Supplemental Methods

RNA extraction and conventional sequencing library preparation

The RNA used for SSV sequencing was extracted using a low molecular weight RNA extraction kit (mirVana, Invitrogen) as previously described (Deschamps-Francoeur et al. 2014), and from these samples, cDNA libraries were prepared using the TruSeq small RNA Sample Prep kit (Illumina), which includes adapter ligation, reverse transcription and PCR amplification. The RNA used for FAV and FRV sequencing was isolated and purified from 5 ug DNA-free total RNA extracted using either a NEBNext Poly(A) mRNA Isolation Module (New England Biolabs) in the case of FAV sequencing or Ribo-Zero Gold (Illumina) in the case of FRV according to the manufacturers' protocol. Library preparations were performed using the NEBNext Ultra directional RNA library Prep Kit for Illumina (New England Biolabs) in order to generate an RNA-seq library from 100 ng of purified RNA. The RNA and library quality were verified using an Agilent Bioanalyzer (Agilent Technologies). The resulting cDNA libraries were paired-end sequenced on Illumina HiSeq sequencers at the McGill University and Genome Quebec Innovation Centre. The number of cycles was 2x126, 2x100 and 2x100 nucleotides for FRV, FAV and SSV, respectively. For the SSV libraries, samples were multiplexed resulting in between 15M and 24M paired-end reads per dataset while the FAV and FRV samples were each run on an entire lane yielding > 150M paired-end reads. Bioinformatics sampling of the resulting data was used to ensure that the difference in read depth does not affect the overall distribution of RNA abundance in the different data sets.

Construction and sequencing of TGIRT-seq libraries

TGIRT-seq libraries were constructed as previously described (Nottingham et al. 2016; Qin et al. 2016). Purified total RNA, extracted from SKOV3ip1 as described above, was first ribodepleted by using a Ribo-Zero Gold (Human/Mouse/Rat) kit (Illumina). The resulting ribodepleted RNAs were either (1) used directly as input for TGIRT template-switching reactions or (2) fragmented with an NEBNext Magnesium RNA Fragmentation Module (New England Biolabs) by incubation at 94°C for 7 min and then treated with T4 polynucleotide kinase to remove 3' phosphates prior to template-switching (Nottingham et al. 2014). All RNA modifying reactions were cleaned up using a modified version of the Zymo RNA Clean & Concentrator (Zymo Research) protocol (with the addition of 8 sample volumes of ethanol to increase retention of very small RNA species). cDNAs were synthesized via TGIRT template-switching with either 0.5 or 1 μM TGIRT-III reverse transcriptase (InGex, LLC) for 15 min at 60° C, during which a DNA oligonucleotide containing the complement of an Illumina Read 2 sequencing primer-binding site becomes seamlessly linked to the 5' cDNA end. After reaction cleanup, a 5' adenylated DNA oligonucleotide containing the complement of an Illumina Read 1 sequencing primer-binding site was then ligated to the 3' cDNA end with Thermostable 5' AppDNA / RNA Ligase (New England Biolabs). Properly ligated cDNAs were amplified by PCR (12 cycles) to synthesize the second strand and add Illumina flow cell capture and index sequences. Libraries were size-selected with Ampure XP beads (Beckman-Coulter) and evaluated on an Agilent 2100 Bioanalyzer. TGIRT-seq libraries were sequenced either on the Illumina NextSeq 500 platform (2 x 150 paired-end reads), the Illumina HiSeq 2500 platform (2 x 125 paired-end reads) or the Illumina HiSeq 4000 platform (2 x 150 paired-end reads) at the Genome Sequencing and Analysis Facility (GSAF) at the University of Texas at Austin.

RNAseq analysis

All datasets were passed through a quantification pipeline to obtain counts per million (CPM) and transcript per million (TPM) values. Fastq files were checked for quality using FastQC and trimmed using Cutadapt (Didion et al. 2017) (with --minimum-length 2 --match-read-wildcards -q 3 --paired-output) and Trimmomatic (Bolger et al. 2014) (with TRAILING:30) to remove adapters and portions of reads of low quality, respectively. Read pairs were then aligned to the human genome build hg38 using an annotation file obtained from Ensembl (described below) with the splice-aware RNA-seq aligner STAR v2.5.1b (Dobin and Gingeras 2016) using the following parameters: --outSAMtype BAM SortedByCoordinate, --outSAMprimaryFlag AllBestScore, --alignIntronMax 1250000, all other parameters at default values. Reads not aligned using STAR were aligned once again using Bowtie v2.2.9 Langmead and Salzberg 2012 (Langmead and Salzberg 2012), which performs well for the alignment of shorter reads. Parameters used for Bowtie are the following: --local, -p 24, -q, -l 13. Read pairs successfully aligned by STAR or Bowtie were merged into a BAM file and separated into two groups: those that align to one genomic position and those that align to more than one genomic position, qualified as uniquely and multimapped reads.

