AWARD NUMBER: W81XWH-13-1-0088

TITLE: Exploring the Presence of microDNAs in Prostate Cancer Cell Lines, Tissue, and Sera of Prostate Cancer Patients and its Possible Application as Biomarker

PRINCIPAL INVESTIGATOR: Pankaj Kumar

CONTRACTING ORGANIZATION: University of Virginia
CHARLOTTESVILLE, VA 22903

REPORT DATE: August 2015

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE August 2015	**2. REPORT TYPE** Annual	**3. DATES COVERED** 1 Aug 2014 – 31 Jul 2015

4. TITLE AND SUBTITLE

Exploring the Presence of microDNAs in Prostate Cancer Cell Lines, Tissue, and Sera of Prostate Cancer Patients and its Possible Application as Biomarker

5a. CONTRACT NUMBER W81XWH-13-1-0088

5b. GRANT NUMBER PC121591

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Pankaj Kumar

E-Mail: pk7z@virginia.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Virginia
Charlottsville, VA 22904-4195

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

MicroDNAs are extra chromosomal circular DNA present in normal mammalian somatic cells. To find the prostate tissue-specific microDNA a panel of human prostrate (LnCap (PSA, hK2 and AR positive), C4-2, and PC-3 (non-tranformed prostate epithelium)) and ovarian (ES2 and OVCAR-8) cancer cell lines were examined for microDNA. The identified microDNAs in all the cell lines were mostly ~200bp or ~400bp in size, arising from the GC rich regions in the genome, and mostly mapped to genic regions and are not associated with repetitive DNA sequences. MicroDNA are enriched in area of genome that had high exon density suggesting a role of splicing in microDNA generation. Comparison of microDNA loci across the cell lines identified hot spots of microDNA generation that are present on every chromosome, and correspond to areas of high gene density and high GC content. However, hierarchical clustering on the basis of microDNA co-ordinates classified the prostate and ovarian cancer cell lines into two separate groups suggesting that at least some microDNAs are tissue-specific and so their sites of origin are affected by tissue-specific gene expression patterns or epigenetic marks. The microDNA was also observed in mouse serum and cancer patient which suggest that microDNA could be surveyed for biomarker for cancer detection in future large study.

15. SUBJECT TERMS

microDNA; eccDNA; Prostate Cancer; Biomarker; Serum microDNA

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	46	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

INTRODUCTION:

Along with colleagues in lab, I have recently discovered a new type of extra-chromosomal circular DNA (eccDNAs, also called microDNA) in mouse tissue as well as in mouse and human cell lines (*1*). These eccDNAs are mostly 100-400 bases long, high in GC content, presence of micro homology at the start and end and arise from tens of thousands of unique genomic loci and could serve as disease biomarkers. Discovering a new biomarker for any cancer, in this case prostate cancer is significant because early detection and accurate prognosis is very important to cure the disease without over treating the many patients who do not have life-threatening disease. In this project I proposed potential of microDNA as cancer biomarker was explored. The circular DNAs are expected to be stable due to absence of free 5' or 3' end (resistant to exo-nuclease) and could be amplified by PCR based method. Finally microDNA were identified in various human cell lines and were compared with the prostate cancer cell lines. It would be great if we could detect microDNA in circulation and in this project microDNA were also isolated and identified in serum isolated from cancer patients.

KEYWORDS: eccDNA; microDNA; high-through put sequencing; prostate cancer; serum; biomarker

OVERALL PROJECT SUMMARY:

High-throughput sequencing of extra chromosomal circular DNA (eccDNA): Major Task 1 (1-8 months):

Isolation and high-throughput sequencing of eccDNA from prostate and non-prostate derived cell lines

Extraction of circular DNA: The steps involved in the isolation of circular DNA are shown in **Fig. 1**. In brief, the nuclei from the cells were extracted as described (*1*). To avoid contamination by mitochondrial DNA only the nuclei of the cell lines were used for the extraction of eccDNA (*1*). Contaminating linear DNA was removed by an ATP-dependent exonuclease (*1, 2*). Purified extra-chromosomal fraction was treated sequentially with proteinase K and RNase, with phenol-chloroform extraction and ethanol precipitation. Multiple displacement amplification (MDA) with random hexamers (*1, 3, 4*) was used to enrich circular DNA by rolling circle amplification. This procedure was applied to isolate eccDNA from three prostate cell lines: LNCaP (PSA, hK2 and AR positive), PC-3 & C4-2, (non-tranformed prostate epithelium) and two ovarian cancer cell lines (ES2 and OVCAR-8). The summary of isolation of microDNA in ovarian and prostate cell lines and its yield is shown in Table 1.

Table 1: Summary of microDNA isolation in various cancer cell lines.

Cell Line	ES2	OVCAR8	LnCap	PC-3	C4-2
Cell Count	$1.8X10^8$	$1x10^8$	$1.1X10^8$	$1.24X10^8$	$1.1X10^8$
Episomal DNA (ug)	21.3	23.7	26	15.6	20.4
Starting DNA (ug)	21.3	23.7	26	10	20
ExoVII (ng)	5600	9632	6074	2640	9352
ATP-dependent DNase (ng)	350	530	680	466	326
Rolling Circle Amplification (RCA) Starting DNA (ng)	88	133	120	116.5	81.5
RCA Ending DNA (ug)	9.2	4	8.368	6.8	8.515
DNA Shearing (400-600bp) (ng)	1472	1312	936	1116	1090
MicroDNA Library (ng)	402	423	738	528	1224
MicroDNA Library conc (ng/uL)	13.4	14.1	24.8	17.6	40.8

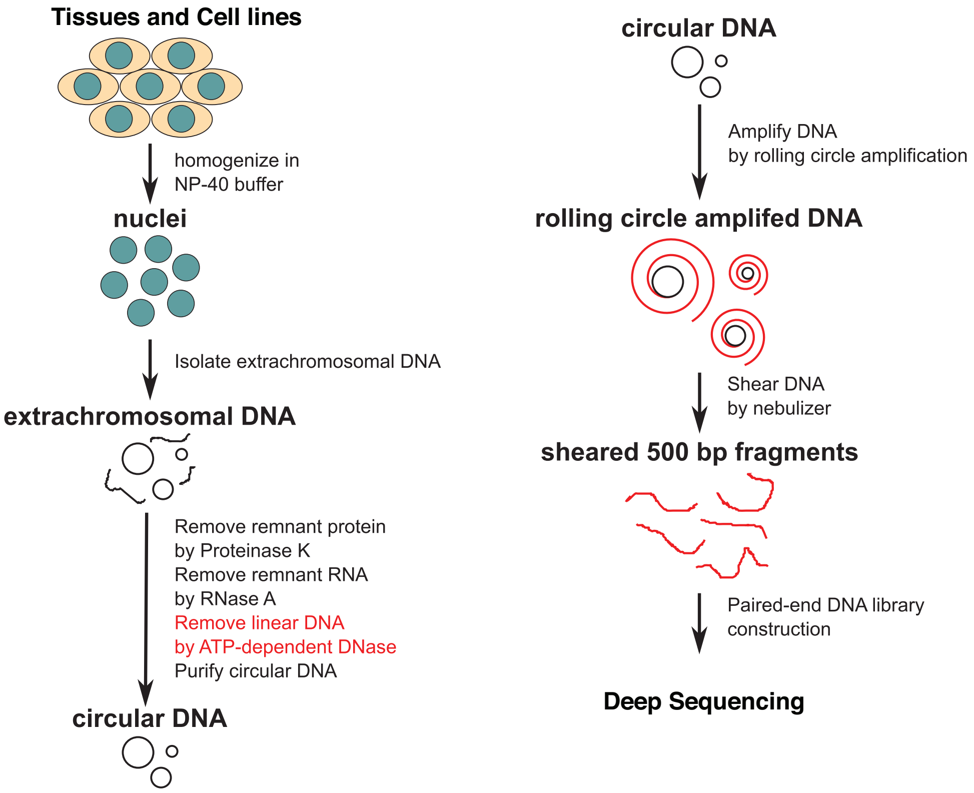

Tissues and Cell lines
homogenize in NP-40 buffer
nuclei
Isolate extrachromosomal DNA
extrachromosomal DNA
Remove remnant protein by Proteinase K
Remove remnant RNA by RNase A
Remove linear DNA by ATP-dependent DNase
Purify circular DNA
circular DNA
circular DNA
Amplify DNA by rolling circle amplification
rolling circle amplifed DNA
Shear DNA by nebulizer
sheared 500 bp fragments
Paired-end DNA library construction
Deep Sequencing

Figure 1: Illustration of method of circular DNA isolation and library preparation from various cell lines of prostate and ovarian tissue. ATP-dependent DNase-resistant DNA from nuclei (eccDNA) was amplified by multiple-displacement amplification (MDA). The amplified DNA was sheared to obtain 500 bp fragments and sequenced by the Illumina sequencing.

MicroDNA library preparation and sequencing: Enriched eccDNA was fragmented, selected and sequenced (Sanger sequencing) to verify the presence of circular DNA. Cloning and sequencing of 500 bps long MDA product confirmed circular nature of DNA (**Fig. 2**). Once circular nature of DNA was confirmed then paired-end (PE) library was prepared as described (*1*). The 500bp size selection was done on nebulized DNA. The ends of the library fragments were modified as per Illumina paired-end protocol and paired-end high-throughput sequencing (64 bases long reads) was performed according to the manufacturer's protocol (Illumina). Summary of microDNA sequencing and mapping in prostate and ovarian cancer cell lines is shown in **Table 2**.

Table 2: Summary of PE sequencing and mapping to human genome.

Sample	**Paired End Reads**	**Pairs Aligned**	**Read Sequences**	**Aligned Sequences**	**Unique Alignment**	**Unique microDNA (Complexity)**
ES2	61.9	26.8	123.9	96.4	86.7	**114,752**
OVCAR8	50.2	28.8	100.4	84.5	75.8	**57,327**
C4-2	41.1	21.4	82.2	69.3	63.2	**41,410**
LnCap	56.1	24.8	112.1	89.1	82.3	**84,841**
PC3	43.5	10.7	87.0	41.6	38.8	**14,705**

*all values in millions except microDNA

```
GCAGCACCATTTACAATGATGCTGCACATTAAATTCAACAGGGAGAAATCCTCTCTGCCCCTCAG
ACTGCCCATCAGGCTTGGGAGGTGTCGGGAGACAGGCGTTCATCCTGGTCGCTGCTTTGGGTAG
CAGCTTGCAGTGCTGAAACAGTCAAAGATGGCTGTCCCTCAGCCCTGCCACCTCCCATTCAAGCG
CCTGCTCTGAAAGCTCCTGAGCAGATGGGCCTGAGATGCAGACAGGGGTGCTCGTGGCAGCACC
ATTTACAATGATGCTGCACATTAAATTCAACAGGGAGAAATCCTCTCTGCCCCTCAGACTGCCCA
TCAGGCTTGGGAGGTGTCGGGAGACAGGCGTTCATCCTGGTCGCTGCTTTGGGTAGCAGCTTGC
AGTGCTGAAACAGTCAAAGATGGCTGTCCCTC
```

Figure 2: Presence of repeat sequence in the 500 bps long cloned and Sanger sequenced fragment. Circular DNA sequence is in purple.

Major Task 2: Identification of circular DNA from various samples (9-15 months)

MicroDNA identification from the paired-end sequencing: The details of the different steps of identification of microDNA are shown in **Fig. 3**. Sequence tags were mapped on the human reference genome using the Novoalign software. Only those tags that were mapped uniquely were considered for the identification of circular DNA. The sequence coverage of each base pair was profiled for each chromosome. An island of interest (potential circle) was delineated on the basis of two consecutive sequenced bases. In other words any stretch of continuously sequenced bases was considered as a part of an island and the start and end of the stretch was considered as start and end of the island respectively. The islands were considered further for the identification of circles. The creation of circular microDNAs would bring together the ends of the linear islands to create a novel junctional sequence that does not exist in the genome. Thus the PE-sequence of a fragment that breaks at or very close to a junction will have one end that maps to the island and another end that maps to the junction and will not map to the reference genome (**Fig. 3b**). Those PE-tags where one tag maps uniquely to an island and the other remains unmapped, but passes the sequence quality filter, was considered for the validation of circular nature of the identified islands. As mentioned earlier, the creation of circular DNAs bring together the two ends of the linear DNA, and thus generated hypothetical junctional tags was created by ligation of the two ends of each island. If the mapped tag of a PE read falls in an island and the un-mapped tag matches a hypothetical junctional tag of the same island, then the island was annotated as a circle. Summary of microDNA sequencing and mapping and number of microDNA identified in prostate and ovarian cancer cell lines is given in **Table 2**.

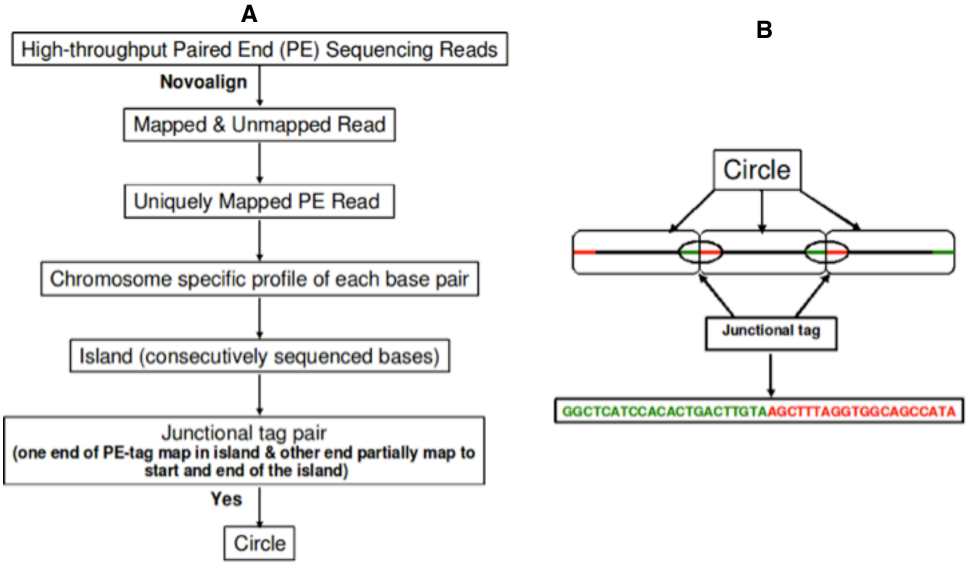

A
High-throughput Paired End (PE) Sequencing Reads
Novoalign
Mapped & Unmapped Read
Uniquely Mapped PE Read
Chromosome specific profile of each base pair
Island (consecutively sequenced bases)
Junctional tag pair
(one end of PE-tag map in island & other end partially map to start and end of the island)
Yes
Circle
B
Circle
Junctional tag
GGCTCATCCACACTGACTTGTAAGCTTTAGGTGGCAGCCATA

Figure 3: Illustration of different steps in the identification of microDNA by Island method (a) and schematic representation of junctional tag (b).

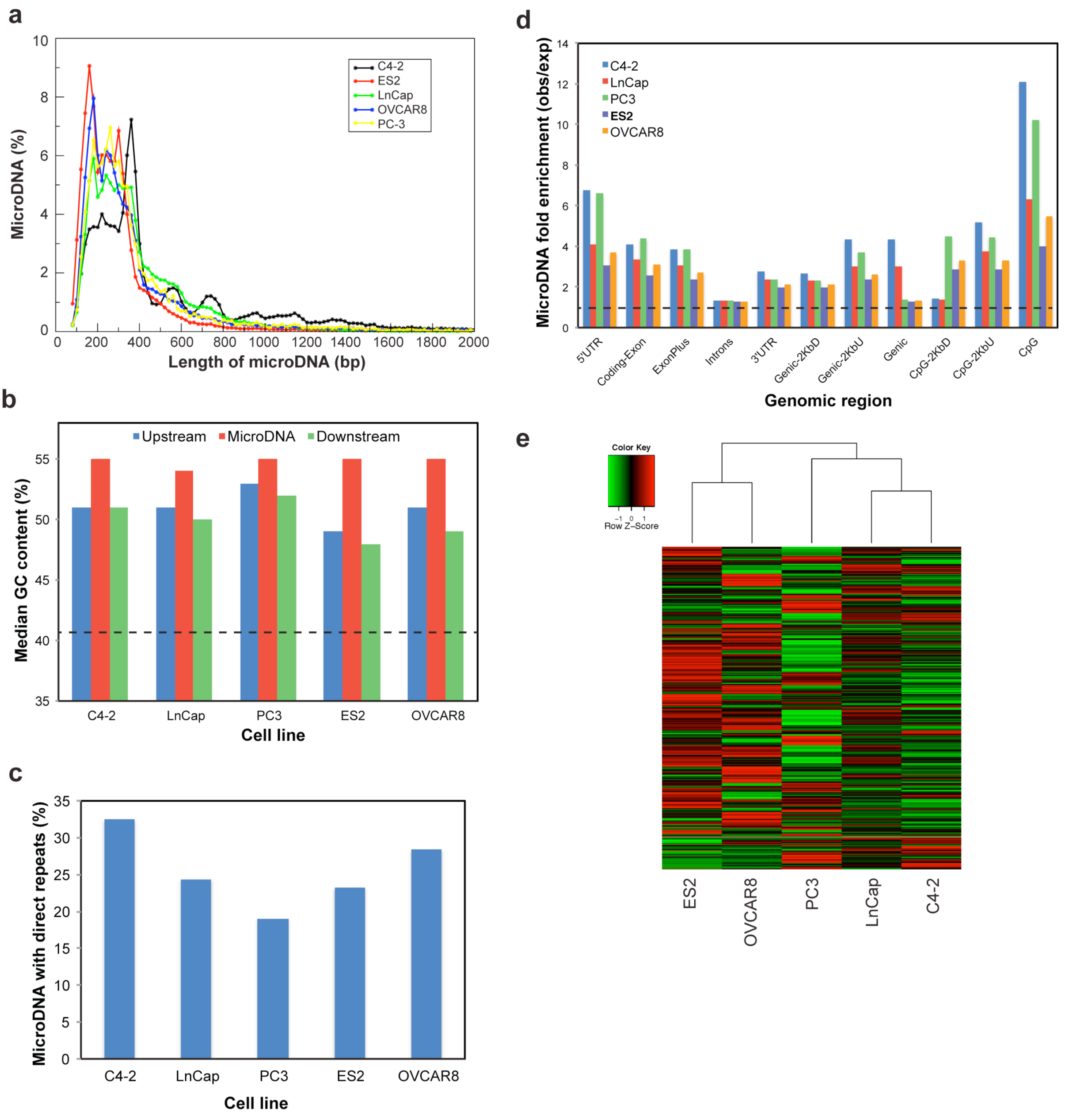


Figure 4: Properties of microDNA identified in human prostate and ovarian cancer cell lines. (a) Length distribution of microDNAs identified in cancer cell lines. (b) Median percent GC content of microDNAs and the genomic sequences up- or downstream of the source loci are enriched relative to the average GC content of the human genome (dashed line). (c) Direct repeats near the start and end of microDNA sequences (2- to 15-bp) are enriched in all cell lines compared to a random model (RM). (d) Enrichment of microDNAs in the indicated genomic region relative to the expected percentage based on random distribution. (e) MicroDNA loci were grouped into 5-Mb bins stepwise across the human genome and the percentage of all microDNA located within each bin was calculated for each cancer cell line and compared using hierarchical clustering.

Identification and analysis of prostate-tissue-specific and prostate-cancer-specific microDNAs:

First, all microDNAs were studied for general properties: size distribution (**Fig. 4a**), GC content (**Fig. 4b**), the presence of 2-10 base direct repeats at the ends of microDNA (**Fig. 4c**), and locations relative to genomic features (exons, introns, UTRs, CpG islands) (**Fig. 4d**). It could be seen that most of the microDNAs are of 200-400 bps long, have high GC content compared to the genomic average and frequently have direct repeat at the ends. These features are similar to the features that have all been observed in the microDNAs identified in normal mouse tissue, and mouse NIH3T3 and human HeLa cells (*1*). This confirms that the microDNAs obtained from normal tissue and human cell lines of different tissue origin conforms to the general properties of microDNAs.

The genomic origins of the circles and the abundance of circles from each locus were compared by hierarchical clustering. For this whole human genome was divided in 5-mega base windows and in each bin the fraction of microDNA in each bin was calculated and compared using hierarchical clustering. It is interesting to note that prostate cancer cell lines are clustering together (**Fig. 4e**) and distinct from the ovarian cell line cluster indicating that some of the genomic loci are differentially producing microDNA between prostate and ovarian cell line. The common and abundant circles identified across all the prostate cancer cell lines have the potential to qualify as a marker for prostate cancer however this tissue specificity of microDNA need to be further checked in patient sera.

Characterization of microDNA across a panel of adult mouse tissues:

MicroDNA was isolated from a battery of mouse tissues (brain, heart, kidney, liver, lung, skeletal muscle, spleen, sperm, testis and thymus) (*5*). EccDNA sequences were then enriched by multiple displacement amplification (MDA) using random primers, and the rolling-circle amplification products were converted to 500-bp long fragments for paired-end sequencing.

The sequences generating the microDNAs map mostly to unique sequences in the mouse genome and are not extensively derived from repetitive elements. Thus, microDNAs are generated universally across all tissue types and high GC content encourages their generation and the presence of short direct repeats flanking the segment that forms the circle.

