Supplemental Material for:

Sno-derived RNAs are prevalent molecular markers of cancer immunity

Ryan D. Chow, AB [1,2,3] and Sidi Chen, PhD [1,2,3,4,5,6,7,8,#]

Affiliations

[1]Department of Genetics,
Yale University School of Medicine, New Haven, Connecticut, USA.

[2]System Biology Institute,
Yale University School of Medicine, West Haven, Connecticut, USA.

[3]Medical Scientist Training Program,
Yale University School of Medicine, New Haven, Connecticut, USA.

[4]Biological and Biomedical Sciences Program,
Yale University School of Medicine, New Haven, Connecticut, USA.

[5]Immunobiology Program,
Yale University School of Medicine, New Haven, Connecticut, USA.

[6]Comprehensive Cancer Center,
Yale University School of Medicine, New Haven, Connecticut, USA.

[7]Stem Cell Center,
Yale University School of Medicine, New Haven, Connecticut, USA.

[8]Lead Contact

Correspondence:
SC (sidi.chen@yale.edu)
+1-203-737-3825 (office)
+1-203-737-4952 (lab)
Integrated Science & Technology Center
Yale University School of Medicine
850 West Campus Drive, Room 361, West Haven, CT 06516, USA

Supplemental Results

DNA level copy number alterations of the snoRNAome

Copy number variation is a major factor in the genetic alterations driving cancer. Considering our findings that sdRNAs are correlated with numerous features of cancer immunity, we wondered which of the parental snoRNAs, if any, are recurrently subject to copy number amplification or deletion in human cancers. By analyzing copy number data from TCGA, we focused on genomic regions of significant amplification or deletion (GISTIC 2.0 $q < 0.05$) and intersected these loci with the coordinates of each snoRNA (Figure S7a-b, Tables S16-17). Of note, *snoID_0379* and *SNORD69* were found to be significantly amplified ($q = 1.55 * 10^{-7}$) and deleted in lower grade glioma ($q = 8.92 * 10^{-4}$), respectively; these snoRNAs are 2 of 8 total snoRNAs with an ImmuneSurv score of 4 in lower grade gliomas. *SnoID_0379* expression was positively correlated with *PD-L1*, negatively correlated with intratumoral CD8+ T cell abundance, and associated with poorer survival. Given that patients with high *snoID_0379* expression have poorer survival, it is logical that amplification of *snoID_0379* would be positively selected for in lower grade glioma. In the complete opposite direction, *SNORD69* expression was negatively correlated with *PD-L1* expression, positively correlated with CD8+ T cell abundance, and associated with better survival. In stark contrast to *snoID_0379*, *SNORD69* was instead found to be recurrently deleted in lower grade gliomas, which is consistent with the mirrored directionality of its ImmuneSurv associations. It is therefore plausible that *snoID_0379* gain and *SNORD69* loss are selected for in lower grade glioma by influencing the tumor-immune microenvironment.

SdRNAs distinguish primary tumors and metastases in melanoma

Understanding the progression of cancer towards metastatic disease is critical for improving patient outcomes. As our prior results indicated that sdRNA expression signatures can be utilized to parse out different types and subtypes of cancers, we wondered if certain sdRNA expression patterns are specifically associated with metastasis. In the TCGA dataset, metastatic melanoma (SKCM) samples are well represented (n = 353). We compared sdRNA expression in metastatic samples to primary tumors (n = 97),

and identified 68 differentially expressed sdRNAs (adjusted $p < 0.05$) (Figure S10a, Table S6). sdRNAs that were downregulated in metastases include those derived from *ZL23* ($p = 6.96 * 10^{-8}$), *SNORD173* ($p = 4.68 * 10^{-6}$), and *ZL7* ($p = 1.38 * 10^{-6}$), while the top upregulated sdRNA is derived from *SNORD30* ($p = 9.47 * 10^{-6}$) (Figure S10b). These data indicate that a host of sdRNAs are differentially expressed in metastatic melanoma compared to primary tumors, pointing to sdRNAs as potential players in the metastatic progression.

