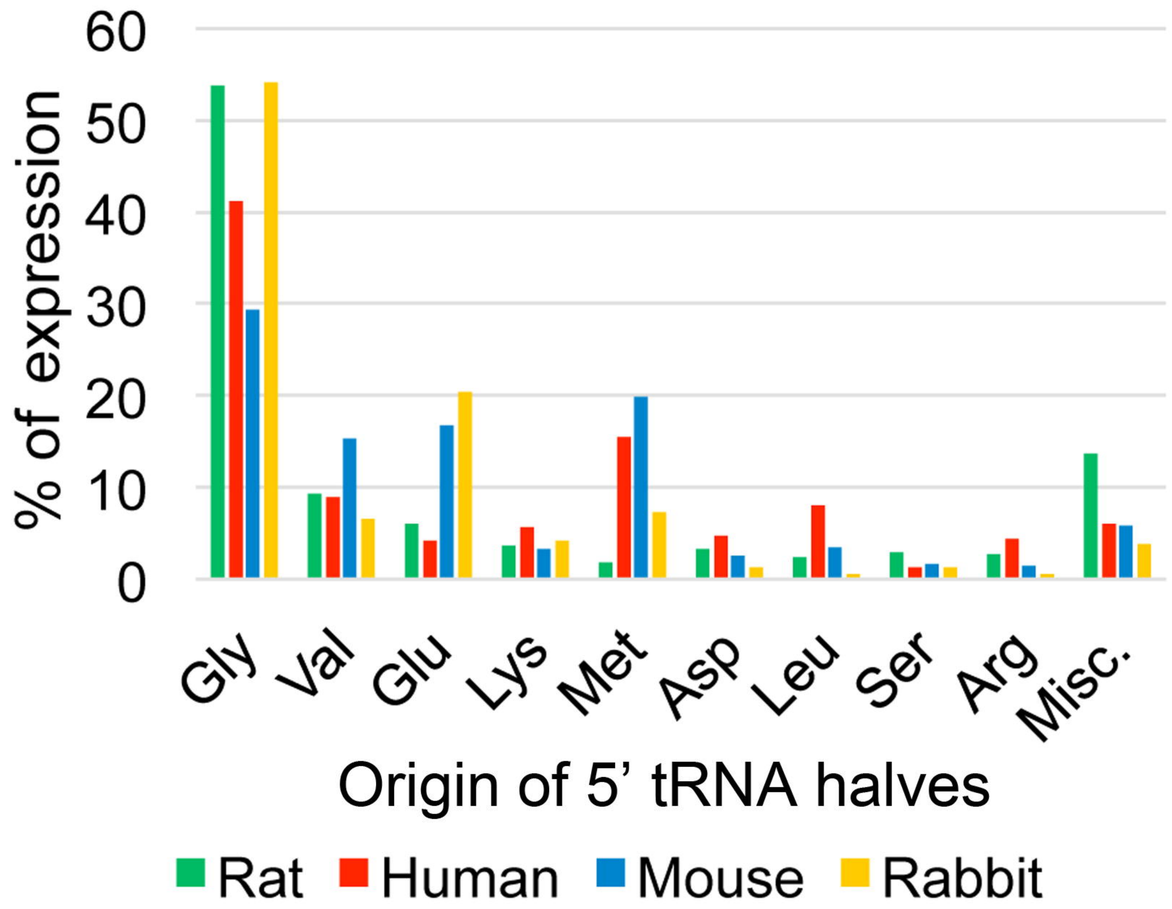
mmu-miR-34a-5p
3’ UG U 5’
UUGGUCGA UCU GUGACGGU
AACCAGUU AGA CGCUGUUA
5’ 3’
A UCU AGC C
Dnmt3a 3’ UTR
mmu-miR-27b-3p
3’ C U 5’
GUCUUGA AUCGG UGACACU
CAGAGCU UAGCC ACUGUGG
5’ 3’
A G A C
Gata4 3’ UTR
mmu-miR-149-3p
3’ C C CA 5’
GUGG GGGGG GGGAGGGAG
UACC CUUCC CCUUCUCUC
5’ 3’
C U U
Ezh2 3’ UTR