Next we analyzed the genomic regions that commonly generate microDNAs and how they compare between tissue types. Interestingly, when each chromosome is divided into 1-Mb windows and the average GC content, gene density or percentage of microDNA per Mb is calculated, there is a positive correlation of microDNA density with GC content (R^2 = 0.86) and gene density (R^2 = 0.69), indicating a non-random distribution of microDNA loci throughout the genome. This is strikingly visualized when each chromosome is divided into 1-Mb windows and the percentage of unique microDNA located within each window is plotted. For example, on chromosome 10 four large "hot-spots" of microDNA generation can be identified that overlap between all the tissue types (Figure 5). These "hot-spots" of microDNA correlate with regions of high GC content and gene density (Figure 5).

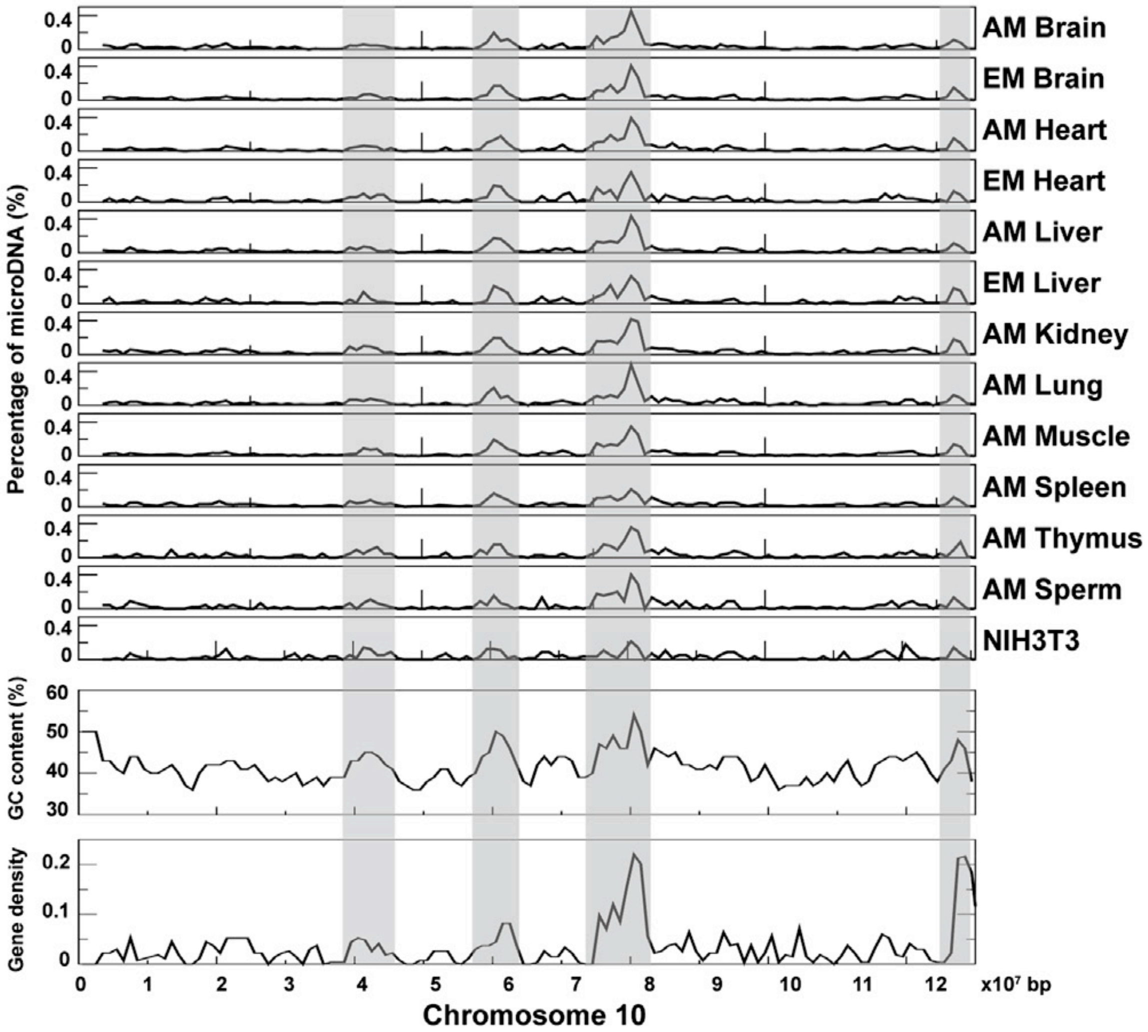


Figure 5. Distribution of microDNA along chromosomes is conserved across tissue types. MicroDNA loci were grouped into bins of 1-Mb stepwise across mouse chromosome 10 and the percentage of all microDNA located within each bin was calculated for each tissue type. MicroDNA clustering patterns are similar across all tissue types and "hotspot" regions (grey bars) correlate with a high GC content and gene density.

Identification of disease specific eccDNA: Major Task 16-24 months

Isolation and high throughput sequencing of eccDNA from mouse serum: To check if microDNA are present in circulation microDNA was isolated from mouse serum. MicroDNAs are present in the circulation and can be isolated and sequenced from mouse and human serum. Furthermore, in mice containing human-derived prostate xenograft tumors, we were able to detect human-derived microDNAs, indicating that microDNAs can be released from tumors or tissues into the bloodstream. The presence of microDNAs in circulation was very interesting and it supported the present proposal that microDNA could serve as a novel biomarker for cancer. The mapping summary of high throughput sequencing reads and microDNA identified in two different conditions are given in Table 3. The length distribution and GC content of microDNA identified in serum sample was similar to our previous observation in normal tissues and cell lines.

Table 3: Sequencing summary of serum-derived microDNAs

Serum sample	**Normal mouse 1**	**Normal mouse 2**	**C4-2 xenograft mouse 1**		**C4-2 xenograft mouse 2**	
Genome	mouse	mouse	mouse	human	mouse	human
Paired end reads	23.4	42.3	51.2		62.7	
Pairs aligned	2.2	12.3	10.5		14.2	
Read sequences	46.8	84.6	102.3		125.3	
Reads aligned	8.2	42.9	39.8	1.5	50	29.2
Unique alignment	7.4	37.6	35.4	1.5	46.4	28.2
MicroDNA*	1,606	5,091	7,463	464	4,761	4,929

*all values in the millions except for microDNA

In the final level of my study microDNA was isolated from sera of ovarian and lung cancer patients (I was fortunate to find the serum sample from cancer patients). It would have been great to include and compare the serum microDNA from prostate cancer patient but unfortunately the samples were not available in our local repository, hence study was done in ovarian and lung cancer patients. This study was done mainly to find microDNA in cancer patient. The isolation and high throughput sequencing of isolated microDNA was done described earlier. It was motivating to find the microDNA in serum of cancer patient which itself was a novel discovery. The length distribution, GC content of microDNA identified in serum samples was similar to our previous observation in mouse tissues and human cell lines.

Completion on analysis and preparation of manuscripts to report the results: Major Task 16-24 months

I published three first author articles during this award period. One is directly related to proposed work. The reference of publications is given below.

Dillon LW, **Kumar P**, Shibata Y, Wang YH, Willcox S, Griffith JD, Pommier Y, Takeda S, Dutta A: **Production of Extrachromosomal MicroDNAs Is Linked to Mismatch Repair Pathways and Transcriptional Activity**. *Cell Rep* 2015, **11**(11):1749-1759.

KEY RESEARCH ACCOMPLISHMENTS:

- MicroDNAs are present in cancer cell lines
- MicroDNAs identified in cancer cell lines have similar features (length distribution, GC content, genomic enrichment etc.) that have been observed in the microDNAs identified in normal mouse tissue, and mouse NIH3T3 and human HeLa cells
- Hierarchical clustering of prostate and ovarian cancer cell lines based on the microDNA loci in the genome indicates some of the microDNAs are tissue specific
- microDNA are present in mouse serum and therefore it can be explored for disease biomarker
- microDNA are present in serum of cancer patient
- The cancer specific microDNA could be identified by further extending this study at large sample size including the microDNA analysis in various human cancer type

CONCLUSION:

To find the tissue specific microDNA we examined a panel of human prostrate (C4-2, LnCap and PC-3) and ovarian (ES2 and OVCAR-8) cancer cell lines. Hierarchical clustering on the basis of microDNA co-ordinates classified the prostate and ovarian cancer cell lines into two separate groups suggesting that microDNA are tissue specific. The tissue specificity of these microDNA could be further explored to find prostate tumor specific microDNAs that could serve as biomarkers for cancer detection and its prognosis. DNA, especially circular DNA, is extremely stable and is also expected to survive in the blood once it is released from cancer cells. The other important finding of this study was the presence of microDNA in the serum of mouse and cancer patient. Even the identification of microDNAs in serum was a novel discovery. The preliminary data from this part of the project are critical to propose a more definitive project on these lines.

PUBLICATIONS, ABSTRACTS, AND PRESENTATIONS:

a. List all manuscripts submitted for publication during the period covered by this report resulting from this project.

1. Lay Press: "Nothing to report."
2. Peer-Reviewed Scientific Journals:

> Dillon LW, **Kumar P**, Shibata Y, Wang YH, Willcox S, Griffith JD, Pommier Y, Takeda S, Dutta A: **Production of Extrachromosomal MicroDNAs Is Linked to Mismatch Repair Pathways and Transcriptional Activity**. *Cell Rep* 2015, **11**(11):1749-1759.
>
> **Kumar P**, Anaya J, Mudunuri SB, Dutta A: **Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets**. *BMC Biol* 2014, **12**:78.
>
> **Kumar P**, Mudunuri SB, Anaya J, Dutta A: **tRFdb: a database for transfer RNA fragments**. *Nucleic Acids Res* 2015, **43**(Database issue):D141-145.

Note: Department of Defense fellowship is fully acknowledged in the all the manuscript published by PI and is attached in appendix.

3. Invited Articles: "Nothing to report."
4. **Abstracts:**

P Kumar, Laura W. Dillon, Y Shibata, and A Dutta., **MicroDNA (Extra Chromosomal Circular DNA) in Normal Mammalian Tissues and Cancer Cell Lines**; (Accepted for poster presentation at the 64th Annual Meeting of The American Society of Human Genetics, October 20, 2014 in San Diego, CA.

P Kumar, Laura W. Dillon, Y Shibata, and A Dutta., **MicroDNA (Extrachromosomal Circular DNA) in Mammalian Tissues and Chicken & Human Cancer Cell Lines are generated by DNA repair pathways Cell Growth & Proliferation Gordon Research Conference** New Discoveries and Approaches to the Study of Cell Proliferation, with an Emphasis on Cancers **July 12-17, 2015, Mount Snow, West Dover, VT**

b. List presentations made during the last year (international, national, local societies, military meetings, etc.).
MicroDNA (Extrachromosomal Circular DNA) in Mammalian Tissues and Chicken & Human Cancer Cell Lines are generated by DNA repair pathways Cell Growth & Proliferation Gordon Research Conference New Discoveries and Approaches to the Study of Cell Proliferation, with an Emphasis on

Cancers **July 12-17, 2015, Mount Snow, West Dover, VT**

INVENTIONS, PATENTS AND LICENSES: "Nothing to report."

REPORTABLE OUTCOMES: "Nothing to report."

OTHER ACHIEVEMENTS: "Nothing to report."

REFERENCES:

1. Y. Shibata *et al.*, Extrachromosomal MicroDNAs and Chromosomal Microdeletions in Normal Tissues. *Science* **336**, 82-86 (2012).
2. H. Yamagishi *et al.*, Purification of Small Polydisperse Circular DNA of Eukaryotic Cells by Use of Atp-Dependent Deoxyribonuclease. *Gene* **26**, 317-321 (1983).
3. F. B. Dean *et al.*, Comprehensive human genome amplification using multiple displacement amplification. *P Natl Acad Sci USA* **99**, 5261-5266 (2002).
4. L. Lovmar, A. C. Syvanen, Multiple displacement amplification to create a long-lasting source of DNA for genetic studies. *Hum Mutat* **27**, 603-614 (2006).
5. L. W. Dillon *et al.*, Production of Extrachromosomal MicroDNAs Is Linked to Mismatch Repair Pathways and Transcriptional Activity. *Cell Rep* **11**, 1749-1759 (2015).

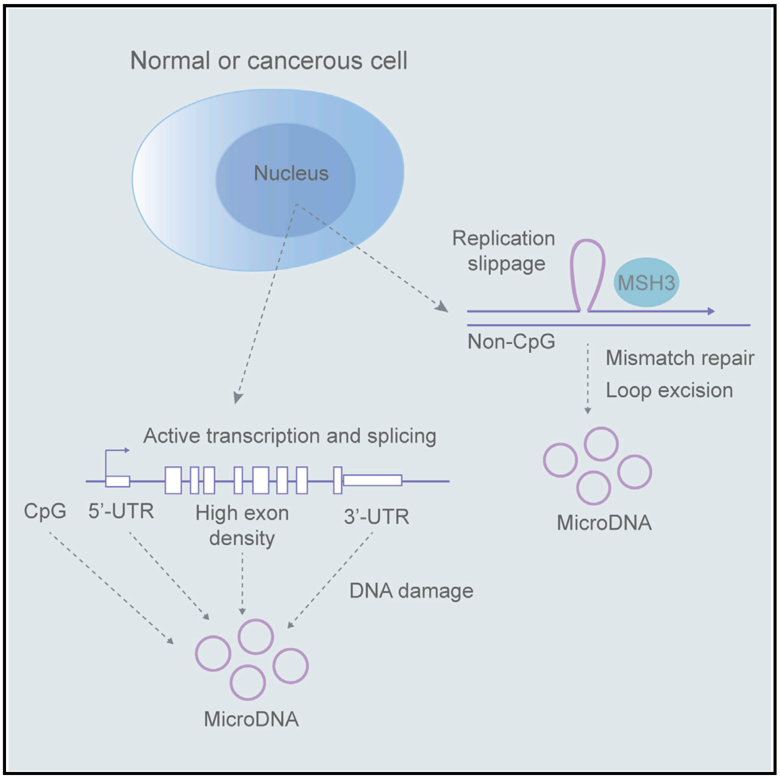
APPENDICES:

Cell Reports

Article

Production of Extrachromosomal MicroDNAs Is Linked to Mismatch Repair Pathways and Transcriptional Activity

Graphical Abstract



Authors

Laura W. Dillon, Pankaj Kumar,
Yoshiyuki Shibata, ..., Yves Pommier,
Shunichi Takeda, Anindya Dutta

Correspondence

ad8q@virginia.edu

In Brief

Through isolating and sequencing small extrachromosomal circular microDNAs across multiple species and cell types, Dillon et al. identify specific genomic features associated with the generation of microDNAs and link the DNA mismatch repair pathway to this process.

Highlights

- Extrachromosomal circular microDNAs are found in all tissue types
- MicroDNAs preferentially arise from genomic regions with active chromatin marks
- High GC content and gene or exon density are associated with microDNA production
- The mismatch repair pathway is linked to microDNA generation

Accession Numbers

GSE68644

Dillon et al., 2015, Cell Reports *11*, 1749–1759
June 23, 2015 ©2015 The Authors
http://dx.doi.org/10.1016/j.celrep.2015.05.020

Cell Reports
Article

OPEN ACCESS
CellPress

Production of Extrachromosomal MicroDNAs Is Linked to Mismatch Repair Pathways and Transcriptional Activity

Laura W. Dillon,[1,5] Pankaj Kumar,[1,5] Yoshiyuki Shibata,[1,5] Yuh-Hwa Wang,[1] Smaranda Willcox,[2] Jack D. Griffith,[2] Yves Pommier,[3] Shunichi Takeda,[4] and Anindya Dutta[1,*]

[1]Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, VA 22908, USA
[2]Lineberger Comprehensive Cancer Center, Department of Microbiology and Immunology, University of North Carolina, Chapel Hill, NC 27514, USA
[3]Laboratory of Molecular Pharmacology and Developmental Therapeutics Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892-4255, USA
[4]CREST Research Project, Japan Science and Technology Corporation, Radiation Genetics, Faculty of Medicine, Kyoto University, Konoe Yoshida, Sakyo-ku, Kyoto 606-8501, Japan
[5]Co-first author
*Correspondence: ad8q@virginia.edu

http://dx.doi.org/10.1016/j.celrep.2015.05.020

SUMMARY

MicroDNAs are <400-base extrachromosomal circles found in mammalian cells. Tens of thousands of microDNAs have been found in all tissue types, including sperm. MicroDNAs arise preferentially from areas with high gene density, GC content, and exon density from promoters with activating chromatin modifications and in sperm from the 5′-UTR of full-length LINE-1 elements, but are depleted from lamin-associated heterochromatin. Analysis of microDNAs from a set of human cancer cell lines revealed lineage-specific patterns of microDNA origins. A survey of microDNAs from chicken cells defective in various DNA repair proteins reveals that homologous recombination and non-homologous end joining repair pathways are not required for microDNA production. Deletion of the MSH3 DNA mismatch repair protein results in a significant decrease in microDNA abundance, specifically from non-CpG genomic regions. Thus, microDNAs arise as part of normal cellular physiology—either from DNA breaks associated with RNA metabolism or from replication slippage followed by mismatch repair.

INTRODUCTION

For a long time, eukaryotic genomes were considered to be stable and relatively conserved, but advances in genome technology have revealed genetic diversity between individuals, such as SNPs and copy-number variations (Beckmann et al., 2007; Flores et al., 2007; Frazer et al., 2009; Lupski, 2010; Stankiewicz and Lupski, 2010). Furthermore, evolution of an organism's genome occurs during its lifespan, resulting in genetic mosaicism among somatic cells. One such example of genomic variation is extrachromosomal circular DNA (eccDNA) (Cohen and Segal, 2009).

EccDNA is observed universally in eukaryotic genomes. Previous studies of eccDNA revealed them to be several hundred to millions of bases in length and to originate from viral genomes, intermediates of mobile elements, or repetitive chromosomal sequences (Cohen and Segal, 2009). Recently, we discovered a class of eccDNA, dubbed microDNAs, in mouse tissues and mouse and human cell lines that exhibits specific features that differ greatly from previously described eccDNA (Shibata et al., 2012). MicroDNAs are short (∼100–400 bp long), circular DNAs derived mostly from unique non-repetitive genomic sequences. They preferentially appear from genic regions, have a high GC content, and exhibit microhomology (2- to 15-bp direct repeats) at the ends of the sequences that circularize to form the microDNAs (Shibata et al., 2012). Our initial discovery of microDNA raised many important questions regarding this class of unusual nucleic acids, including the extent of their existence across all tissue types and the mechanism of their formation.

Because microDNAs are seen even in adult mouse brain, which has low levels of cell proliferation, one possibility is that microDNAs are generated by some kind of repair process arising from DNA damage that occurs in quiescent cells. We hypothesized that an exhaustive examination of various tissues and cell lines with mutations in select DNA repair pathways would allow us to resolve the types of DNA damage and repair pathways involved in the production of microDNAs.

In this report, we characterize features of microDNA across a panel of tissues from normal adult mice. We find that microDNAs are present in all tissue types examined, including germ cells (sperm), and there is very little correlation with the extent of cell proliferation. The microDNAs arise preferentially from regions of the genome with very specific characteristics: a high GC content, gene density, and exon density. Furthermore, microDNAs are highly enriched from promoters with activating chromatin modifications and areas of the genome associated with RNA polymerase II, but depleted in inactive lamin-associated heterochromatin. The preferential production of microDNAs from genomic windows with high exon density and from the extreme 5′ ends of full-length LINE-1 retrotransposon elements suggests that areas with a propensity to form RNA-DNA hybrids, especially near DNA breaks, can lead to the kind of damage that produces microDNAs. Because of the large number of sites in the genome that give rise to microDNAs (complexity), there most likely exists a copying mechanism that produces excess DNA, which is removed as microDNAs without leaving corresponding deletions in the genomic DNA.

A striking feature of microDNAs is the frequent presence of short direct repeats of 2–15 bases at the beginning and end of the genomic sequence that gives rise to the microDNA, leading us to test whether homology-dependent repair pathways are important for microDNA generation. An analysis of cell lines deficient in various DNA repair proteins reveals that no singular DNA repair pathway is responsible for microDNA production. Most likely, if double-strand breaks occur, redundant pathways using homologous recombination (HR) or nonhomologous end-joining (NHEJ) and microhomology-mediated end-joining (MMEJ) contribute to the generation of microDNAs. Short direct repeats in the genome are also known to be sites of replication slippage that give rise to a loop of DNA in the product or template strand during DNA replication or repair, which is usually corrected by the mismatch repair (MMR) pathway (Schofield and Hsieh, 2003). Strikingly, mutation in MMR significantly decreases the abundance of microDNAs and alters the distribution of genomic sites producing the residual microDNAs, suggesting that a significant fraction of the microDNAs is generated by replication slippage and the MMR pathway. In summary, the unexpectedly ubiquitous and abundant microDNAs are the products of break repair or MMR following DNA damage associated with transcription or splicing.

RESULTS

Characterization of MicroDNA across a Panel of Adult Mouse Tissues

To determine whether there is any normal tissue that is bereft of microDNAs, microDNA was isolated from a battery of tissues (including brain, heart, kidney, liver, lung, skeletal muscle, spleen, sperm, testis, and thymus) by first purifying extrachromosomal DNA from the nuclei of homogenized tissues from normal adult C57BL/6 mice followed by removal of linear DNA by digestion with exonucleases. Using electron microscopy, the presence of both double- and single-stranded microDNA in the remaining eccDNA was confirmed (Figure 1A). EccDNA sequences were then enriched by multiple displacement amplification (MDA) using random primers, and the rolling-circle amplification products were converted to 500-bp long fragments for paired-end sequencing. Paired ends where a genomic sequence is paired with an unmapped sequence that does not map anywhere in genome were indexed (Shibata et al., 2012). If the unmapped sequence could be explained as a junctional sequence created by the circularization of the neighboring linear genomic DNA, the sequence was recognized as deriving from a microDNA. MicroDNAs were observed in every tissue type examined with sequences originating from tens of thousands of unique loci within the mouse genome (Table 1).