Aligned reads were then assigned to genomic features using Rsubread v1.20.6 (Liao et al. 2014) and an annotation file described below generated using the correct_annotation module of our quantification pipeline CoCo which ensures that nested genes (such as most snoRNA, several snRNA and tRNA) and multimapped reads are properly quantified. FeatureCounts parameters were set to (countChimericFragments=FALSE, GTF.featureType="exon", GTF.attrType="gene_id", useMetaFeatures=TRUE, largestOverlap=TRUE, isPairedEnd=TRUE, requireBothEndsMapped=TRUE, allowMultiOverlap=FALSE, strandSpecific={2 for FRV, 1 for all the other datasets}(Deschamps-Francoeur et al. 2014), minOverlap=10, countMultiMapping=TRUE, nthreads=20, reportReads=TRUE). The assigned multimapped reads where grouped using a modified version of count_from_bed.pl (available at https://github.com/mw55309/RNAfreak). The multimapped reads were distributed within their group based on the ratio of uniquely mapped reads assigned to each member of the group. Normalized read counts are given in CPM and TPM, as calculated as described below.

Conversion from counts to CPM and TPM

The read counts assigned to genes are normalized by the total number of read pairs assigned to all genes to give counts per million (CPM):

$$CPM\ (i) = \frac{count(i)}{\sum_{j\,\epsilon\,J} count(j)} * 10^6$$

where count(i) represents the number of read pairs aligning to gene i and J represents the set of all genes in the annotation.

To obtain a value that is representative of the abundance of the transcripts in the sample, counts were first normalized by the length of the main transcript per gene (highest transcript support level in Ensembl and highest value source, prioritizing ensembl_havana (Yates et al. 2016) and then are normalized by the total of these normalized counts to get transcripts per million values (TPM):

$$TPM\ (i) = \frac{count(i)/length(i)}{\sum_{j\,\epsilon\,J}(\frac{count(j)}{length(j)})} * 10^6$$

where count(i) and length(i) represent respectively the number of read pairs aligning to gene i and the length of gene i, and J represents the set of all genes in the annotation. The length normalization can be biased when the gene length falls under the maximum read length (ie the number of cycles used for sequencing; for the datasets used, 150nt). To address this problem, all counts obtained from genes smaller than the cycle number were normalized by the cycle number for the specific dataset instead of by the transcript length. The rationale behind this method of calculating TPM, as opposed to the commonly used effective length calculated from the mean fragment length distribution (Li and Dewey 2011), is that genes that are longer than the specified cycle number have more windows to produce fragments while any gene equal or smaller to that length can only produce one fragment. **Table S4** gives the average abundance values in TPM over the samples considered for all genes considered.

Annotation modification

An annotation file in gene transfer format (.gtf) was obtained from Ensembl (Yates et al. 2016) (hg38, v87). The annotation file was supplemented with tRNA genes from GtRNAdb (Chan and Lowe 2016) and with snoRNA genes from Refseq (O'Leary et al. 2016) that were missing in Ensembl annotations. In total, 20 snoRNA and 628 tRNA were added to the Ensembl annotations (listed in **Table S7**). In addition, 63 gene annotations were removed from the gtf file because they have highly similar or identical coordinates to another gene, and thus keeping them would result in any reads aligning to them being labelled as ambiguous. These 63 redundant genes consist of 17 snoRNA, 43 miRNA, 2 lincRNA and 1 antisense RNA. Details are given in **Table S7**.

Gene biotype pooling

Gene biotypes as given by the Ensembl annotation files were pooled for simplicity. The groups “Protein_coding”, “Pseudogene” and “Long_noncoding” were obtained by pooling biotypes as recommended by Ensembl (http://useast.ensembl.org/Help/Faq?id=468). The group “Other” corresponds to any other biotype not listed.

End-point RT-PCR, quantitative RT-PCR analysis and digital PCR

Quantitative RT-PCR, primer design and validation were performed by the Université de Sherbrooke RNomics Platform (http://rnomics.med.usherbrooke.ca/) as previously described (Brosseau et al. 2010). RNA integrity was assessed with an Agilent 2100 Bioanalyzer (Agilent Technologies). Reverse transcription was performed using 1 μg total RNA with Transcriptor reverse transcriptase, random hexamers, dNTPs (Roche Diagnostics), and 10 units of RNaseOUT (Invitrogen) following the manufacturer’s protocol in a total volume of 10 μl. Primers were individually resuspended to 20–100 µM stock solution in Tris-EDTA buffer (IDT) and diluted as a primer pair to 1 µM in RNase DNase-free water (IDT). Reactions were performed in 10 μl reaction volumes placed in 384 well plates on a CFX-384 thermocycler (BioRad) with 5 µL of 2X iTaq Universal SYBR Green Supermix (BioRad), 10 ng (3 μl) cDNA, and 200 nM final (2 μl) primer pair solutions. Reactions were performed in three technical replicates and 2-3 biological replicates. The following cycling conditions were used: 3 min at 95°C; 50 cycles: 15 sec at 95°C, 30 sec at 60°C, 30 sec at 72°C. Relative expression levels were calculated using the qBASE framework (Hellemans et al. 2007). Relative RNA abundance was calculated using three reference genes YWHAZ, MRPL19 and SDHA. No-template control was included in each run to control for DNA contamination. The amplification efficiency of the different primers was tested by digital PCR. In general, digital and quantitative RT-PCR gave the same trends when compared to sequencing data. The primers used are listed in Table S8.