Reanalysis of independent smRNA-seq datasets confirms dynamic expression of sdRNAs in cancer

By reanalyzing independent smRNA-seq datasets, we further confirmed the expression of sdRNAs in lung cancer (GSE33858; Figure S11a-c), colon cancer [64] (GSE46622; Figure S11d-e), and pancreatic cancer [65] (E-MTAB-3494; Figure S11f-h). These smRNA-seq libraries were size-selected for 17-27nt, 18-30nt, and < 40nt RNA species, which are approximately consistent with the TCGA smRNA-seq datasets and would be expected to encompass sdRNAs but not full-length snoRNAs. Importantly, we again confirmed the 5' or 3' bias of smRNA-seq reads that mapped to snoRNAs, consistent with the asymmetric processing of snoRNAs into sdRNAs. Additionally, the expression of sdRNAs was sufficient to partition different subtypes of lung cancer (Figure S11c), while also distinguishing normal pancreatic tissue from pancreatic adenocarcinomas (Figure S11g). In aggregate, the reanalysis of these smRNA-seq datasets validated the widespread and dynamic expression of sdRNAs in multiple human cancers.

Supplemental Figure Legends

Figure S1: Construction and characterization of the pan-cancer sdRNA atlas

a. Snapshot of smRNA-seq reads from TCGA-C8-A26V-01A mapping to several snoRNAs encoded within the human SNHG1 locus. The mapped reads demonstrate a 5' or 3' bias, indicative of sdRNAs.

b. Snapshot of smRNA-seq reads from TCGA-EA-A3HS-01A mapping to SNORD60 within the human SNHG19 locus. The mapped reads demonstrate a 5' or 3' bias, indicative of sdRNAs.

c. Scatterplot detailing the distribution of median sdRNA expression in 32 different cancer types (total n = 10,262 tumor samples), grouped by parental snoRNA. Values shown are in terms of $\log_2$ transcripts per million (tpm), with the median value shown for each parental snoRNA (n = 942 snoRNAs).

Figure S2: Lengths of reads mapping to snoRNA loci across cancer types

Bar plots detailing the lengths of reads mapping to snoRNA loci in the TCGA smRNA-seq datasets. Data are expressed as percentages of total mapped reads to snoRNA loci. These read lengths are consistent with the expected size range of sdRNAs.

Figure S3: Distributions of reads mapping to C/D snoRNAs across cancer types

Average profiles and heat maps of the mapped read distributions from all expressed C/D snoRNAs across 32 cancer types (the plots of KICH, LGG and OV from Figure 1g are redisplayed here). The read distributions generally clustered into three groups (k1, k2, k3). Values shown are normalized to maximum read depth for each snoRNA.

Figure S4: Distributions of reads mapping to H/ACA snoRNAs across cancer types

Average profiles and heat maps of the mapped read distributions from all expressed H/ACA snoRNAs across 32 cancer types (the plots of ACC, BRCA and UCS from Figure 1h are redisplayed here). The read distributions generally clustered into three groups (k1, k2, k3). Values shown are normalized to maximum read depth for each snoRNA.

Figure S5: Correlation analysis of high-variance sdRNAs reveals clusters of co-expression modules

a. Heat map of the standard deviation in expression for each sdRNA within individual cancer types, grouped by parental snoRNA. Even within a cancer type, sdRNA expression signatures were highly dynamic (n = 942 parental snoRNAs).

b. Heat map of pairwise Pearson correlation coefficients for high variance sdRNAs (n = 300). Right panel, the median $\log_2$ tpm expression across all tumors for each snoRNA. Examples of sdRNA clusters are outlined by boxes in the heat map.

Figure S6: t-SNE rendering of patient populations based on sdRNA expression signatures across individual cancer types

t-SNE plots of sdRNA expression in tumors from 32 different cancer types (total n = 10,262), grouped by parental snoRNA. t-SNE dimensions are the same as shown in Figure 2, but now each cancer type is visualized separately. These data further demonstrate the heterogeneity of sdRNA expression both across and within cancer types.

Figure S7: Additional t-SNE analysis

a. t-SNE plot of normal adjacent tissues (n = 675 samples). Samples are color-coded according to the adjacent cancer type.

b. t-SNE plot of cancers derived from the gastrointestinal tract, lung, kidney, and melanocytes similar to Figure 3b, with the difference that here the samples are color-coded according to histologic cancer type rather than the tissue of origin.

c-f. t-SNE plots of sdRNA expression in tumors from 32 different cancer types (total n = 10,262), colored by PD-L1 expression (**c**), GZMA expression (**d**), endothelial cell abundance (**e**), and patient survival (**f**). For patient survival, to avoid issues with censored data, only patients who had deceased were considered.