MicroDNA from the mouse tissues have features similar to those described in our initial publication (Shibata et al., 2012). The lengths range from 60 to 2,000 bp, with the majority (≥84%) between 100 and 400 bp (Figure 1B). The sequences generating the microDNAs map mostly to unique sequences in the mouse genome and are not extensively derived from repetitive elements. In all tissues, the microDNA sequences are significantly more GC-rich than the genomic average (Figure 1C). The sequences directly flanking the starts and ends of the microDNA have a significant enrichment in 2 to 15-bp direct repeats of homology compared with a random model (Figure 1D). Furthermore, the sources of microDNA are highly enriched in genic regions, especially 5′-UTRs of genes, exons, and CpG islands (Figure 1E). Within genes, microDNAs originate more often from the 5′ or 3′ ends than the main body of the gene (Figure S1). Thus, microDNAs are generated universally across all tissue types, and their generation is encouraged by high GC content and the presence of short direct repeats flanking the segment that forms the circle.

MicroDNAs Overlap with Repetitive Elements in Mouse Tissues

While microDNAs map uniquely to the genome, we also wanted to investigate differences in microDNA originating from repetitive elements. Therefore, we compared the percentage of uniquely mapped microDNAs from each tissue type that originate from the four major classes of repetitive elements: LINEs (long interspersed nuclear element), SINEs (short interspersed nuclear element), LTRs (long terminal repeat), and repetitive DNA elements, as defined by RepeatMasker (Smit et al., 1996–2010). Approximately 40%–50% of microDNAs map to repetitive elements, consistent with the fraction of the genome covered by such elements, suggesting that microDNAs are not preferentially enriched from repetitive elements. MicroDNAs originate nearly equally from SINE, LTR, and DNA elements in all tissue types, with the exception that sperm microDNAs are enriched ∼2-fold from LINE elements (Figure 2A). Upon further analysis, we found this enrichment is almost entirely due to microDNAs from full-length LINE-1 retrotransposons (L1) (Penzkofer et al., 2005), accounting for 5% of all microDNAs in sperm (Figure S2A). Specifically, sperm microDNAs are highly enriched in full-length L1 elements of the L1Md_T class (26.5-fold over random expectation) (Figures 2B and S2B). Additional tissues (liver, lung, testis, thymus, and embryonic mouse brain) also exhibited a significant enrichment in this full-length L1 element, but to a lesser extent than sperm (Figures 2B and S2). Previously, full-length L1 transcripts and L1-encoded proteins have been detected in prepubertal mouse spermatocytes (Branciforte and Martin, 1994). The putatively active mouse L1 element is >7 kb long and is composed of the 5′-UTR, an internal CpG-rich promoter, two open reading frames (ORF1 and ORF2), and a 3′-UTR including a poly(A) tail (Ostertag and Kazazian, 2001). The length of the mouse L1 5′-UTR element can differ due to varying tandem repeats of ∼200-bp monomers. Interestingly, we found that 95% of all sperm microDNAs originating from L1 elements map to the 5′-UTR and almost exclusively to the monomer repeat sequences (Figures 2C and 2D). MicroDNAs from other mouse

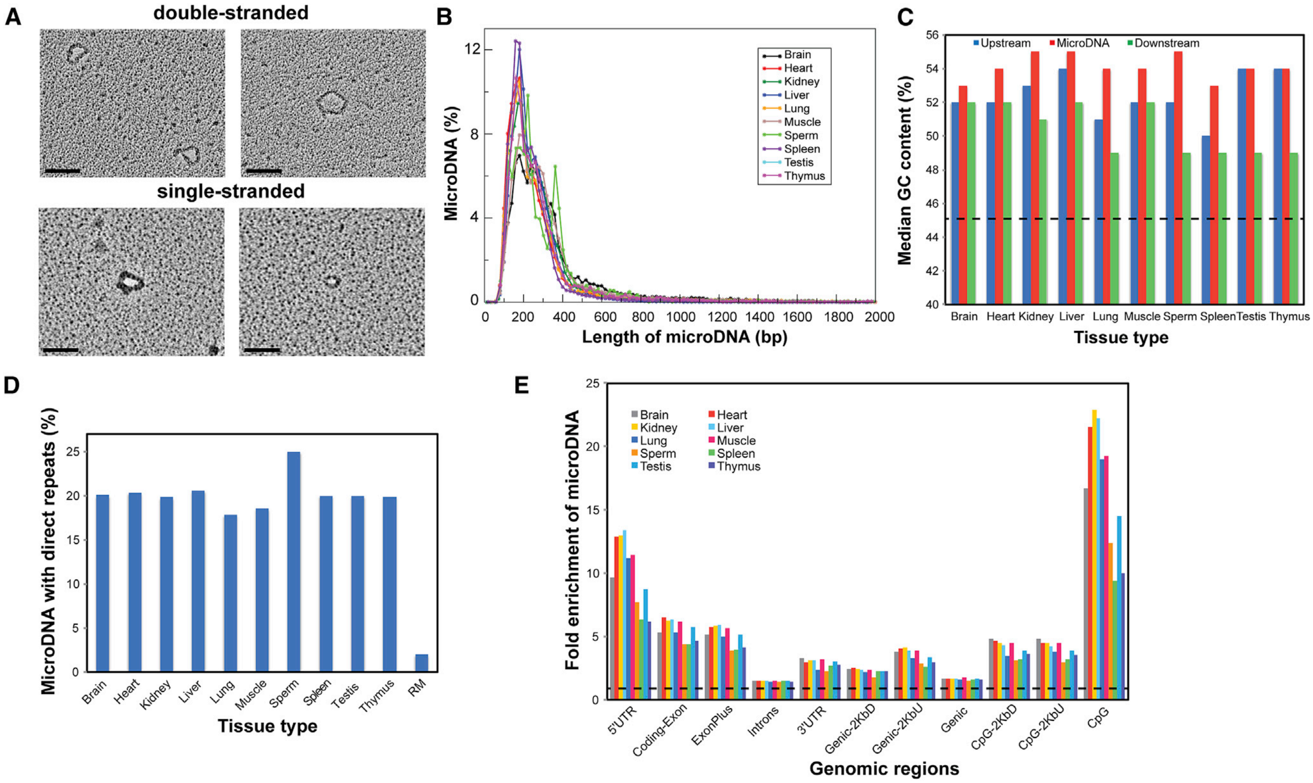


Figure 1. Properties of MicroDNAs in Normal Adult Mouse Tissues

(A) EM of double-stranded microDNA from adult mouse kidney tissue and single-stranded microDNA from spleen tissue after binding with the T4 gene 32 single-stranded DNA binding protein. Black scale bar represents 100 nm.

(B) Length distribution of microDNAs identified in adult mouse tissues.

(C) Median percent GC content of microDNAs and the genomic sequences upstream or downstream of the source loci are enriched relative to the average GC content of the mouse genome (dashed line).

(D) Direct repeats near the start and end of microDNA sequences (2- to 15-bp) are enriched in all tissues compared with a random model (RM).

(E) Enrichment of microDNAs in the indicated genomic region relative to the expected percentage based on random distribution. The black dashed line at 1 indicates the randomly expected level.

See also Figure S1.

tissues that originate from L1 elements also appear primarily from the 5′-UTR element of the full-length L1 elements (Figure 2C). Since an intermediate structure during L1 transposition has the newly transposed L1 attached at its 3′ end to the receptor site in the genome, while the 5′ end is unattached and resembles a double-strand DNA or DNA-RNA break, we speculate that microDNAs are preferentially generated near double-strand break ends created during L1 element transposition.

Tissue MicroDNA Genomic Hotspot Features

Next we analyzed the genomic regions that commonly generate microDNAs and how they compare between tissue types. On a chromosomal level, there is a correlation between the length of a chromosome and the percentage of microDNA that originate from that chromosome (R^2 = 0.44; Figure 3A). Interestingly, when each chromosome is divided into 1-Mb windows and the average GC content, gene density, or percentage of microDNA per Mb is calculated, there is a positive correlation of microDNA density with GC content (R^2 = 0.86) and gene density (R^2 = 0.69), indicating a non-random distribution of microDNA loci throughout the genome (Figure 3A). This is strikingly visualized when each chromosome is divided into 1-Mb windows and the percentage of unique microDNA located within each window is plotted. For example, on chromosome 10, four large "hotspots" of microDNA generation can be identified that overlap between all the tissue types (Figure 3B). In agreement with the analyses in Figure 3A, these hotspots correlate with regions of high GC content (Figure 3C) and gene density (Figure 3D).

Because of the non-random distribution of microDNA throughout the genome and a strong correlation with gene density and 5′-UTRs of genes, we next tested whether the generation of microDNA is linked to transcription and its associated chromatin states. MicroDNAs are enriched over random expectation by >10-fold at promoters with activating or bivalent marks (poised) and at RNA Polymerase II-occupied regions (Figure 3E). There is a lesser enrichment on active enhancers and within the body of active genes across numerous tissue types. In contrast, microDNAs are depleted from lamin-associated domains, which are genomic regions that are in contact with the nuclear lamina and are typified by low gene-expression levels (Guelen et al.,

Table 1. Summary of MicroDNA Sequencing and Mapping in Normal Adult Mouse Tissues, Human Cancer Cell Lines, and DT40 Cell Lines

Sample	Paired End Reads	Pairs Aligned	Mapped-Unmapped Pairs[a]	Read Sequences	Aligned Sequences	Unique Alignment	Unique microDNA (Complexity)[b]
Mouse tissue type							
Brain	24.7	7.8	6.7	49.3	32.9	31.6	24,312
Heart	30.8	8.89	8.8	61.5	41.4	36.0	15,876
Kidney	35.2	12.9	10.4	70.5	49.9	45.2	39,481
Liver	29.7	8.4	5.5	59.4	33.1	28.0	45,958
Lung	41.5	14.7	14.7	83.0	55.0	49.6	19,659
Skeletal muscle	36.3	12.9	10.7	72.6	50.8	46.1	38,503
Sperm	29.2	7.3	3.6	58.5	24.5	20.9	5,271
Spleen	37.6	11.9	11.6	75.1	49.0	45.2	54,481
Testis	25.4	8.3	6.4	50.8	30.5	28.2	48,267
Thymus	38.1	12.0	14.5	76.1	45.9	42.9	91,204
Human cancer cell line							
ES2	61.9	26.8	15.1	123.9	96.4	86.7	114,752
OVCAR8	50.2	28.8	8.9	100.4	84.5	75.8	57,327
C4-2	41.1	21.4	8.3	82.2	69.3	63.2	41,410
LnCap	56.1	24.8	12.5	112.1	89.1	82.3	84,841
PC3	43.5	10.7	7.4	87.0	41.6	38.8	14,705
DT40 cell line							
WT	33.6	11.3	7.6	67.3	43.0	40.0	106,983
BRCA1−/−	43.7	13.2	8.9	87.3	57.2	52.8	122,403
BRCA2−/−	38.6	11.8	9.0	77.2	51.7	47.5	112,199
CtIP−/−	47.4	16.3	10.3	94.7	63.0	60.0	149,006
Ku70−/−	38.7	11.9	9.0	77.4	50.7	46.7	124,433
Lig4−/−	52.8	15.8	10.4	105.6	60.4	56.1	138,463
MSH3−/−	67.3	23.9	16.5	134.7	87.6	75.3	115,221
NBS1−/−	43.5	12.2	10.6	87.1	54.9	51.2	128,693
Rad54−/−	37.4	12.3	8.4	74.8	47.7	44.2	112,530

[a] Mapped-unmapped pairs with junctional sequences indicative of circularization.

[b] All values in millions except unique microDNA.

2008). Furthermore, microDNAs producing loci are significantly enriched in the core regions of active promoters compared to their flanking regions (Figure 3F) and at transcription start sites (+/− 1-Kb) with activating (H3K4me3+) chromatin marks (Figure 3G). Combined, these data indicate that the generation of microDNAs is in part linked to transcription and RNA metabolism. Consistent with this, there is a progressive enrichment of microDNA yield with the number of bases that are transcribed, up to about 1,500 bases transcribed in a 2,000 base window (Figure 3H). However, we were struck by the sharp drop-off of microDNA yield in windows that were transcribed for greater than 1,500 bases. We speculated that the difference might stem from whether the transcription was over an exon (usually <1,500 bases in length) or an intron (which is often >1,500 bases long). Indeed, we have noted in Figure 1E that exons are much more enriched in microDNA yield than introns. Thus, we examined whether microDNA yield increases in areas with high exon density and discovered a striking increase in yield of microDNAs with increasing numbers of exons in the 2,000 base window (Figure 3I). Thus RNA transcription with splicing appears to favor microDNA production, and the microDNAs produced tend to overlap more with exons than with introns. This result suggests that a high level of pre-mRNA splicing at a genomic locus contributes to microDNA production.

MicroDNA in Human Ovarian and Prostate Cancer Cell Lines

Next, we examined human cancer cell lines of two origins, prostate (LNCaP, C4-2, and PC-3) or ovarian (OVCAR8 and ES-2), to determine whether microDNAs are selectively generated from sites that are expressed differentially between the two lineages. Tens to hundreds of thousands of unique microDNAs were identified within each cancer cell line, mapping to unique non-repetitive regions of the human genome (Table 1). Consistent with our observations in the mouse tissues, microDNAs from the human cancer cell lines are primarily 100 to 400 bp in length (Figure 4A), GC rich (Figure 4B), have a high frequency of 2- to 15-bp repeats at the starts and ends of the loci generating microDNAs

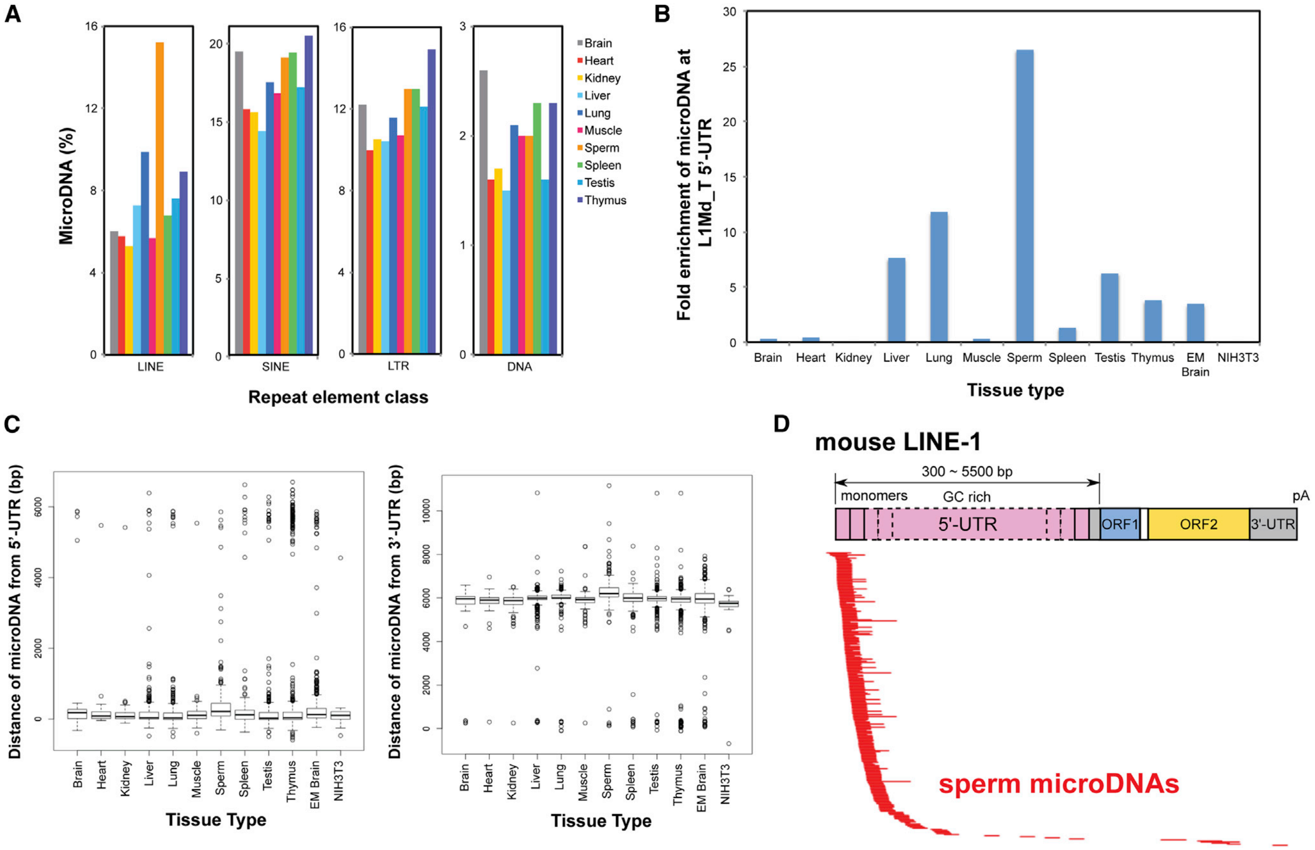


Figure 2. MicroDNAs in Mouse Sperm Are Enriched from LINE-1 Elements

(A) Percentage of microDNA that map to sequences corresponding to each repetitive DNA class using RepeatMasker.

(B) Fold enrichment of microDNAs from adult mouse tissues in the 5′-UTR of full-length intact L1Md_T L1 elements relative to random expectation.

(C) Boxplot distribution of the distance from tissue L1Md_T-mapped microDNA coordinates to the 5′-UTR (left) or 3′-UTR (right) of the element.

(D) Mapped positions of sperm microDNAs (red lines) corresponding to the mouse full-length intact LINE-1 element (diagram on top). MicroDNAs map almost entirely to the tandem repeat monomers (pink boxes) within the 5′-UTR (gray).

See also Figure S2.

(Figure 4C), and are highly enriched in 5′-UTRs, exons, and CpG islands (Figure 4D).

Given the correlation noted earlier between transcription, splicing and active promoters with microDNA production, we predicted that the origins of the microDNAs may be predictive of the lineage of a cancer cell line. To test this we divided the genome into 5-Mb windows and calculated the frequency at which different microDNA sequences in the five cancer cell line libraries were observed in each window. When this site-specific frequency of microDNAs was used to cluster the five data sets by unsupervised hierarchical clustering (Figure 4E), the microDNAs from the two ovarian cancer cell lines clustered together relative to those from the prostate cancer cell lines, suggesting that the sites at which microDNAs formed have some dependence on the lineage of the cancer cell line.

Deletion of DNA Repair Proteins Alters MicroDNA Production

Because of the presence of microhomology at the starts and ends of many microDNA genomic loci, we expected that DNA repair pathways might be involved in microDNA generation. Therefore, we isolated and characterized microDNAs from chicken DT40 cell lines deficient in a variety of important DNA repair proteins, including DNA ligase IV (Lig4) (Adachi et al., 2001) and Ku70 (Takata et al., 1998) involved in non-homologous end joining (NHEJ); BRCA1 (Martin et al., 2007), BRCA2 (Hatanaka et al., 2005), Rad54 (Bezzubova et al., 1997), and CtIP (Nakamura et al., 2010) required for HR, NBS1 (Tauchi et al., 2002) involved in both HR and NHEJ and MSH3 involved in DNA MMR. We found that all mutant strains were capable of producing microDNA from hundreds of thousands of unique genomic loci (Table 1). As we observed in the mouse tissues and human cancer cell lines, microDNAs from the DT40 cell lines are primarily 100 to 400 bp (Figure 5A) and possess a high GC content (Figure 5B).