Splicing index were determined using end-point RT-PCR as previously described (Klinck et al. 2012). The reactions were performed using 10 ng cDNA in 10 µL final volume containing 0.2 mmol/L each dNTP, 1.5 mmol/L MgCl2, 0.6 µmol/L each primer, and 0.2 units of Platinum Taq DNA polymerase (Thermo Scientific). An initial incubation of 2 min at 95°C was followed by 35 cycles at 94°C 30 s, 55°C 30 s, and 72°C 60 s. The amplification was completed by 2-minutes incubation at 72°C. PCR reactions were carried out using thermocyclers SimpliAmp PCR System (Applied Biosystems, Life Technologies), and the amplified products were analyzed by automated chip-based microcapillary electrophoresis on Labchip GX Touch HT instruments (Perkin Elmer). Amplicon sizing and relative quantitation was performed by the manufacturer's software. The reactions were performed using 3 technical and 2-3 biological replicates. The primers used are listed in Table S9.

Digital PCR was performed by the Université de Sherbrooke RNomics Platform (http://rnomics.med.usherbrooke.ca/) using Bio-Rad QX200. The reactions were carried out using SYBR Green and 3 different housekeeping genes were used as internal control. The list of primers used are listed in Table

S10. Droplet Digital PCR (ddPCR) reactions are composed of 10ul of 2X QX200 ddPCR EvaGreen Supermix (Bio-Rad) ,10 ng (3 μl) cDNA,100 nM final (2 μl) primer pair solutions and 5ul molecular grade sterile water (Wisent) for a 20ul total reaction.

Each reaction mix (20ul) was converted to droplets with the QX200 droplet generator (Bio-Rad). Droplet-partitioned samples were then transferred to a 96-well plate, sealed and cycled in a C1000 deep well Thermalcycler (Bio-Rad) under the following cycling protocol: 95 °C for 5 min (DNA polymerase activation), followed by 50 cycles of 95 °C for 30 s (denaturation), 59 °C for 1 min (annealing) and 72°C for 30 s (extension) followed by post-cycling steps of 4°C for 5 min and 90 °C for 5 min (signal stabilization) and an infinite 12 degrees hold. The cycled plate was then transferred and read using the QX200 reader (Bio-Rad) either the same or the following day post-cycling. The concentration reported is copies/ul of the final 1x ddPCR reaction (using QuantaSoft software from Bio-Rad).

References:

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Brosseau JP, Lucier JF, Lapointe E, Durand M, Gendron D, Gervais-Bird J, Tremblay K, Perreault JP, Elela SA. 2010. High-throughput quantification of splicing isoforms. *RNA (New York, NY)* **16**: 442-449.

Chan PP, Lowe TM. 2016. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res* **44**: D184-189.

Deschamps-Francoeur G, Garneau D, Dupuis-Sandoval F, Roy A, Frappier M, Catala M, Couture S, Barbe-Marcoux M, Abou-Elela S, Scott MS. 2014. Identification of discrete classes of small nucleolar RNA featuring different ends and RNA binding protein dependency. *Nucleic Acids Res* **42**: 10073-10085.

Didion JP, Martin M, Collins FS. 2017. Atropos: specific, sensitive, and speedy trimming of sequencing reads. *PeerJ* **5**: e3720.

Dobin A, Gingeras TR. 2016. Optimizing RNA-Seq Mapping with STAR. *Methods Mol Biol* **1415**: 245-262.

Hellemans J, Mortier G, De Paepe A, Speleman F, Vandesompele J. 2007. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome biology* **8**: R19.

Klinck R, Chabot B, Abou Elela S. 2012. *High-Throughput Analysis of Alternative Splicing by RT-PCR*. Wiley.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923-930.

Nottingham RM, Wu DC, Qin Y, Yao J, Hunicke-Smith S, Lambowitz AM. 2016. RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. *Rna* **22**: 597-613.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733-745.

Qin Y, Yao J, Wu DC, Nottingham RM, Mohr S, Hunicke-Smith S, Lambowitz AM. 2016. High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. *Rna* **22**: 111-128.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L et al. 2016. Ensembl 2016. *Nucleic Acids Res* **44**: D710-716.