Figure S8: sdRNAs are correlated with tumor vascularization across multiple cancers

a. Heat map of sdRNAs positively correlated with endothelial cell abundance (EndothelialScore) (adjusted $p < 0.05$, adjusted within each cancer type), as determined by the xCell deconvolution algorithm. For visibility, only sdRNAs that were positively correlated in four or more cancer types are shown. Boxes are colored according to the Spearman correlation with EndothelialScore. Parental snoRNAs without annotated names are instead labeled by their host gene in parentheses. SnoRNA classifications are annotated on top based on a color legend on the right panel.

b. Scatter plots depicting the correlation between EndothelialScore and *SNORD114-1* sdRNA expression in breast adenocarcinoma (BRCA, n = 965), colorectal adenocarcinoma (COAD, n = 265), head and neck squamous cell carcinoma (HNSC, n = 435), sarcoma (SARC, n = 230), skin cutaneous melanoma (n = 89), stomach adenocarcinomas (STAD, n = 335), thymomas (THYM, n = 111), and uterine corpus endometrial carcinoma (UCEC, n = 159). Spearman correlation coefficients and associated p-values are noted on each plot. *SNORD114-1* sdRNA abundances are shown as transcripts per million (tpm), while EndothelialScores were determined by the xCell algorithm applied to the RNA-seq data.

c. Bar plot depicting the number of significant sdRNAs in each cancer type, in relation to EndothelialScore. Red, positive correlation; blue, negative correlation.

Figure S9: Pan-cancer copy number variation in snoRNAs

a. Heat map of snoRNAs that are recurrently amplified in at least 2 cancer types ($q < 0.05$). Cells are colored by $-\log_{10}$ q-values. Right panel, bar plot detailing the number of cancer types in which a given snoRNA is significantly amplified.

b. Heat map of snoRNAs that are recurrently deleted in at least 3 cancer types ($q < 0.05$). Cells are colored by $-\log_{10}$ q-values. Right panel, bar plot detailing the number of cancer types in which a given snoRNA is significantly deleted.

Figure S10: Differential expression of sdRNAs in metastases and primary melanomas

a. Volcano plot detailing log fold change and -$\log_{10}$ adjusted p-values for sdRNAs in metastases compared to primary melanomas, grouped by parental snoRNA. SdRNAs more highly expressed in metastases are colored red, while sdRNAs more highly expressed in primary melanomas are colored blue.

b. Box plots comparing expression of *ZL23, SNORD173, ZL7*, and *SNORD30* in metastases (red) compared to primary melanomas (blue). Associated p-values are indicated on the plot. Data are shown as $\log_2$ tpm.

Figure S11: Analysis of independent smRNA-seq datasets confirms expression of sdRNAs in cancer

a. Snapshot of smRNA-seq reads from a lung cancer (Sample 159T; GSE33858) mapping to *SNORD43* encoded within the *RPL3* locus. The mapped reads demonstrate a 5' bias, indicative of sdRNAs.

b. Violin plot of the top 30 sdRNAs with the highest variance across the GSE33858 dataset. Data are shown in terms of $\log_2$ tpm.

c. Principal component analysis of lung adenocarcinoma, lung squamous carcinoma, and lung adenosquamous carcinoma samples in GSE33858, based on sdRNA expression.

d. Snapshot of smRNA-seq reads from a colon cancer (Sample P3met; GSE46622) mapping to *SNORD98* encoded within the *CCAR1* locus. The mapped reads demonstrate a 3' bias, indicative of sdRNAs.

e. Violin plot of the top 10 sdRNAs with the highest variance across the GSE46622 dataset. Data are shown in terms of $\log_2$ tpm.

f. Snapshot of smRNA-seq reads from a pancreatic cancer (Sample P4; E-MTAB-3494) mapping to *SNORD26* encoded within the *SNHG1* locus. The mapped reads demonstrate a 3' bias, indicative of sdRNAs.

g. Principal component analysis of pancreatic tumors and normal pancreas in E-MTAB-3494, based on sdRNA expression.

h. Violin plot of the top 10 sdRNAs with the highest variance across the E-MTAB-3494 dataset. Data are shown in terms of $\log_2$ tpm.

List of supplemental tables

Cancer type abbreviations

ACC	Adrenocortical carcinoma
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Diffuse large B-cell lymphoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and neck squamous cell carcinoma
KICH	Kidney chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute myeloid leukemia
LGG	Brain lower grade glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin cutaneous melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular germ cell tumors
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine corpus endometrial carcinoma
UCS	Uterine carcinosarcoma
UVM	Uveal melanoma