Furthermore, practically every genomic locus (92%–98%) generating microDNA in all the DT40 lineages exhibits microhomology (2–15 bp) at the sequences directly flanking the starts and ends of the microDNA (Figure 5C), which is a much higher frequency than observed in the mouse tissues (Figure 1D) and

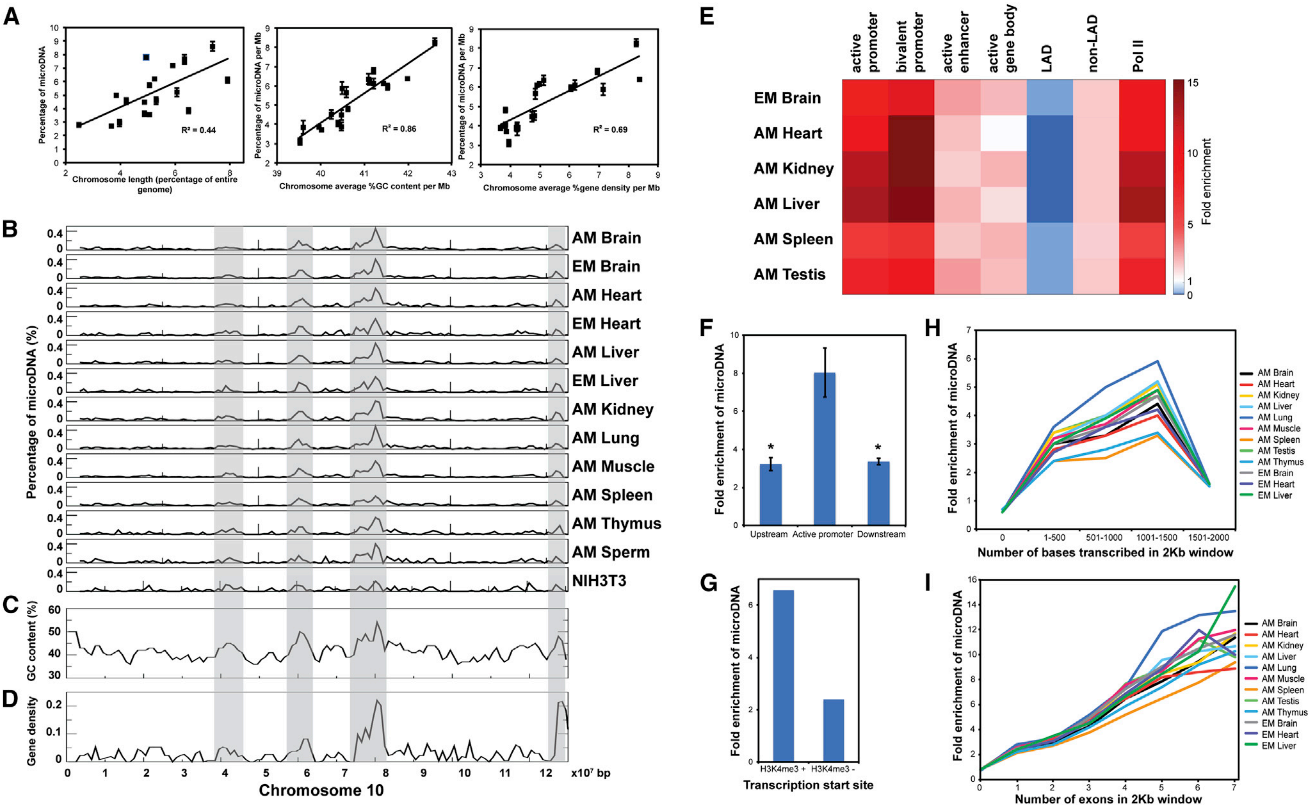


Figure 3. Distribution of MicroDNA along Chromosomes Is Conserved across Tissue Types and Correlates with Active Chromatin Marks and High Exon Density

(A) The length of each chromosome as a percentage of the entire genome versus the average percentage of total microDNA originating from that chromosome (left). The average percent GC content (center) or average percent gene density (right) per Mb for each chromosome versus the average percentage of total microDNA per Mb on the chromosome. Error bars represent SD.

(B–D) MicroDNA loci were grouped into bins of 1-Mb stepwise across mouse chromosome 10, and the percentage of all microDNA located within each bin was calculated for each tissue type. MicroDNA clustering patterns are similar across all tissue types, and "hotspot" regions (gray bars) correlate with a high (C) GC content and (D) gene density.

(E) Fold enrichment of microDNAs in mouse tissues at genomic regions with various chromatin modifications, lamin-associated domains (LADs), or Pol II binding.

(F) Average microDNA enrichment was calculated from three replicates of embryonic mouse brain at active promoters (H3K4me3+,H3K27ac+,H3K27me3−) and flanking sequences of the same length directly upstream or downstream. Error bars represent SD. *$p < 0.005$ (Student's t test).

(G) MicroDNA enrichment within a 2-kb window surrounding transcription start sites that are either ± for H3K4me3 in embryonic mouse brain.

(H) Genic regions were divided into 2-kb windows and grouped based on the number of bases transcribed. Fold enrichment of microDNA loci is presented for each group.

(I) Using the same 2-kb windows, the number of exons per window was calculated, and the fold enrichment of microDNA loci was calculated.

human cancer cell lines (Figure 4C). Although it was unlikely that HR pathways would act on the very short sequences of microhomology to bring the ends of the microDNAs together, we can now definitively rule out such a hypothesis because of the sustained incidence of microhomology at the ends of the microDNAs in the cells with mutations in HR genes. Upon further analysis of the distribution of repeats across the different DT40 cell lines we found that >75% of microDNA loci have 4 to 8 bp of microhomology, with 6 bp being the most frequently observed (Figure S3). Furthermore, no significant differences were observed in the microhomology distribution patterns between DT40 WT cells and the various knockouts.

The DT40 MSH3−/− cell line was unique in that microDNAs that are produced are highly enriched from CpG islands and their neighborhoods (Figure 5D) compared with WT. After observing this alteration in the genomic location of microDNAs from MSH3−/− cells, we examined whether the overall abundance of microDNAs was also altered in this cell line. Double-stranded microDNAs from DT40 WT and MSH3−/− cells were quantified and their lengths measured using electron microscopy. The number of ds microDNAs per nucleus was reduced 81% in MSH3−/− cells compared with WT (Figures 5E and S4A), implicating the MMR pathway in the generation of a significant portion of microDNAs. Furthermore, by counting the number of molecules observed on the grids when we load known numbers of similar length DNA molecules, we estimate that DT40 WT cells contain ~120 ds microDNAs per nucleus while the MSH3−/− DT40 cells contain ~20 microDNAs per nucleus. The EM-based

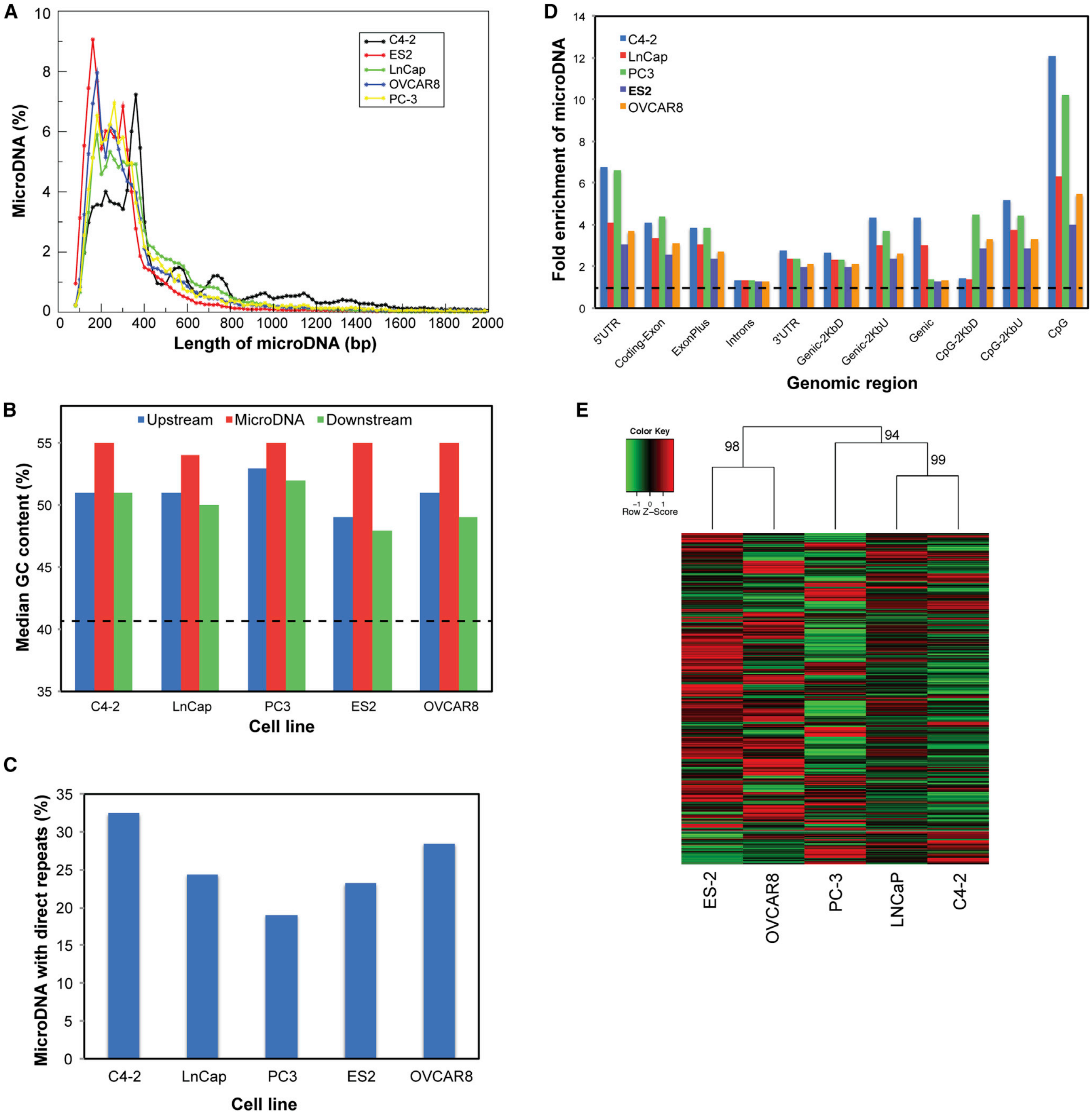


Figure 4. MicroDNA from Human Ovarian and Prostate Cancer Cell Lines Cluster Based on Subtype

(A) Length distribution of microDNAs identified in cancer cell lines.

(B) Median percent GC content of microDNAs and the genomic sequences upstream or downstream of the source loci are enriched relative to the average GC content of the human genome (dashed line).

(C) Direct repeats near the start and end of microDNA sequences (2- to 15-bp) are enriched in all cell lines compared with random expectation.

(D) Enrichment of microDNAs in the indicated genomic region relative to the expected percentage based on random distribution.

(E) MicroDNA loci were grouped into 5-Mb bins stepwise across the human genome and the percentage of all microDNA located within each bin was calculated for each cancer cell line and compared using hierarchical clustering. Values at each branch point indicate the confidence interval of the cluster (approximately unbiased p value calculated by pvclust package in R), where a confidence interval >95 is considered highly significant.

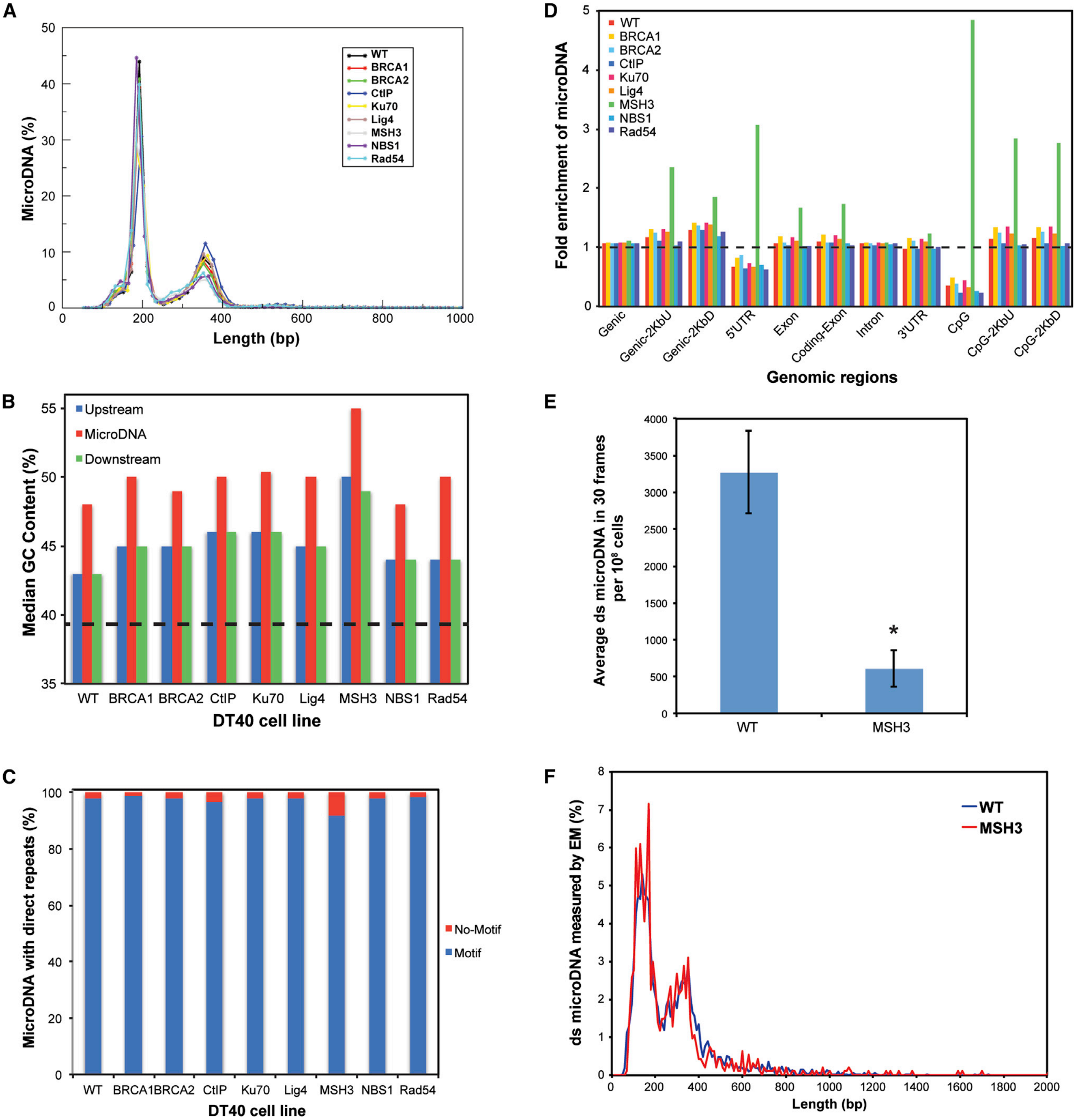


Figure 5. Properties of DT40 MicroDNA Reveal that Multiple DNA Repair Pathways Are Involved in MicroDNA Generation

(A) Length distribution of microDNAs identified in DT40 cell lines.

(B) Median percent GC content of microDNAs and the genomic sequences upstream or downstream of the source loci are enriched relative to the average GC content of the chicken genome (dashed line).

(C) Percentage of microDNA with (blue) or without (red) 2- to 15-bp direct repeats at the genomic source loci.

(D) Enrichment of microDNAs in the indicated genomic region relative to the expected percentage based on random distribution.

(E) ds microDNAs were visualized by EM and the abundance quantitated for 30 random frames. Quantities were normalized based on the total microDNA per 10^8 cells mounted on each grid and averaged for three independent experiments. Error bars represent SD. *p = 0.001 (Student's t test).

(F) Length (bp) distribution of individual ds microDNAs as determined by EM using a DNA standard of known length. Data are a combination of three independent experiments. Results from each individual replicate are given in Figure S4B.

See also Figures S3 and S4.

lengths of the microDNAs (Figure 5F) were very similar to the lengths determined by high-throughput sequencing (Figure 5A) providing strong support for the sequencing method adopted to identify microDNAs. There was no alteration in the length distribution of microDNA in the MSH3−/− cells relative to the WT cells (Figures 5F and S4B).

One hypothesis for the generation of microDNAs is that the microhomology encourages slippage of the replicative DNA polymerase, and the resulting loops are excised (and ligated into circles) by MSH3-dependent MMR pathways (Figure 6, left), with the single-stranded circles being converted to double-stranded circles by primed DNA synthesis. The nature of the microDNAs from MSH3−/− cells suggests that replication slippage and MMR are involved in microDNA production at regions of the genome that are not in CpG islands, such that mutation of MSH3 decreased microDNA production from non-CpG parts of the genome, while sparing microDNAs generated from CpG islands (thus enriching for microDNAs from CpG islands).

DISCUSSION

Together, these studies reveal that microDNAs are a widespread phenomenon found across numerous vertebrate species and are present in all tissue types, and different cellular processes can alter their generation. The frequency and widespread nature of these extrachromosomal DNAs, along with their persistence in non-dividing tissues, indicate that microDNA make up a potentially important fraction (up to ∼10–50 Kb per cell) of uncharacterized DNA within the cell. It is striking that in three disparate biological sources—mouse tissues, human cancer cell lines, and chicken DT40 cells—microDNAs had identical properties: lengths of 100 to 400 bases, high GC content and enriched in short direct repeats flanking the genomic source. Additionally, mammalian microDNAs differed from chicken microDNAs in that only mammalian microDNAs were highly enriched relative to random expectation from genic regions, 5′UTRs, exons, and CpG islands.

MicroDNA loci are enriched in regions of active RNA metabolism with activating chromatin marks and high density of exons. MicroDNA association with genes extends to GC rich sequences, especially within the 5′- and 3′-UTRs. Many of these genomic features are shared with regions susceptible to the formation of R loops, three stranded RNA:DNA hybrid structures formed as a byproduct of transcription that can lead to genomic instability and are implicated in the regulation of gene expression (Skourti-Stathaki and Proudfoot, 2014; Sollier et al., 2014). G-rich DNA, especially at the 5′ and 3′ ends of genes, has a propensity to form R-loops (Ginno et al., 2012, 2013; Roy and Lieber, 2009; Skourti-Stathaki et al., 2011). Like R loops, we often observed microDNA at CpG islands and the 5′ and 3′ ends of genes (Figures 1E, 4D, and S1). Furthermore, loss of the SRSF1 splicing factor has been found to result in increased R-loop formation and subsequent DNA damage, illustrating a connection between R-loop formation and splicing (Li and Manley, 2005). The fact that we find microDNA enriched in genomic regions with activating chromatin marks and high exon density also suggests a connection between microDNA production and mRNA processing. Together this leads to the interesting possibility that R-loop formation predisposes certain parts of the genome (with activating chromatin modifications, bound RNA-polymerase II, high density of intron-exon junctions) to microDNA formation.

Based on the data presented here, there most likely exist multiple mechanisms for the generation of microDNA (Figure 6). For example, if polymerase slippage occurs during DNA replication at succeeding short direct repeats, DNA loops can form on the product or template strand (Figure 6, left). MMR pathways excise these DNA loops (Schofield and Hsieh, 2003), but ligation of the excised product could form an ss microDNA. Excision of a loop on the newly replicated product strand will not leave a deletion in the genome, while excision of a loop from the template strand will lead to a microdeletion in the genome. The greater than 80% decrease in microDNA abundance observed in the DT40 MSH3−/− cell line (Figure 5E) suggests this mechanism may contribute to the majority of microDNA formation within the cell, but not all. Therefore, another possibility is that a DNA break or replication fork stalling allows the newly synthesized nascent DNA strand to circularize with help from the short stretches of microhomology on the template (Figure 6, center). Ligation of such a circle will form an ss microDNA, and displacement of the circle during subsequent repair will not leave a deletion behind in the genome. In both of these cases, the ss microDNA could later be converted to ds microDNA by DNA polymerase. As discussed earlier, the prevalence of microDNAs at the 5′ end of intact LINE-L1 elements in a tissue where the elements are known to transpose suggest a relationship between ds break ends and microDNA generation. Furthermore, hotspots of microDNA generation often have chromosomal microdeletions that also appear to be generated by microhomology-mediated end joining (Shibata et al., 2012). Therefore, two DNA ds breaks followed by microhomology-mediated circularization of the released fragment could lead to the generation of a ds microDNA molecule and a microdeletion within the genome (Figure 6, right).

In our previous paper, we speculated that the generation of microDNA could affect cellular processes by leaving behind microdeletions in the genomic DNA. In general, the extraordinarily high complexity (number of sites in the genome producing microDNAs) and abundance (over 100 ds microDNAs per cell in DT40 WT cells) of the microDNAs suggest that most microDNAs are generated by copying mechanisms during replication or repair and will not always result in a corresponding microdeletion in the genome. However, our discovery that there are hotspots in the genome that produce microDNA will make it easier to search for such somatically mosaic microdeletions in those parts of the genome in normal tissues. Our results also point to the ubiquity and abundance of the microDNAs, suggesting that these extrachromosomal copies of a genomic sequence can also alter a cell's function by potentially titrating cellular proteins or by producing abnormal short RNAs, hypotheses that we will explore in the future. Overall, these results add to our understanding of the plasticity and diversity of what was previously believed to be a static genome, particularly in normal cells and tissues.

EXPERIMENTAL PROCEDURES

See the Supplemental Experimental Procedures for additional details.

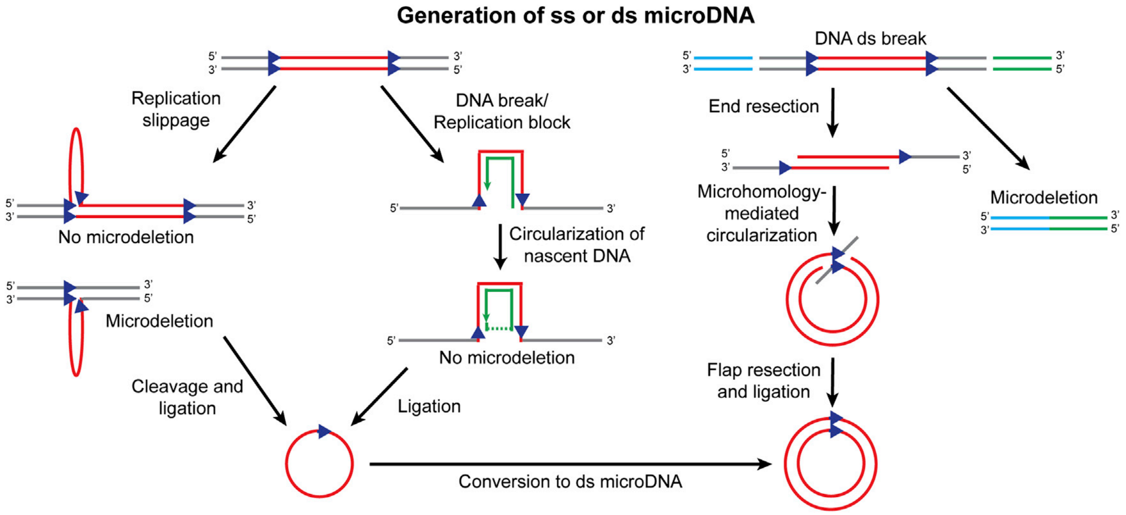


Figure 6. Models for the Generation of ss or ds MicroDNA

(Left) Polymerase slippage during DNA replication at succeeding short direct repeats (blue arrowheads) can result in the formation of DNA loops on the product or template strand (far left). Excision of the loop and subsequent ligation could result in the formation of ss microDNA and leave behind a deletion if occurring on the template strand. (Center) A DNA break or replication fork stalling allows the newly synthesized nascent DNA strand to circularize with help from the short stretches of microhomology on the template. Ligation of such a circle will form an ss microDNA, and displacement of the circle during subsequent repair will not leave a deletion behind in the genome. In both the processes described in the left and center, the ss microDNA could later be converted to ds microDNA by DNA polymerase. (Right) Two DNA ds breaks followed by microhomology-mediated circularization of the released fragment could lead to the generation of a ds microDNA molecule and a microdeletion within the genome.