Figure 5

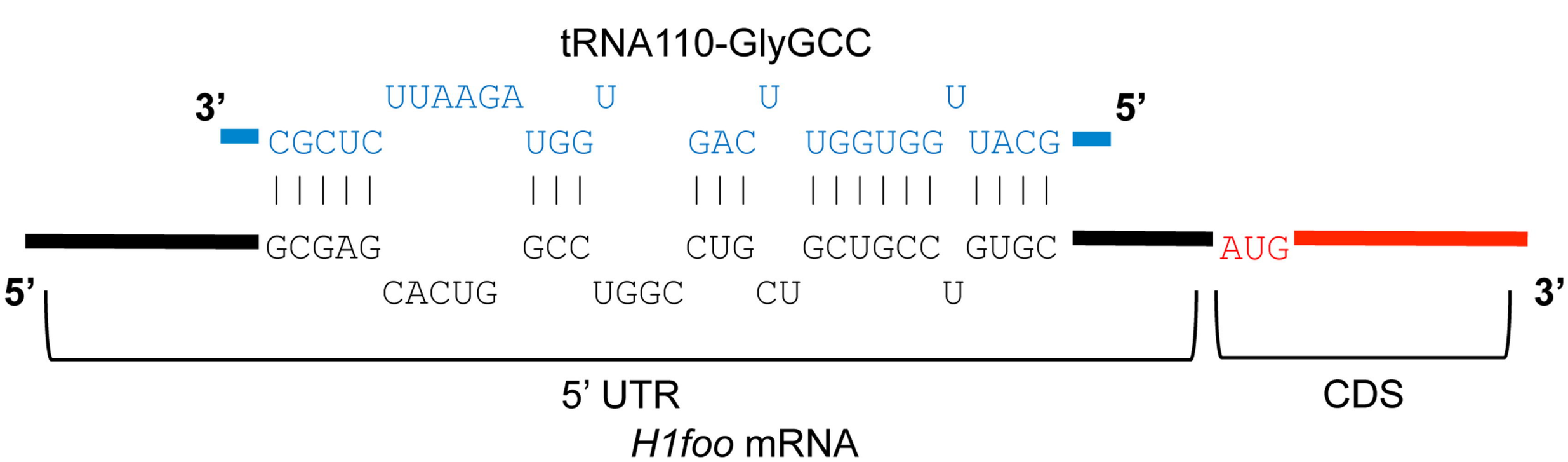
A



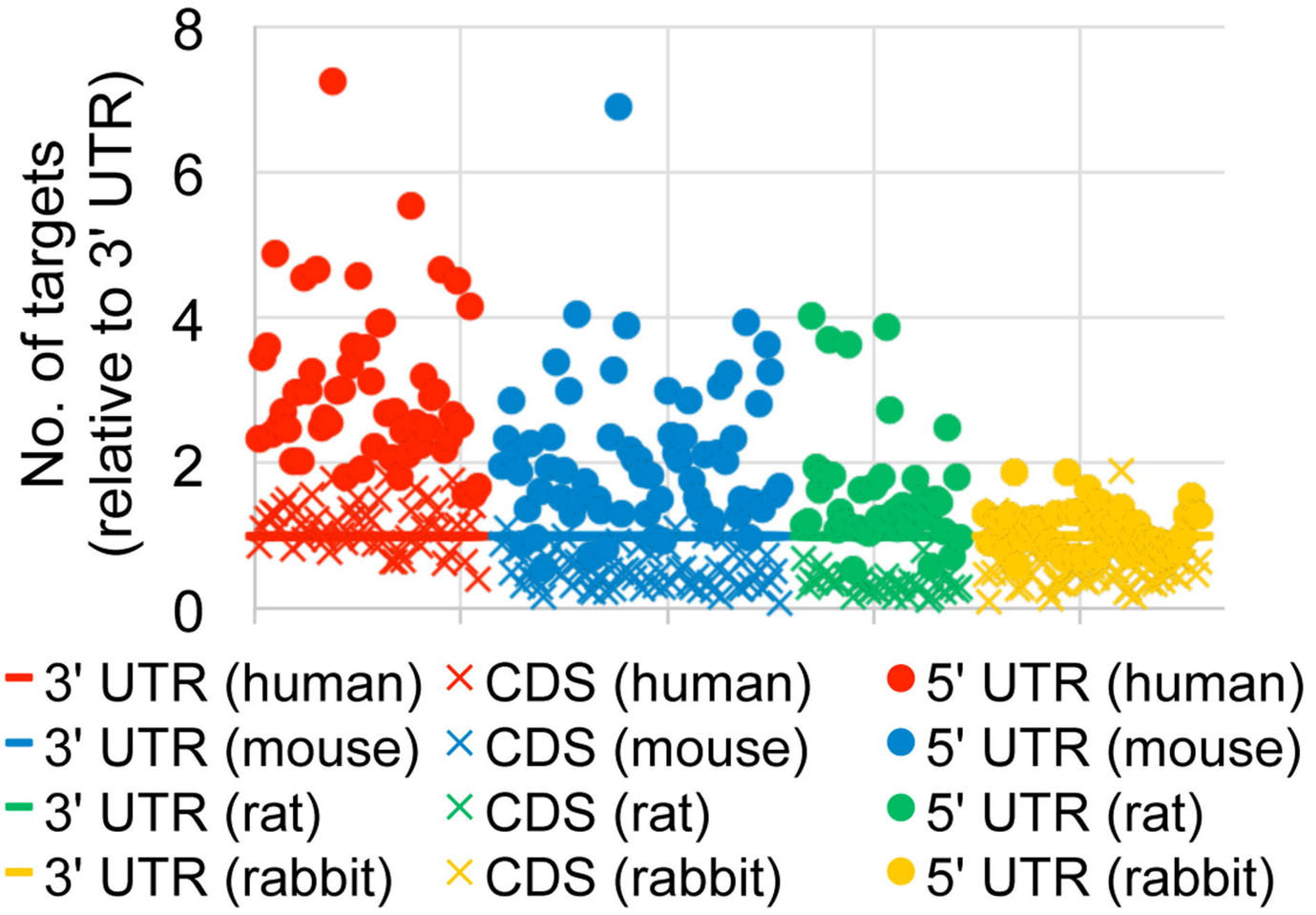
B



C



D



E