Animal Care and Use

Animal studies were performed according to protocols approved by the University of Virginia Institutional Animal Care and Use Committee. All tissues were collected from 6-month-old male C57BL/6 mice.

MicroDNA Isolation and Purification

MicroDNA were isolated and purified as described in (Shibata et al., 2012). In short, nuclei were extracted from mouse tissues and cell lines, and extrachromosomal DNA was isolated. MicroDNAs were purified from the total extrachromosomal DNA fraction by removal of linear DNA by exonucleases.

MicroDNA Library Preparation and Sequencing

Purified eccDNA was amplified using MDA and DNA libraries generated. Paired-end DNA sequencing (50 cycles) was performed on the Illumina platform.

Identification of MicroDNA by Paired-End Sequencing

The algorithm used for the identification of microDNAs from paired-end sequencing data is the same as described in (Shibata et al., 2012). In short, paired-end reads are mapped to the reference genome, and using a combination of the island and split-read method, unique circular microDNAs are identified.

Epigenetic Marks

Histone H3 and RNA polymerase II ChIP-seq data for mouse tissues were downloaded from ENCODE/LICR and LAD coordinates from NKI Nuclear Lamina Associated Domains Track in the UCSC browser. Source classes were defined by overlapping coordinate signatures as follows: active promoter = H4K3me3+,H3K27ac+,H3K27me3−, bivalent promoters = H3K4me3+,H3K27me3+, H3K27ac−, active enhancer = H3K4me1+ and H3K4me3−, H3K27ac and H3K27me3−, active gene body = H3K36me3 without active promoter marks.

Electron Microscopy for Quantitating Abundance of MicroDNAs per Cell

Extrachromosomal DNA was prepared for visualization by electron microscopy by direct mounting as described previously (Shibata et al., 2012). For quantification, microDNAs from a defined number of cells were mounted, and 30 randomly selected images were captured from across the grid and the number of circles counted and normalized to the cell count. A DNA standard of known length and quantity was used to determine the lengths of the microDNAs and the number of molecules present per sample.

ACCESSION NUMBERS

Sequencing data were submitted to NCBI GEO and are available under accession number GEO: GSE68644.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and four figures and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.05.020.

AUTHOR CONTRIBUTIONS

L.W.D., P.K., Y.S., and A.D. designed the study. L.W.D. performed microDNA isolation and paired-end DNA library construction and next generation sequencing for mouse tissues and human cancer cell lines and Y.S. for DT40 cell lines. P.K. performed the majority of the bioinformatics analyses with help from Y.S., S.W., Y.-H.W., L.W.D., and J.D.G performed electron microscopy of microDNA from mouse tissues and DT40 cell lines. S.T. and Y.P. generated and provided DT40 mutant cell lines. L.W.D., P.K., Y.S., and A.D. wrote the manuscript.

ACKNOWLEDGMENTS

We thank E. Rissman from the University of Virginia for providing mouse sperm samples, A. Jazaeri from MD Anderson Cancer Center for providing the human ovarian cancer cell lines OVCAR8 and ES-2, and all members of the A.D. lab for helpful discussions. This work was supported by grants from the NIH (RO1CA60499 and PO1CA104160 to A.D., T32CA009109 to L.W.D., GM31819 and EM13773 to J.D.G., and RO1GM101192 to Y.-H.W.), Department of Defense (W81XWH-13-1-0088 to P.K.), NCI Intramural Program, Center for Cancer Research (Z01-BC-006150 to Y.P.), Ovarian Cancer Research Fund (290250 to L.W.D.), and Kaleidoscope of Hope Foundation (to L.W.D.).

Received: March 12, 2015
Revised: April 20, 2015
Accepted: May 11, 2015
Published: June 4, 2015

REFERENCES

Adachi, N., Ishino, T., Ishii, Y., Takeda, S., and Koyama, H. (2001). DNA ligase IV-deficient cells are more resistant to ionizing radiation in the absence of Ku70: Implications for DNA double-strand break repair. Proc. Natl. Acad. Sci. USA *98*, 12109–12113.

Beckmann, J.S., Estivill, X., and Antonarakis, S.E. (2007). Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat. Rev. Genet. *8*, 639–646.

Bezzubova, O., Silbergleit, A., Yamaguchi-Iwai, Y., Takeda, S., and Buerstedde, J.M. (1997). Reduced X-ray resistance and homologous recombination frequencies in a RAD54-/- mutant of the chicken DT40 cell line. Cell *89*, 185–193.

Branciforte, D., and Martin, S.L. (1994). Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. Mol. Cell. Biol. *14*, 2584–2592.

Cohen, S., and Segal, D. (2009). Extrachromosomal circular DNA in eukaryotes: possible involvement in the plasticity of tandem repeats. Cytogenet. Genome Res. *124*, 327–338.

Flores, M., Morales, L., Gonzaga-Jauregui, C., Domínguez-Vidaña, R., Zepeda, C., Yañez, O., Gutiérrez, M., Lemus, T., Valle, D., Avila, M.C., et al. (2007). Recurrent DNA inversion rearrangements in the human genome. Proc. Natl. Acad. Sci. USA *104*, 6099–6106.

Frazer, K.A., Murray, S.S., Schork, N.J., and Topol, E.J. (2009). Human genetic variation and its contribution to complex traits. Nat. Rev. Genet. *10*, 241–251.

Ginno, P.A., Lott, P.L., Christensen, H.C., Korf, I., and Chédin, F. (2012). R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol. Cell *45*, 814–825.

Ginno, P.A., Lim, Y.W., Lott, P.L., Korf, I., and Chédin, F. (2013). GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. Genome Res. *23*, 1590–1600.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., and van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948–951.

Hatanaka, A., Yamazoe, M., Sale, J.E., Takata, M., Yamamoto, K., Kitao, H., Sonoda, E., Kikuchi, K., Yonetani, Y., and Takeda, S. (2005). Similar effects of Brca2 truncation and Rad51 paralog deficiency on immunoglobulin V gene diversification in DT40 cells support an early role for Rad51 paralogs in homologous recombination. Mol. Cell. Biol. *25*, 1124–1134.

Li, X., and Manley, J.L. (2005). Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. Cell *122*, 365–378.

Lupski, J.R. (2010). New mutations and intellectual function. Nat. Genet. *42*, 1036–1038.

Martin, R.W., Orelli, B.J., Yamazoe, M., Minn, A.J., Takeda, S., and Bishop, D.K. (2007). RAD51 up-regulation bypasses BRCA1 function and is a common feature of BRCA1-deficient breast tumors. Cancer Res. *67*, 9658–9665.

Nakamura, K., Kogame, T., Oshiumi, H., Shinohara, A., Sumitomo, Y., Agama, K., Pommier, Y., Tsutsui, K.M., Tsutsui, K., Hartsuiker, E., et al. (2010). Collaborative action of Brca1 and CtIP in elimination of covalent modifications from double-strand breaks to facilitate subsequent break repair. PLoS Genet. *6*, e1000828.

Ostertag, E.M., and Kazazian, H.H., Jr. (2001). Biology of mammalian L1 retrotransposons. Annu. Rev. Genet. *35*, 501–538.

Penzkofer, T., Dandekar, T., and Zemojtel, T. (2005). L1Base: from functional annotation to prediction of active LINE-1 elements. Nucleic Acids Res. *33*, D498–D500.

Roy, D., and Lieber, M.R. (2009). G clustering is important for the initiation of transcription-induced R-loops in vitro, whereas high G density without clustering is sufficient thereafter. Mol. Cell. Biol. *29*, 3124–3133.

Schofield, M.J., and Hsieh, P. (2003). DNA mismatch repair: molecular mechanisms and biological function. Annu. Rev. Microbiol. *57*, 579–608.

Shibata, Y., Kumar, P., Layer, R., Willcox, S., Gagan, J.R., Griffith, J.D., and Dutta, A. (2012). Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. Science *336*, 82–86.

Skourti-Stathaki, K., and Proudfoot, N.J. (2014). A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. Genes Dev. *28*, 1384–1396.

Skourti-Stathaki, K., Proudfoot, N.J., and Gromak, N. (2011). Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. Mol. Cell *42*, 794–805.

Smit, A.F.A., Hubley, R., and Green, P. (1996–2010). RepeatMasker Open-3.0. http://www.repeatmasker.org/.

Sollier, J., Stork, C.T., García-Rubio, M.L., Paulsen, R.D., Aguilera, A., and Cimprich, K.A. (2014). Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. Mol. Cell *56*, 777–785.

Stankiewicz, P., and Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. Annu. Rev. Med. *61*, 437–455.

Takata, M., Sasaki, M.S., Sonoda, E., Morrison, C., Hashimoto, M., Utsumi, H., Yamaguchi-Iwai, Y., Shinohara, A., and Takeda, S. (1998). Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. EMBO J. *17*, 5497–5508.

Tauchi, H., Kobayashi, J., Morishima, K., van Gent, D.C., Shiraishi, T., Verkaik, N.S., vanHeems, D., Ito, E., Nakamura, A., Sonoda, E., et al. (2002). Nbs1 is essential for DNA repair by homologous recombination in higher vertebrate cells. Nature *420*, 93–98.

Published online 11 November 2014 *Nucleic Acids Research, 2015, Vol. 43, Database issue* ***D141–D145***
doi: 10.1093/nar/gku1138

tRFdb: a database for transfer RNA fragments

Pankaj Kumar[1], Suresh B. Mudunuri[2], Jordan Anaya[1] and Anindya Dutta[1,*]

[1]Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, VA 22901, USA and [2]Department of Computer Science and Engineering, Grandhi Varalakshmi Venkatarao Institute of Technology (GVIT), Bhimavaram, Andhra Pradesh 534207, India

Received August 14, 2014; Revised October 15, 2014; Accepted October 27, 2014

ABSTRACT

We have created tRFdb, the first database of transfer RNA fragments (tRFs), available at http://genome.bioch.virginia.edu/trfdb/. With over 100 small RNA libraries analyzed, the database currently contains the sequences and read counts of the three classes of tRFs for eight species: *R. sphaeroides*, *S. pombe*, *D. melanogaster*, *C. elegans*, Xenopus, zebra fish, mouse and human, for a total of 12 877 tRFs. The database can be searched by tRF ID or tRF sequence, and the results can be limited by organism. The search results show the genome coordinates and names of the tRNAs the sequence may derive from, and there are links for the sequence of the tRF and parental tRNA, and links for the read counts in all the corresponding small RNA libraries. As a case study for how this database may be used, we have shown that a certain class of tRFs, tRF-1s, is highly upregulated in B-cell malignancies.

INTRODUCTION

tRNA-derived RNA fragments approaching the size of microRNAs were first appreciated as a class of small noncoding RNA in 2009 by three different laboratories (1–3). Sequences mapping to the 5′ ends of tRNAs (transfer RNA fragment (tRF)-5s), the 3′ ends of tRNAs (tRF-3s) and the trailer sequence (tRF-1s) were observed in LNCap and C4-2 cells, and a tRF-1 was shown to be involved in cell proliferation (Figure 1) (2). tRFs were also found to be present in HeLa nuclei and associated with Argonautes, and present in HEK 293 cells and involved in RNA silencing (1,3). Since then tRFs have been found in all domains of life, and there are now a few reviews cataloging the known functions of tRFs (4,5). Despite this work on tRFs, it is not known how tRF-5s or tRF-3s are generated, and the function of the majority of tRFs is unknown.

tRFs are present in similar abundance to miRNAs, are more evolutionarily conserved and interesting functions for tRFs continue to be discovered, yet the handful of papers on tRFs clearly shows that research on this class of small RNA is lagging behind other small RNAs. This may be due to a general confusion about these sequences and lack of standardized nomenclature, for example the same tRF has been referred to by different names, and a tRF has been mistaken as an miRNA (1,6). In addition, currently there is not a searchable database where researchers can compare tRFs from various experiments.

Because a tRF may be derived from several different tRNAs, and several distinct tRFs may come from the same tRNA, it is not practical to use the name of the parental tRNA in the tRF identifier. As a result, we have decided to improve upon the nomenclature already established in the field (2). In each organism, tRFs are named in the order they are identified, with the first tRF-5 named 5001, the first tRF-3 named 3001 and the first tRF-1 named 1001. In the case of tRF-5s and tRF-3s, there are multiple distinct subclasses (7). When there are two or more tRF-5s that differ only in length: an ‘a’, ‘b’ or ‘c’ is appended for tRF-5s of ∼15, ∼22 or ∼31 bases. All tRF-5as, -bs and -cs share a common seed sequence. Similarly, when there are two distinct tRF-3s mapping to the same tRNA, the tRF-3s of length ∼18 have an ‘a’ appended, while tRF-3s of length ∼22 have a ‘b’ appended. The latter is of particular importance since tRF-3-as and tRF-3-bs have different 5′ ends and therefore different seed sequences.

The field of tRF research is in its infancy and we present the first attempt to classify and tabulate tRF sequences, tRFdb, a database for tRFs. In its current form, the database is a simple way for researchers to view the tRF sequences present in various organisms and compare their read counts in multiple experiments. We hope that our database spurs research into this novel class of small RNA and we invite suggestions from the community on what additions should be added to future versions of tRFdb.

MATERIALS AND METHODS

Analysis of the small RNA data

Source and processing of small RNA high-throughput sequencing data. The small RNA high-throughput sequencing data were downloaded from the GEO (http://www.ncbi.nlm.nih.gov/geo/) and NCBI SRA databases (http://www.ncbi.nlm.nih.gov/sra). Information about the

[*]To whom correspondence should be addressed. Tel: +1 434 924 2466; Fax: +1 434 924 5069; Email: ad8q@virginia.edu

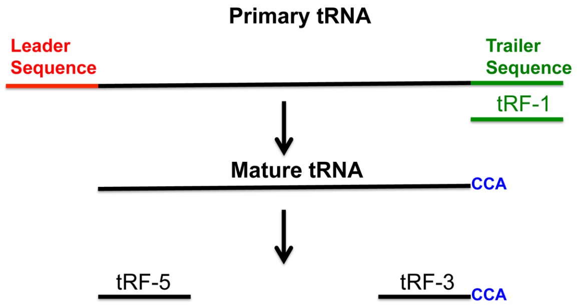


Figure 1. Illustration of primary tRNA, mature tRNA and tRF-5, -3 and -1. tRF-1 (shown in green) is generated from primary tRNA. tRF-5 and -3 are produced from mature tRNA. The tRF-3s always have 'CCA' at their 3′ end.

tRNA genes for species *Rhodobacter sphaeroides* (Bacteria; ATCC_17025), *Schizosaccharomyces pombe* (schiPomb1), Drosophila (dm3), *Caenorhabditis elegans* (ce6), Xenopus (xenTro3), zebra fish (Zv9), mouse (mm9) and human (hg19) was either downloaded from the 'Genomic tRNA database' (http://gtrnadb.ucsc.edu) or NCBI (http://www.ncbi.nlm.nih.gov/). We extracted mature tRNA genes from the same genome assembly on which the tRNA gene coordinates were built and added 'CCA' at the end. In addition to that we also add 50 bases downstream sequences to find the tRFs generated from the tRNA trailer sequences. The genomic sequences were extracted based on the strand information of tRNA gene transcription. A species-specific tRNAdb blast database was built to query the small RNA sequences and the small RNAs were mapped using BLASTn (8). We considered only those alignments where the query sequence (small RNA) was mapped to the database sequence (tRNA) along 100% of its length with 100% identity. To eliminate any false positives, the small RNAs that mapped on to the 'tRNAdb' were again searched against the whole genome database using blast search excluding the tRNA loci. Only those small RNAs that mapped exclusively to tRNA loci were included as probable tRFs.

Quality filter to remove random degradation products of tRNA genes. The ends of the small RNA mapped on tRNA genes was used to assess the significant enrichment of any mapped small RNA on tRNA. The small RNA mostly (>90% of total mapped reads on individual tRNA) mapped on three specific regions: extreme 5′ end (tRF-5), extreme 3′ end (tRF-3) of mature tRNA and 3′ trailer region (tRF-1) of primary tRNA genes. Therefore, tRFs mapped only to these specific locations were considered for building of database. As shown in Figure 2, for each tRF, there is one or two most abundant RNA sequenced (e.g. the tRF-5 'GCATTGGTGGTTCAGTGGTAGA' was sequenced at 8258 reads per million (RPM)) accounting for more than 80% of the reads mapping to that site. This distinguishes the main tRF from other low abundance products created by nucleases digesting the main tRF, or possibly from random degradation of tRNAs and tRFs.

RESULTS

tRFs are non-random tRNA fragments

The presence of one highly abundant clone for each tRF (tRF-5, -3 or 1-series) (Figure 2) gives support that the tRF is generated from an individual tRNA with specificity. Furthermore the 5′ end of tRF-1 exactly corresponded to the base immediately downstream from the RNaseZ cleavage site and the 3′ end of tRF-1 contained Pol III transcription termination motifs, also supporting the view that tRF-1s are specific and stable small RNA generated from processing of the precursor tRNA. Interestingly, the leader sequence of the precursor tRNA was rarely found with the abundance of the other tRFs.

For humans and mice, tRF-5s are mainly 15, 22 and 32 nts, whereas tRF-3s are 18 and 22 bases long, while in other organisms tRF-5s and tRF-3s do not contain clear subclasses (7). Depending upon the distance of the transcription termination site, a variable length of small RNA with 3′ poly 'U' tract is generated for different tRNA genes (2,9) and therefore the tRF-1s are more variable in length. All the tRFs that originate from the 5′ end of mature tRNA were found to start with the first base of the mature tRNA, indicating that these tRFs are generated after the removal of the leader sequence from the pre-tRNA. The tRFs, which originate from the 3′ end of mature tRNAs, always had a 'CCA' at their 3′ end (2,9). All the tRF-1s exactly start (5′ end of tRF) with the first unpaired base (discriminator base) at the 3′ end of the acceptor stem of pre-tRNA while the end base (3′ end of tRF) is within a RNA pol III transcription termination sites (poly 'U' tract), indicating that tRF-1s are generated by tRNA 3′ processing enzymes during tRNA maturation (2).

Exploration of database

A snapshot of the database search page is shown in Supplementary Figure S1. The output can either be viewed online or the user may download the output after searching on the selected parameters. The database can be searched either by tRF type (tRF-5, -3 or -1) or tRF-ID (5001, 5001a, 5001b, 5001c, 3001, 3001a, 3001b, 1001 etc.). The database has an option to select the species of interest or search all species together. The output is displayed as a table: tRF-ID, organism name, tRF type, tRNA gene co-ordinate, tRNA gene name and hyperlinks to the tRF sequence itself and the small RNA experiments that detected the tRF. The 'Sequence' hyperlink provides the length and sequence of a given tRF and the originating tRNA gene. The 'Experiment' hyperlink provides the GEO ID of the experiments where the tRF was identified, the abundance of the tRF in those datasets and their source (cell line name or tissue). Sublinks from the 'Experiment' page provide additional information: the 'View Alignment' hyperlink from each experiment displays the alignments (and read frequencies) of all short RNAs that map to the tRNA gene that yields the tRF (similar to Figure 2). The 'View Graph' link displays a summary of the alignments as a histogram of the number of times a base on the tRNA gene is represented in the short RNA library. An example of such a histogram is in Supplementary Figure S2 and demonstrates graphically

TRF-1

```
Mapping of small RNA on the 3′ trailer sequence of primary tRNA.
(1003:chr17-8090184-8090265:chr17.trna7-SerGCT)
GACGAGGTGGCCGAGTGGTTAAGGCGATGGACTGCTAATCCATTGTGCTCTGCACGCGTGGGTTCGAATCCCATCCTCGTCGGCTAAGGAAGTCCTGTGCTCAGTTTT
...............................................................................(5)GCTAAGGAAGTCCTGTGC........
...............................................................................(5)GCTAAGGAAGTCCTGTGCTCAG....
...............................................................................(7)GCTAAGGAAGTCCTGTGCTCAGT...
..............................................................................(15)GCTAAGGAAGTCCTGTGCTCAGTT..
.............................................................................(340)GCTAAGGAAGTCCTGTGCTCAGTTT.
.............................................................................(622)GCTAAGGAAGTCCTGTGCTCAGTTTT
.................................................................................(6)TAAGGAAGTCCTGTGCTCAGTTT.
................................................................................(16)TAAGGAAGTCCTGTGCTCAGTTTT
......................................................................................(31)AGTCCTGTGCTCAGTTTT
```

TRF-3

```
Mapping of small RNA on the 3′ end of mature tRNA.
(3001:chr6-28446481-28446400:chr6.trna126-LeuAAG)

GGTAGCGTGGCCGAGTGGTCTAAGACGCTGGATTAAGGCTCCAGTCTCTTCGGGGGCGTGGGTTTGAATCCCACCGCTGCCACCAGGTTTATT
.............................................................(10)GAATCCCACCGCTGCCACCA
..............................................................(12)AATCCCACCGCTGCCACCA
..............................................................(536)ATCCCACCGCTGCCACCA
...............................................................(112)TCCCACCGCTGCCACCA
.................................................................(19)CCCACCGCTGCCACCA
..................................................................(10)CCACCGCTGCCACCA
```

TRF-5

```
Mapping of small RNA on the 5′ end of mature tRNA.
(5049:chr1-16872504-16872434:chr1.trna133-GlyCCC)

GCATTGGTGGTTCAGTGGTAGAATTCTCGCCTCCCACGCGGGAGACCCGGGTTCAATTCCCGGCCAATGCACCA
GCATTGGTGGTTCAG(123)......................................................
GCATTGGTGGTTCAGT(174).....................................................
GCATTGGTGGTTCAGTG(93).....................................................
GCATTGGTGGTTCAGTGG(331)...................................................
GCATTGGTGGTTCAGTGGT(193)..................................................
GCATTGGTGGTTCAGTGGTA(302).................................................
GCATTGGTGGTTCAGTGGTAG(135)................................................
GCATTGGTGGTTCAGTGGTAGA(8258)..............................................
GCATTGGTGGTTCAGTGGTAGAA(470)..............................................
GCATTGGTGGTTCAGTGGTAGAAT(149).............................................
GCATTGGTGGTTCAGTGGTAGAATT(773)............................................
GCATTGGTGGTTCAGTGGTAGAATTC(68)............................................
GCATTGGTGGTTCAGTGGTAGAATTCT(44)...........................................
GCATTGGTGGTTCAGTGGTAGAATTCTC(224).........................................
GCATTGGTGGTTCAGTGGTAGAATTCTCG(177)........................................
```

Figure 2. The patterns of small RNA deep sequencing reads (GSM416733) mapping to tRNA genes. Examples of small RNA reads that mapped to specific tRNAs are shown. More than 80% of reads for a given tRF-1, -3 or -5 represent one or two most abundant reads and this is the read that is included as the main tRF sequence in the database. The most abundant clones are shown in green. Upper: tRF-1-like sequences mapping to the primary tRNA chr17.trna7-SerGCT. Middle: tRF-3-like sequences mapping to the mature tRNA chr6.trna126-LeuAAG. Lower: tRF-5-like sequences mapping to the mature tRNA chr6.trna126-LeuAAG.

that the tRFs included in tRF-db are extensively enriched relative to other short fragments derived from that tRNA gene. When a tRF-5 or tRF-3 maps to multiple tRNA genes, all the tRNA genes with identical sequence in the relevant area are assigned as a potential source of the tRF. tRF-1s, in contrast, are generated from the 3′ trailer sequence of the primary tRNA and are mostly unique to individual tRNA genes.

Finally, the Help tab provides the help-page that explains the basic concept about tRFs and how to explore and understand the output of the tRFdb. The Statistics tab provides the number of unique tRFs and number of library analyzed for each of the species. Feedback button allows users to provide feedback to the database administrator for any missing or dead link, trouble-shooting or to provide new information or suggestion for the improvement of database.

Future improvements in tRF-db

The 'Sequence search' function already allows the user to identify all tRFs from all species with the exact sequence, even if the tRF-IDs are different between species. In the short term, we will standardize the tRF-IDs such that a given ID will identify tRFs with the same sequence in different species, much like mmu-miR-206 and hsa-miR-206, which identify microRNAs with the same sequence in mouse and human, respectively. Another improvement will be to allow mismatches in the 'Sequence search' function to identify closely related tRFs. This will allow us to group tRF-5s and -3s with the same seed sequence into the same tRF families. In the slightly longer term, this database will be expanded to include targets of tRFs predicted by experimental data such as PAR-Clip or CLASH. We also plan to include the stress-induced tRNA halves (tiRs) that have been widely reported in the literature (10). Finally we propose to provide links to current and future publications that contain functional information on the tRFs as they appear.

tRFs are differentially expressed: a case study of tRFs in normal and cancer B cells

We discovered a very intriguing difference in the expression of tRFs in small RNA isolated from normal and malignant human B cells (11). The small RNA was isolated from each of the four subsets of B cells (naive, germinal center, memory and plasma cell) from normal human subjects in two replicates (from two different individuals). In the same study, small RNAs were isolated from human B-cell-derived tumors for each B-cell subset. tRF-1s, as a class, were more abundant in the malignant compared to normal in all the subsets of B cells. In contrast, the abundance of tRF-5s or tRF-3s was not significantly different in normal and malignant B cells. The differentially expressed tRFs between the normal and malignant B cells that were detected at >20 RPM are shown in Supplementary Figure S3A and B (included in the database). Many of the individual tRF-1s were 100–1000-fold more abundant in the malignant B cells compared to the normal B cells. However, specific tRF-5s or tRF-3s did not exhibit a similar induction in cancer B cells as shown in Supplementary Figure S4 (included in the database). In fact, several tRF-5s or -3s were equal or less abundant in the malignant B cells. Thus, the induction of tRF-1s in malignant B cells is not simply a reflection of higher metabolism of tRNAs in the cancer cells. These differentially expressed tRF-1s need to be further tested to discover the significance of their high expression in tumor cells and to determine if they can be used as biomarkers for B-cell malignancy.

DISCUSSION

tRFs are a newly discovered class of micro-RNA-sized small RNAs that are highly abundant in different human and mouse cell lines, mouse tissues and organisms ranging from bacteria to humans. Individual tRFs are generated with precise ends and are not degradation products of the tRNAs. Mutation of different components of the miRNA biogenesis pathway does not affect tRF levels (7). In human HEK293 cells, tRF-5s and tRF-3s are associated with Argonautes 1, 3, and 4 as evidenced by PAR-CLIP data (7) (included in the database).

Seven to eight base-long seed sequences at the 5′ ends of miRNAs have to be complementary to the target gene 3′UTR to suppress the expression of the target gene. Considering the miRNA-like binding of tRFs to Ago proteins, and CLASH data indicating the tRFs bind to target mRNAs that are complementary to the 5′ seed sequences of tRFs (7), it is important to note that though tRF-3 and tRF-1 have similar 3′ ends (CCA in case of tRF-3 and poly ‘U’ in case of tRF-1) (2), their 5′ ends have much more diversity, thus targeting different 3′UTR regions. However, much like miRNAs, we expect to subclassify the tRFs into tRF families based on similarities of the 5′ seed sequences, and this will be added to the database later.

It has been suggested recently that only highly expressing miRNAs are functional in mammals (12). Thus, it is noteworthy that many of the tRFs are sequenced at an abundance comparable to that of many abundant microRNAs. For example, in the GSM416733 dataset, the five most abundant microRNAs and their RPM are: hsa-miR-106b-5p (4101), hsa-miR-103a-3p (5222), hsa-miR-20a-5p (5316), hsa-miR-16-5p (9630) and hsa-miR-17-5p (9883). In comparison, the RPM of the two most abundant tRF-5s, -3s and -1s in the same dataset are: tRF- 5014a (8226), tRF-5030b (7890), tRF- 3008a (8132) and tRF- 3008b (7313), tRF-1032 (2341) and tRF-1037 (3117)!

tRF-3s associate with PIWI protein Twi12, an Argonaute family protein in Tetrahymena that does not have slicer activity like Ago-2 (13). Many of the human tRF-3 and -5s also bind strongly with Ago1-3-4 compared to Ago-2 protein (7). Since Ago-1-3-4 also do not have slicer activity, the tRFs may be involved in as yet undiscovered functions unique to the non-slicer Ago proteins. It is also possible that the high association of tRFs with non-slicer Argonaute family of proteins may be to sequester the tRFs to prevent them from interfering with Ago-2 containing RNA-induced silencing complexes. This database will help researchers explore how tRFs contribute to gene-regulation circuits through their association with Argonaute proteins and inspire a search for other proteins that associate with and are effectors of this enigmatic class of non-micro-short RNAs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We thank members of the Dutta lab and Dr Uma Devi Paila for helpful discussions.

FUNDING

National Institute of Health [P01 CA104106 to A.D.]; Department of Defense [W81XWH-13-1-0088 to P.K.]; Cell and Molecular Biology Training (CMB) grant [to J.A.]. Funding for open access charge: National Institute of Health [P01 CA104106 to A.D.].

Conflict of interest statement. None declared.

REFERENCES

1. Cole,C., Sobala,A., Lu,C., Thatcher,S.R., Bowman,A., Brown,J.W., Green,P.J., Barton,G.J. and Hutvagner,G. (2009) Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA*, **15**, 2147–2160.
2. Lee,Y.S., Shibata,Y., Malhotra,A. and Dutta,A. (2009) A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.*, **23**, 2639–2649.
3. Haussecker,D., Huang,Y., Lau,A., Parameswaran,P., Fire,A.Z. and Kay,M.A. (2010) Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA*, **16**, 673–695.

4. Tuck,A.C. and Tollervey,D. (2011) RNA in pieces. *Trends Genet.*, **27**, 422–432.
5. Pederson,T. (2010) Regulatory RNAs derived from transfer RNA? *RNA*, **16**, 1865–1869.
6. Schopman,N.C., Heynen,S., Haasnoot,J. and Berkhout,B. (2010) A miRNA-tRNA mix-up: tRNA origin of proposed miRNA. *RNA Biol.*, **7**, 573–576.
7. Kumar,P., Anaya,J., Mudunuri,S.B. and Dutta,A. (2014) Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Biol.*, **12**, 78.
8. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
9. Wang,Q., Lee,I., Ren,J., Ajay,S.S., Lee,Y.S. and Bao,X. (2012) Identification and functional characterization of tRNA-derived RNA fragments (tRFs) in respiratory syncytial virus infection. *Mol Ther.*, **21**, 368–379.
10. Thompson,D.M. and Parker,R. (2009) Stressing out over tRNA cleavage. *Cell*, **138**, 215–219.
11. Jima,D.D., Zhang,J., Jacobs,C., Richards,K.L., Dunphy,C.H., Choi,W.W., Au,W.Y., Srivastava,G., Czader,M.B., Rizzieri,D.A. *et al.* (2010) Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. *Blood*, **116**, e118–e127.
12. Mullokandov,G., Baccarini,A., Ruzo,A., Jayaprakash,A.D., Tung,N., Israelow,B., Evans,M.J., Sachidanandam,R. and Brown,B.D. (2012) High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nat. Methods*, **9**, 840–846.
13. Couvillion,M.T., Sachidanandam,R. and Collins,K. (2010) A growth-essential Tetrahymena Piwi protein carries tRNA fragment cargo. *Genes Dev.*, **24**, 2742–2747.

BMC Biology

RESEARCH ARTICLE Open Access

Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets

Pankaj Kumar[1], Jordan Anaya[1], Suresh B Mudunuri[2] and Anindya Dutta[1,3*]

Abstract

Background: tRFs, 14 to 32 nt long single-stranded RNA derived from mature or precursor tRNAs, are a recently discovered class of small RNA that have been found to be present in diverse organisms at read counts comparable to miRNAs. Currently, there is a debate about their biogenesis and function.

Results: This is the first meta-analysis of tRFs. Analysis of more than 50 short RNA libraries has revealed that tRFs are precisely generated fragments present in all domains of life (bacteria to humans), and are not produced by the miRNA biogenesis pathway. Human PAR-CLIP data shows a striking preference for tRF-5s and tRF-3s to associate with AGO1, 3 and 4 rather than AGO2, and analysis of positional T to C mutational frequency indicates these tRFs associate with Argonautes in a manner similar to miRNAs. The reverse complements of canonical seed positions in these sequences match cross-link centered regions, suggesting these tRF-5s and tRF-3s interact with RNAs in the cell. Consistent with these results, human AGO1 CLASH data contains thousands of tRF-5 and tRF-3 reads chimeric with mRNAs.

Conclusions: tRFs are an abundant class of small RNA present in all domains of life whose biogenesis is distinct from miRNAs. In human HEK293 cells tRFs associate with Argonautes 1, 3 and 4 and not Argonaute 2 which is the main effector protein of miRNA function, but otherwise have very similar properties to miRNAs, indicating tRFs may play a major role in RNA silencing.

Keywords: Small RNA, Non-coding RNA, Regulatory RNA, tRF, tRNA

Background

Small RNAs have been defined as 19 to 31 nucleotide long RNAs present primarily in metazoans and plants, classified as either a miRNA, siRNA or piRNA based on biogenesis, and found to regulate gene expression through association with an Argonaute family member (reviewed in [1]). However, recently it has been shown that bacteria can use CRISPR RNAs as a form of adaptive immunity [2] and that a bacterial Argonaute associates with small RNAs preferentially derived from plasmids [3], suggesting that RNA interference may be more ubiquitous than previously appreciated. Moreover, careful analysis of deep sequencing libraries has revealed many reads that cannot be assigned to known small RNAs, and instead map to mRNAs, snoRNAs, rRNAs, tRNAs or others (reviewed in [4]), suggesting the existence of previously unappreciated classes of small RNAs, some of which appear to be conserved among all domains of life.

The best studied of these new small RNAs are fragments of tRNAs that correspond to half of a mature tRNA. First described in *Escherichia coli* as a response to bacteriophage infection [5], these fragments have been observed in numerous organisms and are commonly referred to as tiRNAs (reviewed in [6]). These molecules are known to accumulate during stress, are generated by Rny1 in yeast, angiogenin (ANG) in humans, and the 5' halves have been shown to be capable of inhibiting protein translation in

* Correspondence: ad8q@virginia.edu
[1]Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, VA 22901, USA
[3]PO Box 800733, 1340 Jefferson Park, Ave Jordan Hall Room 1232, Charlottesville, VA, USA
Full list of author information is available at the end of the article

multiple organisms [7,8], while either the 5' halves or 3' halves could theoretically associate with RNase Z or RNase P, respectively, to slice target RNAs [9,10].

Distinct from tRNA halves are the less well studied small RNAs known as tRNA derived RNA fragments (tRFs). There are three types of tRFs recognized, those derived from the extreme 5' and 3' ends of mature tRNAs (tRF-5s and tRF-3s), and those that map to the 3' trailer fragment of precursor tRNA transcripts (tRF-1s). These classes were first observed in LNCaP and C4-2 cells, and one tRF-1 was found to promote cell proliferation [11]. Soon after, numerous tRF-5s were observed in HeLa cell nucleoli deep sequencing, and these small RNAs were found to be weakly associated with Argonautes 1 and 2, and one was shown to be generated by DICER1 [12]. Consistent with these previous reports, tRF-3 and tRF-1 sequences were reported in HEK293 cells (referred to as Type I and Type II tsRNAs, respectively), and were shown to be primarily cytoplasmic [13]. This study also showed that tRF-1s were formed by RNase Z as expected, tRF-3s and tRF-1s preferentially associated with Argonautes 3 and 4 over 1 and 2, tRF levels could affect the efficacy of miRNAs and siRNAs, and a tRF-3 but not a tRF-1 could act in trans RNA silencing.

Since the initial classification of tRFs there have been multiple studies on tRFs in organisms ranging from archaea to humans, and these studies have been summarized in several recent reviews [14-16]. These reviews highlight the conflicting reports of the biogenesis of tRF-5s and tRF-3s, with some original reports on these molecules implicating DICER1, but a recent paper showing DICER1 is dispensable for the generation of most tRF-5s and tRF-3s [17]. Several recent papers have also shown that tRF-5s or tRF-3s can associate with Argonautes or participate in RNA silencing [18-20], while another paper argues tRF-5s cannot silence a reporter gene but rather function similarly to 5' tRNA halves and inhibit translation [21]. Despite this growing literature on tRFs there are still concerns that tRFs could simply represent degradation products of their extremely abundant parental molecules, or concerns that the reads seen in deep sequencing are biologically relevant since it is known tRNA modifications can affect reverse transcriptase [22].

We have taken advantage of the recent explosion of small RNA-Seq data and novel methods of mapping the miRNA interactome to perform a meta-analysis of tRFs and provide insight into their properties. Our analysis of publicly available data sets clearly shows that tRFs are DROSHA-, DICER1-independent precisely generated fragments present in organisms ranging from bacteria to humans. We find that tRF-5s and tRF-3s, but not tRF-1s, are very abundant in AGO1, 3 and 4 photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) data and use canonical miRNA seed rules to associate with mRNAs. Analysis of AGO1 crosslinking, ligation, and sequencing of hybrids (CLASH) data suggests tRF-5s and tRF-3s may interact with thousands of different RNAs in human cells.

Despite the fact that tRFs are more evolutionarily conserved than miRNAs, are present in similar abundance, and are the only small RNA to display clear Argonaute sorting in humans, there is not a universally accepted nomenclature for tRFs or unique identifiers for different tRF sequences. As a result, it is possible for multiple labs to be working on the same tRF without noticing it, for example, cand45 in [12] is the same molecule as tRF-1001 in [11], or for a tRF to be misannotated as a miRNA [23]. To help the community study this new class of small RNA, we have created tRFdb [24], a relational database of tRNA derived RNA fragments, with all the tRF sequences which we have observed and unique identifiers. The names of datasets analyzed for each figure in this article are given in Table 1.

Results

tRFs are created by specific cleavage sites

We mapped small RNA reads from HEK293 cells to a collapsed tRNA gene [see Additional file 1: Figure S1] and, as expected, observed large numbers of reads that mapped to either the 5' end, 3' end or trailer sequence, corresponding to tRF-5s, tRF-3s and tRF-1s. Surprisingly, when we plotted the frequency of unique tRF reads of different lengths (Figure 1A) we observed three peaks for tRF-5s at ~15, ~22 and ~32 nts, and two peaks for tRF-3s at ~18 and ~22 nts. To the best of our knowledge, these distinct populations of tRF-5s and tRF-3s have never been reported before. The 5' ends of '3' tRNA halves' have 5' hydroxyl rather than a 5' phosphate and are biochemically different from tRFs and other small RNA. In addition, the tRNA halves are cleaved in the middle of the anticodon loop producing a 34 to 36 nt fragment making it possible to distinguish them from most tRF-5s (<32 bases) with 3' ends clearly in the stem and not in the anticodon loop itself. As shown in Figure 1C, we have developed a nomenclature for these subclasses of tRF-5s and tRF-3s: 3' cleavage at +5 (tRF-5a), +22 to +24 (tRF-5b) and +30 to +32 (tRF-5c), 5' cleavage at +55 (tRF-3b) and +59 to +60 (tRF-3a). The tRF-5 cleavage sites are in the D loop, D stem or the 5' half of the anticodon stem, while the tRF-3 cleavage sites are both in the TΨC loop. These tRF subclasses are seen in all human data sets analyzed from [25] and are also conserved in mice, but become less distinct further down the evolutionary tree [see Additional file 1: Figure S2].

Most of the tRF-1s observed in this data set are 15 to 22 bases long and always begin at the end of the tRNA

Table 1 Name of datasets analyzed for each figure

Figure Name	Analyzed library	
Figure 1A	GSM416733	
Figure 1B	GSM416733	
Figure 2A	GSM416733	HEK293
	GSM416753	HeLa
	GSM416754	U2OS
	GSM416755	143B
	GSM416756	A549
	GSM416757	H520
	GSM416758	SW480
	GSM416759	DLD2
	GSM416760	MCF7
	GSM416761	MB-MDA231
Figure 2B	GSM314552	Mouse-Esc
	GSM416732	Mouse-MEF
	GSM466487	Drosophila
	GSM604032	C.elegans
	GSM775340	S.cerevisiae
	GSM757894	S.pombe
	GSM1208316	R.sphaeroides
Figure 2C	GSM510432toGSM510435	Ovary
	GSM51043toGSM510439	Testes
	GSM510440toGSM510444	Brain
	GSM510445toGSM510456	Newborn
	GSM510457toGSM510460	E12.5
	GSM510465toGSM510468	E7.5
	GSM314552	ESC
Figure 3A-B	GSM416733	HEK293
Figure 4A-B	GSM314552	ESC_WT
	GSM314553	ESC_dcr–
	GSM314557	ESC_dgcr8–
Figure 4C-D	SRR029028	WT
	SRR029029	dcr-1
	SRR029030	dcr-2
Figure 4E	GSM466487	WT
	GSM466492	dcr-2
	GSM466496	r2d2
Figure 4F	GSM757894	WT
	GSM757897	dcr
Figure 4G	SRR207111	Whole-cell
	SRR207116	Nucleus
Figure 5A	GSM545212	AGO1
	GSM545213	AGO2
	GSM545214	AGO3
	GSB545215	AGO4
Figure 5B	GSM545212, GSM545213, GSM545214 and GSB545215 combined	
Figure 5C	GSM545212 dataset and the 17,319 CCRs reported in Hafner et al. [36].	
Figure 6	GSM1219487, GSM1219488, GSM1219489, GSM1219490, GSM1219491, GSM1219492 combined	
Figure 7	GSM1219487, GSM1219488, GSM1219489, GSM1219490, GSM1219491, GSM1219492 combined	

gene sequence which becomes a mature tRNA, and end with a RNA polymerase III (RNA pol III) transcription termination signal (UUUUU, UUCUU, GUCUU or AUCUU) [26,27]. Because the termination signal occurs at different locations in each pre-tRNA, tRF-1s are predicted to vary in length and, as expected, we found a broad length distribution (Figure 1A).

It is important to note that when looking at reads that map to a single tRNA, the peaks become much sharper (Figure 1B). The precision with which individual tRFs are generated strongly suggests that tRFs are not generated by random exonucleolytic digestion of longer precursors. In addition, because the method of small RNA sequencing in these data sets requires reverse transcriptase to read through the tRF into the adaptor sequence, tRNA modifications would, if anything, lower the number of reads that we are observing, not create artificial short sequences.

tRFs in different cell lines, organisms, and tissues

We next wanted to compare read counts for tRFs in cell lines other than HEK293 cells. We can observe read counts for all three classes of tRFs for all cell lines in the [25] data sets (Figure 2A). In general tRF-5s were present in higher abundance than tRF-3s, and tRF-3s were more abundant than tRF-1s.

To see if tRFs are present in other species we analyzed the publicly available small RNA data of mice [28], *Drosophila melanogaster* [29], Caenorhabditis *elegans* [30], Schizosaccharomyces *pombe* [31], *Saccharomyces cerevisiae* [32] and the bacterium *Rhodobacter sphaeroides* [3]. tRF-5s and tRF-3s are observed in all the species (Figure 2B). However fewer tRF-1s were observed in *Drosophila* (approximately 500 Reads Per Million (RPM)) and very few in *C. elegans*, *S. cerevisiae* and *R. sphaeroides*, although about 7,000 RPM of tRF-1s were detected in *S. pombe*.

The lower abundance of tRF-1s in the lower eukaryotes and absence in bacteria could be explained if the 3’ trailer sequences of pre-tRNAs were not in the 14 to 36-nucleotide range that was selected for cloning and sequencing. Indeed, 14 to 36 base-long pre-tRNA trailer sequences are ten-fold fewer in *C. elegans* and *S. cerevisiae*

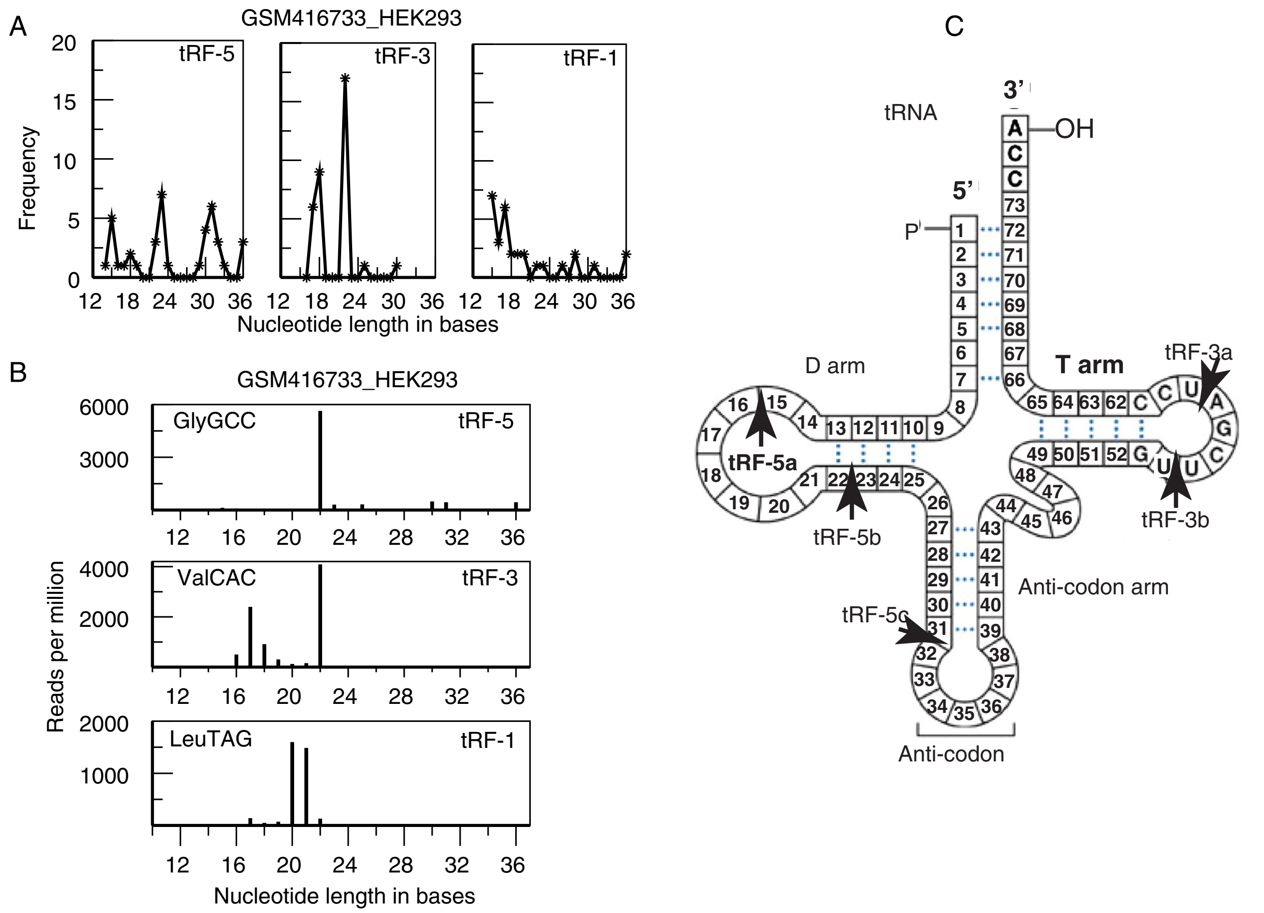


Figure 1 Non-random mapping of small RNAs (tRFs) on tRNA genes (HEK293 human cell line). (A) Numbers of unique tRFs that were present at a minimum of 20 reads per million are plotted against length of the tRF. **(B)** Length distribution of reads that mapped to a specific tRF-5 (GlyGCC), tRF-3 (ValCAC) and tRF-1 of (LeuTAG). **(C)** Illustration of a mature tRNA showing the cut sites that would generate the different subclasses of tRF-5s and tRF-3s.

compared to human and mouse [see Additional file 1: Figure S3], which could account for the fewer tRF-1s in the small RNA libraries from these species. However, *Drosophila* has comparable numbers of 3' trailers in the correct size range, and yet yielded fewer tRF-1, while *S. pombe* had fewer 3' trailers in the correct size range and yielded a large number of tRF-1 clones. Thus, some factor other than the possible number of 3' trailers in the correct size range, such as protein binding partners, helps determine how many tRF-1s are stable and identifiable in each species.

All the analyses of mammalian tRFs until now have been performed against RNA extracted from cell lines. To investigate if tRFs are also expressed in normal mammalian tissues, we analyzed the small RNA isolated from adult mouse ovary, testis and brain, and from mouse embryos and embryonic stem cells [28,33]. tRFs are present in all the tissues analyzed (Figure 2C), but the tRF-5s and tRF-3s were two- to five-fold less abundant in testes and brains compared to embryos, but as abundant in ovaries as in embryos. In contrast, the tRF-1s were less abundant in adult tissues with the highest level seen in brain, and that, too, was five- to twelve-fold less that in mouse embryos and embryonic stem cells.

All tRNAs do not produce three tRFs, and not all tRFs are equally abundant

To determine if all tRNA genes produce all three types of tRFs and, if they do, whether the tRFs are in comparable abundance, we selected those tRNA genes where a tRF-1 was detected in HEK293 cells at >20 RPM. A given tRF-1 has a unique sequence that can be assigned to a specific tRNA gene. When the 5' and 3' ends of more than one tRNA gene are identical in sequence, we classify them as a tRNA family. Thus, we compare the cloning frequency of a specific tRF-1 with that of the tRF-5 or −3 derived from the corresponding family of tRNA genes. A tRNA family represents a group of genes encoding the same tRF-5 or −3 sequences. These could include tRNA isoacceptors but are not necessarily so.

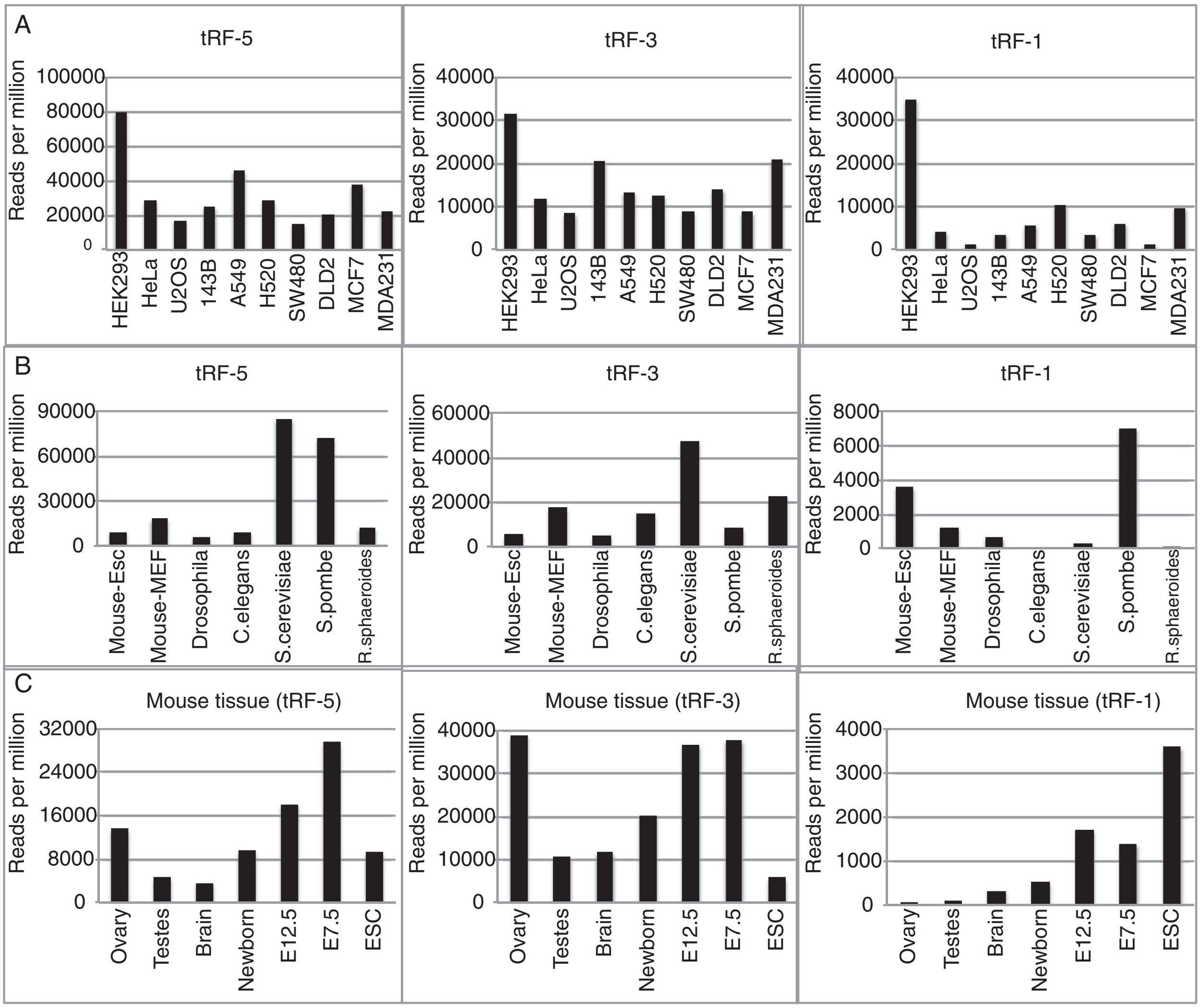


Figure 2 Presence of tRFs in bacteria to human. (A) Frequency of the three types of tRF in different human cell lines. tRF alignments that start with the first or second base of tRNA were collated as tRF-5 and whose 3′ end mapped to the 3′ end of tRNA and have a CCA at their 3′ end were categorized as tRF-3. tRFs whose 5′ end matched with the first or second bases of the 3′ trailer sequence of a tRNA were categorized as tRF-1. The number of tRF-5, tRF-3 and tRF-1 mapped in each cell line was normalized with the total number of reads in the analyzed library. **(B)** Shows the frequency of tRF-5, tRF-3 and tRF-1 in mouse embryonic stem cells, mouse cell line NIH3T3, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, *S. pombe* and *R. sphaeroides.* **(C)** tRF expression in different mouse tissues and embryonic stem cells (ESC).

The sequencing frequencies of these matched sets of tRFs were plotted (Figure 3A). Not all the tRF types are detected for a given tRNA gene and family. For example, tRF-5-Ser^{TGA} or tRF-3-Gly^{TCC} or -Leu^{AAG} are selectively absent even though tRF-1 were detected in all three cases.

When all three tRFs from a given tRNA gene or family are detected, their cloning frequencies are not similar. For example, tRNA4-leu^{TAA} produces a tRF-1 that is nearly 40- to 50-fold more abundant than the tRF-5 or −3 generated from the Leu^{TAA} tRNA family, and the Pearson correlation coefficients between tRF-5s and tRF-3s (R = −0.088), tRF-3s and tRF-1s (R = −0.224) and tRF-5s and tRF-1s (R = −0.099) are very low (Figure 3B). The lack of a correlation between the concentrations of tRF-5, −3 or −1 from a given tRNA gene (or family) further supports the hypothesis that tRFs are non-random, stable products derived from specific tRNAs and pre-tRNAs.

Processing of tRFs is distinct from miRNA biogenesis

To study the role of DICER1 in the generation of tRFs, we investigated the high throughput sequencing data of short RNAs from the wild type and *dicer1* mutants isolated under similar conditions from the same experiments. Such data were available for three species, that is, mouse [28], *S. pombe* [31] and two data sets for *Drosophila* [34,35]. Mutation of DICER1 (or Dicer-1 in *Drosophila*) did not significantly decrease the expression of any of the three

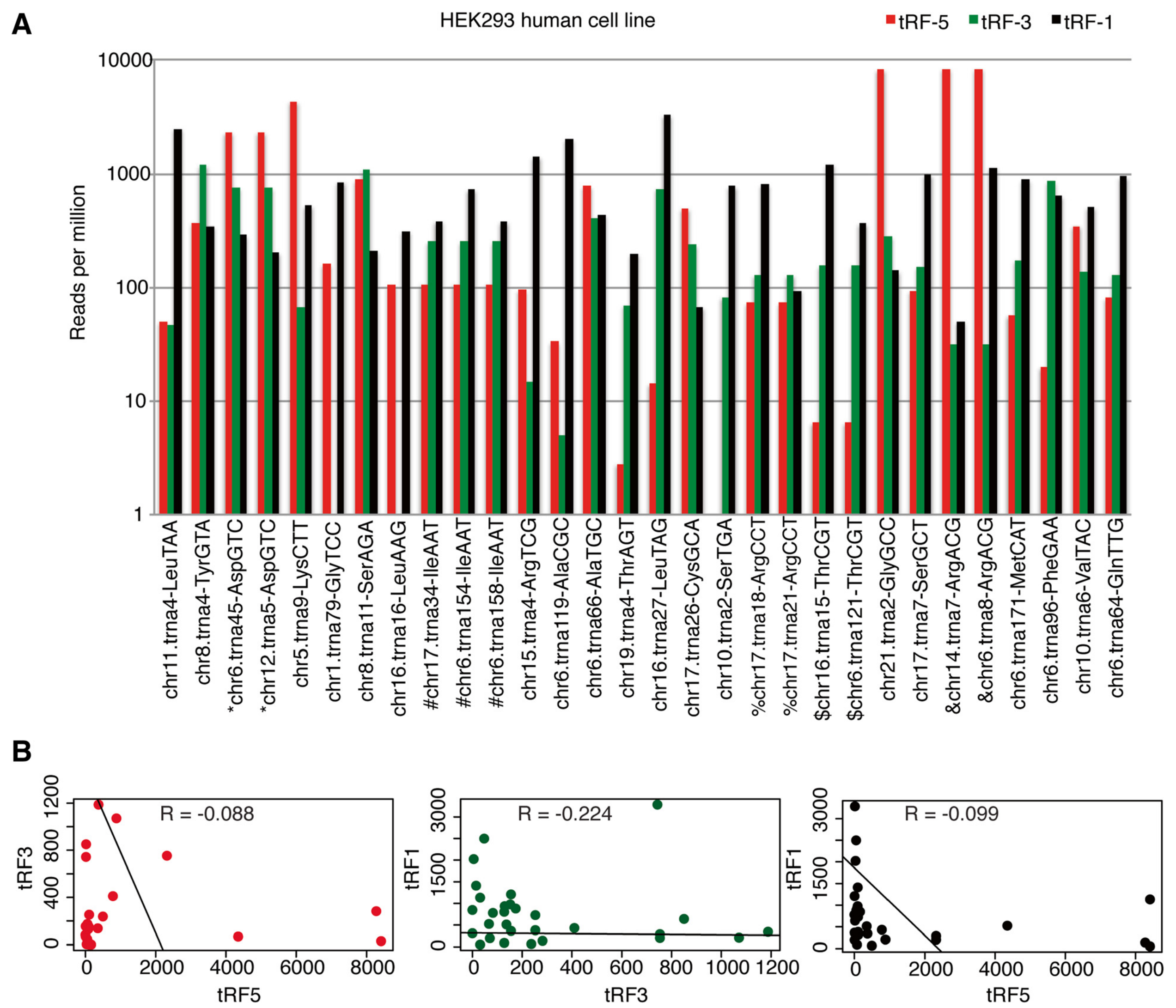


Figure 3 A given tRNA does not yield tRF-5, −3 and −1 at equal abundance. (A) Number of reads per million of tRF-5s, tRF-3s and tRF-1s from selected tRNA genes. The tRNA genes were selected on the basis of tRF-1s that had >20 reads per million in HEK293 human cell line library. The tRF-1s were compared with tRF-5s and -3s from the same tRNA, regardless of whether the tRF-5 or −3 was derived from that specific tRNA gene or other members of the tRNA gene family. The duplicated tRNA genes (tRNAs with the same anticodon) are marked with special character "*", "#", "$", "%" and &. **(B)** Scatter plots of tRF-3s versus tRF-5s, tRF-1s versus tRF-3s, and tRF-1s versus tRF-5s for the tRNA genes shown in **A** along with Pearson correlation coefficients.

classes of tRFs in mice (Figure 4A), *S. pombe* (Figure 4F) and *Drosophila* (Figure 4C and E), in contrast to the nearly hundred-fold suppression of the cloning frequency of several microRNAs in mouse (Figure 4B) and three- to twenty-fold suppression in *Drosophila* (Figure 4D). DGCR8 (an essential partner for the Microprocessor complex that cleaves pri-miRNA to generate pre-miRNA) was similarly dispensable for tRF generation (Figure 4A). Dicer-2 and the double strand RNA binding protein R2d2 are involved in the biogenesis of siRNA in *Drosophila*. Mutation of *dicer-2* or *r2d2* did not decrease the expression of tRF-5 or −1 either (Figure 4D-E). Although *r2d2* mutation decreased tRF-3 levels to about 40%, in the context of all the other mutants, we conclude that the proteins involved in generating canonical miRNAs or siRNAs are dispensable for the generation of tRFs in mice, *Drosophila* and *S. pombe*.

tRF-5s are nuclear while tRF-3s and -1s are cytoplasmic

To determine the cytoplasmic or nuclear location of tRFs we analyzed the small RNA of 18 to 30 bases isolated separately from nuclei and whole cell fraction of HeLa cell line [36] (Figure 4G). The tRF-5s were equally abundant in the whole cell and nuclear fractions, suggesting that they are mostly present in the nucleus, consistent with the observation of large numbers of tRF-5s

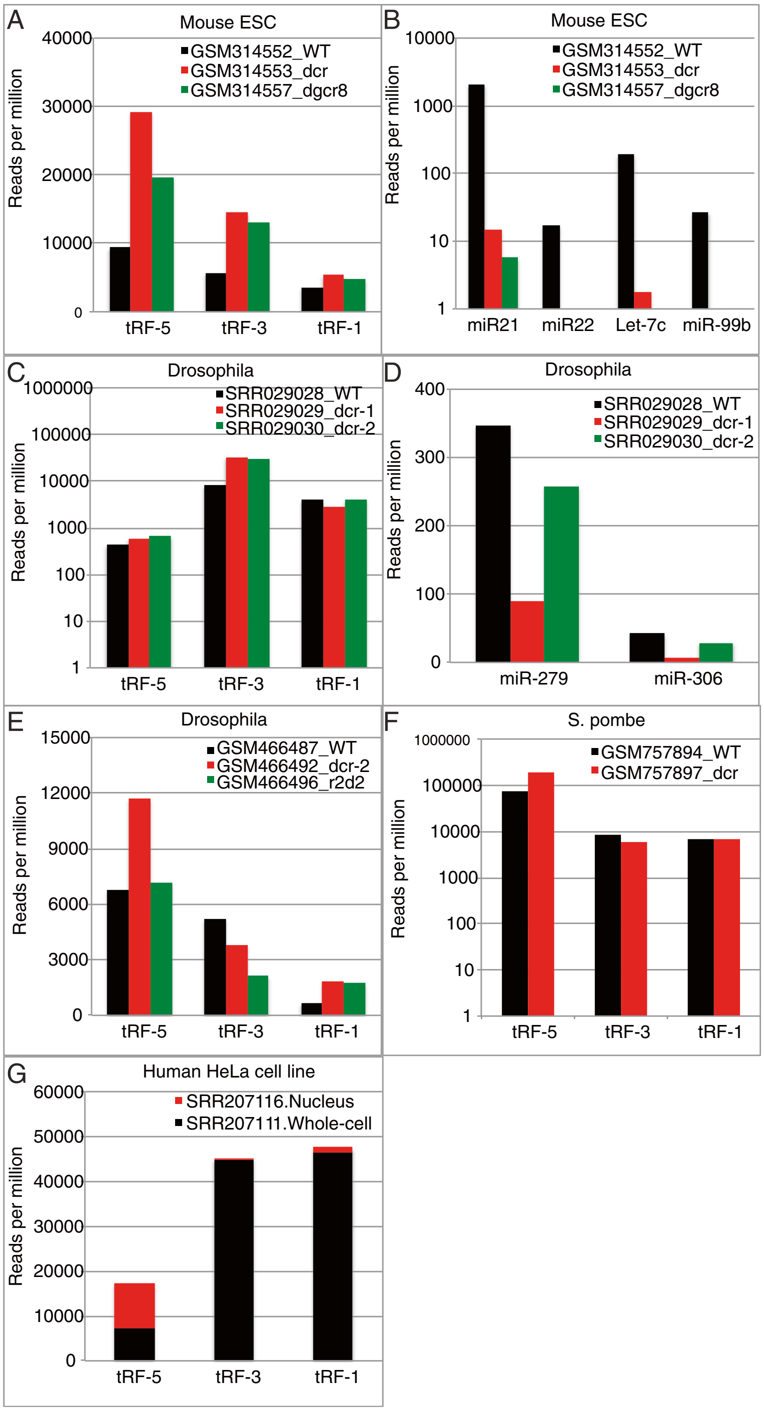


Figure 4 Processing of tRFs is distinct from miRNAs and tRF-3 and tRF-1 are mostly cytoplasmic. (A) tRF read counts in wild type, *dicer1* −/−, and *dgcr8*−/−mouse ES cells. **(B)** Same data sets as A, but read counts of various miRNAs are shown. **(C)** tRF read counts in *Drosophila* S2 cells either mock, *dicer-1* dsRNA, or *dicer-2* dsRNA treated. **(D)** Same data sets as C, but read counts of two miRNAs are shown. **(E)** tRF read counts from fly heads of either wild type, *dicer-2* mutant, or *r2d2* mutant flies. **(F)** tRF reads in wild type or *dcr1* delta *S. pombe*. **(G)** tRF read counts in HeLa cell nuclear fractionation or whole cell.

in Hela cell nucleoli [12]. tRF-3s and tRF-1s were much more abundant in the whole cell fraction compared to the nuclear fraction suggesting that both species are almost exclusively in the cytoplasm, which is consistent with the findings of [13]. The specific subcellular localization of the classes of tRFs raises questions about their biogenesis and functions. We note that in most analyzed short RNA libraries we see an abundance of tRFs in the order tRF-1 < tRF-3 < tRF-5 (Figure 2A); however, the reverse trend was observed in short RNA libraries generated from the whole cell and nuclear fractions (Figure 4G). This may be due to variations in the protocol used by the lab that produced these libraries, but the effect should be the same on both the nuclear and whole cell libraries. Thus, we do not expect such variations to uniquely enrich tRF-5 in the nuclear fraction compared to tRF-3 or tRF-1.

tRF-5s and tRF-3s associate with AGO1, 3, and 4

We investigated the association of tRFs with human Argonautes by analyzing the human AGO1, 2, 3 and 4 PAR-CLIP data isolated from HEK293 cell lines [37]. In PAR-CLIP, when the 4-thiouridine is crosslinked to the protein of interest, it often becomes mutated to a cytidine during library preparation. Positional T to C mutation analysis of the data provides information about the RNA-protein interaction. In the presented analysis we allow 1 T/C mutation and give preference for perfect mappings (see Methods for details). Read counts for tRF-5s and tRF-3s are comparable to miRNAs for AGO1, AGO3 and AGO4, but are nearly absent in AGO2, while there are almost no read counts for tRF-1s for all four Argonautes (Figure 5A). Since this is the first time a class of small RNA has been reported to show differential human Argonaute sorting, we were interested if we could observe this trend in other data sets. AGO1 and AGO2 HEK293 cell PAR-CLIP was also performed by [38] and analysis of this data again showed this same pattern (data not shown). Unfortunately, we are unaware of any mouse AGO1, 3 or 4 CLIP or PAR-CLIP data, so we could not repeat this analysis for mice, but we do note that only very small numbers of tRFs are seen in mouse AGO2 CLIP data [39] (data not shown).

tRF-3s and tRF-5s bind to human Argonautes like miRNAs

As reported in [37], miRNAs are crosslinked to the AGO protein at specific positions, namely positions 9 to 13, and this is borne out by a high T to C mutation frequency at these positions and very low mutational frequency at other positions, particularly the first 7 positions, that constitute the ‘seed’ and are involved in base-pairing with the target RNA. We first checked if we could replicate these results with our algorithms (Figure 5B), and we were able to detect a high percentage of T to C mutations at positions 9 to 13 for miRNAs and a very low frequency at the first 7 positions. We next checked whether tRF-3s display a similar pattern of mutations since this would indicate a similar binding mode and perhaps function. As with miRNAs, we saw a very low mutation frequency for the first 6 positions and peaks between positions 8 to 12. The slight difference in mutational frequencies between miRNAs and tRF-3s could be due to a sampling bias since tRF-3s contain fewer Ts than miRNAs, and the Ts that are present are not as randomly distributed as in miRNAs. Alternatively, this difference could indicate a biologically relevant distinction in the way tRF-3s interact with Argonautes. The T to C mutational frequency of tRF-5s also shows protection from cross-linking of the first six residues (Figure 5B (lower panel), suggesting that these bases are facing away from the Argonaute in an orientation suitable for binding to a target RNA, that is, they represent a seed region. The maximal T-C change is observed at base 7, and not at bases 9 to 13, unlike what is observed with microRNAs and tRF-3s. This is partly because tRF-5s are enriched in Us at base 7, but may also be because some of them (tRF-5c, Figure 1A) form a slightly different complex with Argonautes because they are longer at 32 bases than miRNAs and tRF-3s.

tRF-3s associate with target RNAs via canonical seed sites

In addition to miRNAs being cross-linked at specific positions in PAR-CLIP, target RNAs are also preferentially cross-linked at a certain position with respect to the RISC complex: in the middle of the complex immediately preceding the sequence annealed to the microRNA seed. This information was used in [37] to generate 17,319 crosslink-centered regions (CCRs) of RNAs present in the PAR-CLIP data. CCRs are 41 nt long sequences centered at the T that showed the highest T to C frequency. They demonstrated that the reverse complement of known miRNA seeds is enriched in CCRs directly following this central cross-linked T. We reproduced this observation by taking the 50 most abundant miRNAs in AGO1 PAR-CLIP data and scanning the 17,319 CCRs with different seed definitions (Figure 5C). As expected, the canonical seed sites 7mer-A1 (target has an A at nucleotide position 1 and matches positions 2 to 7 of microRNA) and 7mer-m8 (target matches positions 2 to 8 of microRNA) produce the largest number of matches, and at the expected position in the CCRs, immediately downstream from the cross-linked T. The less canonical seed sequences, such as bases 3 to 9 of the microRNAs, produce fewer matches with the CCRs, and microRNA positions that have never been recognized as seeds produce even fewer.

Given that we saw similar patterns of T to C mutations for tRF-3s and miRNAs, we were emboldened to test whether the 5’ ends of tRF-3s act as seeds to select matching target CCRs, with the match at a location in the CCR that is immediately downstream of the central cross-linked T. As with miRNAs, the two best seed

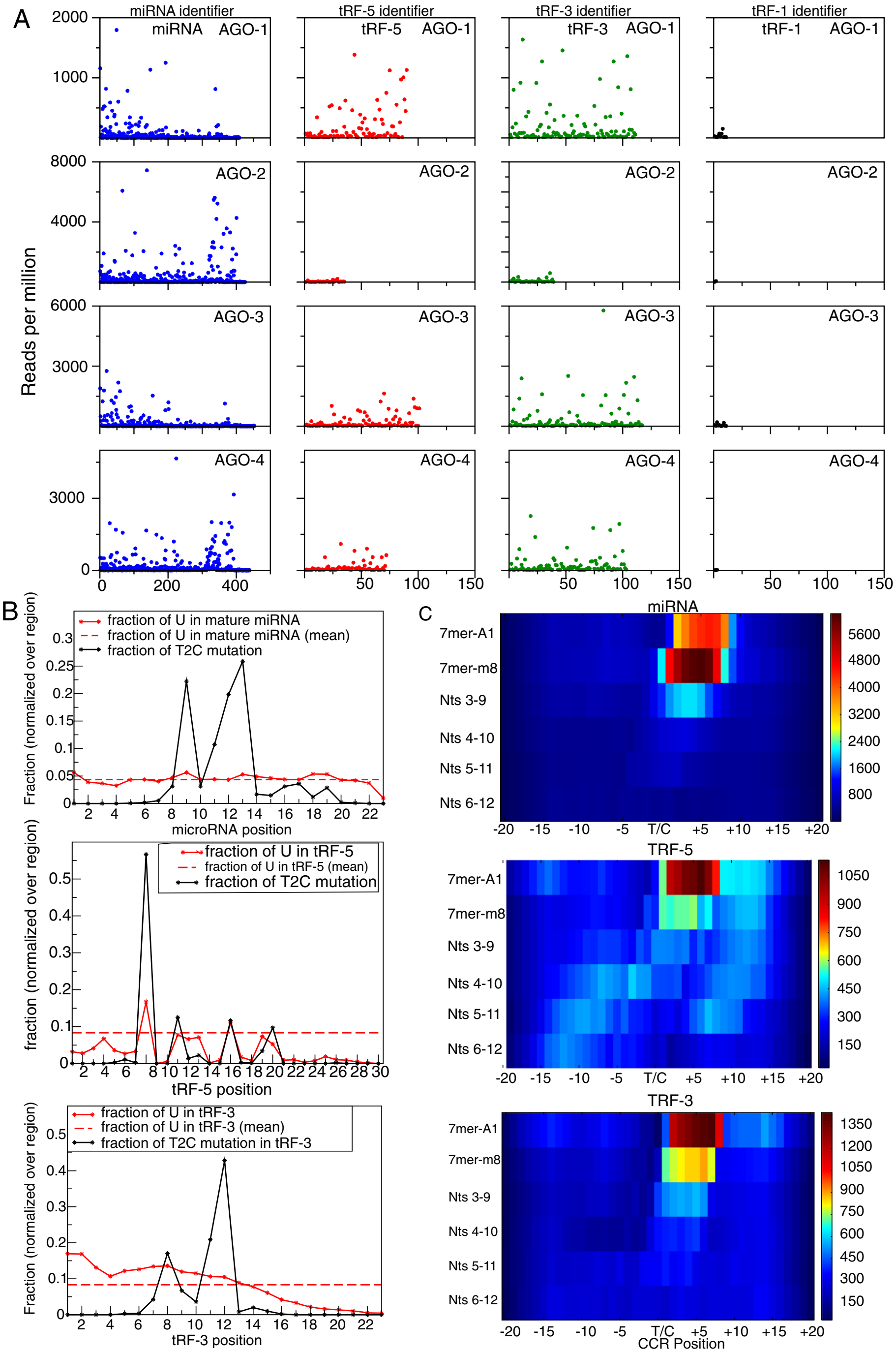


Figure 5 PAR-CLIP analysis of miRNAs and tRFs. (A) Read counts for miRNAs, tRF-5s, tRF-3s and tRF-1s in AGO 1 to 4 PAR-CLIP data from Hafner *et al* [37]. Each microRNA and tRF is given an identifying number. The expression level of each microRNA and tRF is shown on the Y-axis and the assigned number is shown on the X-axis. **(B)** Normalized positional T to C mutation frequencies for miRNA, tRF-5 and tRF-3 reads found in the AGO 1 to 4 PAR-CLIP data. **(C)** Matches of canonical and noncanonical seeds of the 50 most abundant miRNAs, tRF-5s or tRF-3s seen in the AGO1 dataset to the 17,319 CCRs reported in Hafner *et al.*

definitions of tRF-3s that matched CCRs are the canonical 7mer-A1 and 7mer-m8 (Figure 5C). However, while miRNAs seemed to show a preference for the 7mer-m8 site over the 7mer-A1 site, tRF-3s had the opposite preference. Again, this could represent sampling bias, or it may hint at a RNA-induced silencing complex (RISC) that is fundamentally different. Significantly, the best complementary matches to the tRF-3 seeds were also immediately downstream of the central T in the CCRs, just as seen with the miRNAs. Seeds of tRF-5s also show matches above background with canonical sites in the CCRs (Figure 5C), but tRF-1s do not show matches above background [see Additional file 1: Figure S4].

CLASH data indicates tRF-3s and tRF-5s target thousands of RNAs

CLASH is a new technique that has recently been used to study the AGO1-miRNA-target RNA interactome in HEK293 cells [40]. Briefly, the technique is similar to CLIP, but with the addition of a ligation step that connects the 3’ end of the AGO bound small RNA to the 5’ end of the target RNA. To analyze this data we found reads that started with either a miRNA or tRF, and then performed blastn with the rest of the sequence against human Ref-Seq RNA (see Methods). In our analysis we see 187 HOXC8-mir-196a/b chimeric reads, which corresponds closely to the 191 chimeric reads identified by Helwak *et al* [40]. Surprisingly, despite miRNAs being more abundant, we saw more tRF-3-mRNA chimeras than miRNA-mRNA chimeras (Figure 6A-B). We also observed numerous tRF-5-mRNA chimeras, but very few tRF-1-mRNA chimeras, which is consistent with our PAR-CLIP analysis (Figure 6B). Manual observation of some of the more abundant tRF-3-mRNA chimeras shows nice clustering of the mRNA portions of the reads (Figure 6C). Examples of some of the most abundant tRF-mRNA interactions are shown with their predicted mfold structures (Figure 7), and a list of all tRF-mRNA chimeric reads can be found in Additional file 2.

Discussion

The recent increase in small RNA-Seq data and novel methods to investigate the miRNA-interactome allowed us to perform a detailed analysis of the properties of tRFs. We have shown that tRFs are very precisely generated fragments that are present in all cell lines investigated and in organisms ranging from humans to bacteria. Using well-established PAR-CLIP data we showed that tRF-5s and tRF-3s clearly associate with human Argonautes 1, 3 and 4, but show little to no association with AGO2. Although Argonaute sorting is common in some organisms such as plants, this has not been seen in humans before. tRF-5s and tRF-3s also match clusters observed in PAR-CLIP data with canonical seed rules, indicating that AGO-tRF complexes are able to associate with RNAs. CLASH analysis confirmed this hypothesis by the observation of large numbers of tRF-mRNA chimeras.

This investigation raises many questions about the biology of tRFs and small RNAs in general. If DICER1 and DROSHA are not involved in tRF generation then which proteins are? Do all organisms generate tRFs by the same pathway? Do tRFs associate with Argonautes in other organisms? Is the lack of association with AGO2 telling us something about RISC assembly in humans? And will the traditional methods for studying miRNAs (such as antisense RNAs) be applicable to a small RNA whose parental RNA is one of the most abundant RNAs in the cell?

These questions are too much for any one group to answer, but it is tempting to speculate as to the possibilities. tRNAs are heavily modified and it is known that some of these modifications affect tRNA stability. Indeed, lack of a specific tRNA modification has been shown to increase tRNA half generation [41]. This suggests that there may be other modifications which can affect tRF-5 or tRF-3 generation.

The matches to PAR-CLIP clusters and the large number of CLASH chimeras point to a role for tRFs in RNA silencing. In fact, it is known that a large number of CLIP-Seq clusters are not able to be assigned to miRNA seeds [42]. Thus far, the scientific community has explained this fact by proposing noncanonical seeds, such as those that contain bulges, but it may be that these orphan clusters are targeted by tRFs. However, it is important to keep an open mind for the function of tRFs. For example, a recent publication showed that a tRF-3 which our lab previously identified is able to serve as a primer for HTLV-1 reverse transcriptase [43]. In addition, some of the most abundant tRF-3 chimeras we observe are with histone mRNAs. Histone mRNAs are known as the only mRNAs in the cell to not contain a poly-A tail, thus these interactions cannot result in traditional degradation of the mRNA target. However, the binding site for a large number of the interactions is very close to the stem loop, indicating the tRF-3s could compete with Stem-loop binding protein and affect the mRNA stability via this mechanism.

In addition, although miRNA-Argonaute complexes are traditionally thought to function in the cytoplasm, increasingly diverse functions for Argonautes in the nucleus continue to be discovered (reviewed in [44]). For example, there have been recent reports of transcriptional gene silencing by short RNAs and regulation of alternative splicing by Argonaute and related PIWI proteins in mammals. The presence of tRF-5s in the nucleus and the association of tRF-5s with Ago1, Ago3 and Ago4 suggest that tRF-5s may participate in these processes.

Although there has been a steady increase in tRF literature since their discovery, our current knowledge of tRFs clearly pales in comparison to other small RNAs. We have shown that tRFs display similar properties to miRNAs, and given the importance of miRNAs in processes ranging from development to cancer, it is not farfetched to imagine equally important functions for tRFs.

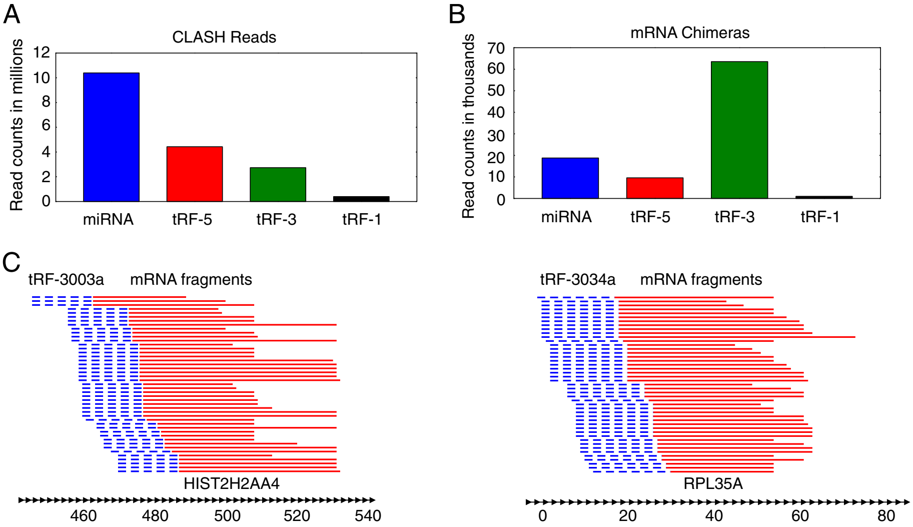


Figure 6 tRF-mRNA chimeras are abundant in AGO1 CLASH data. (A) Numbers of CLASH reads that started with a perfect match to a miRNA, tRF-3, tRF-5 or tRF-1 and deemed to not be pre-miRNA, tRNA or a RNA in Ref-Seq. **(B)** Numbers of miRNA, tRF-5, tRF-3 or tRF-1 chimeras with mRNAs. **(C)** Alignments of the mRNA portion of the 45 most abundant reads for the tRF-3003a-HIST2H2AA4 interaction and the tRF-3034a-RPL35A interaction to the corresponding mRNA. The tRF-3 portion of the reads is depicted in dashed blue while the mRNA fragment is depicted in red. CLASH, cross-linking, ligation, and sequencing of hybrids.

Conclusions

tRFs are a newly discovered class of small RNA that are highly abundant in different human cell lines, mouse tissues and organisms ranging from bacteria to humans. Individual tRFs show a narrow size distribution, suggesting that the fragments are precisely generated and not degradation products of tRNAs. Mutation of different components of the miRNA biogenesis pathway does not have an effect on tRF levels, and tRFs are seen in organisms that do not contain miRNAs, indicating tRF generation is distinct from miRNA biogenesis. In human HEK 293 cells tRF-5s and tRF-3s are associated with Argonautes 1, 3 and 4 as evidenced by PAR-CLIP data. These tRFs contain seed sequences, which match the central portion of large numbers of CCRs. This observation, along with the finding of thousands of tRF-mRNA chimeras in CLASH data, indicates tRF-3s and tRF-5s can target RNAs in a manner similar to miRNAs.

Methods

Analysis of the small RNA data

The data analyzed in this manuscript were downloaded from either the GEO database [45] or NCBI SRA database [46]. We considered only those sets of high throughput sequencing data where small RNAs of 14 to 36 bases long were size selected and then sequenced. For each dataset we looked for the processed sequence along with its cloning frequency. In case of non-availability of processed data, the raw data were used to generate the unique sequence and its cloning frequency. The adaptor sequences from the raw data were removed using the ‘Cutadapt’ (version 1.0) program [47].

Building and mapping of small RNA on ‘tRNAdb’

Information about the tRNA genes for each species (Human hg19; Mouse mm9; Drosophila dm3; C. elegans ce6; S. cerevisiae sacCer1; S. pombe schiPomb1) was downloaded from the ‘Genomic tRNA database’ [48]. For each tRNA gene the DNA sequences ranging from 100 bases upstream of the start of mature tRNA to 200 bases downstream of the end of mature tRNA were extracted from the same genome assembly on which the tRNA gene coordinates were built. A species-specific tRNA database called tRNAdb’ was built. To find the tRNA-related RNA sequences in each library, the small RNAs were mapped on the species-specific tRNAdb, using BLASTn [49]. In general we considered only those alignments where the query sequence (small RNA) was mapped to the database sequence (tRNA) along 100% of its length. The blast output file was parsed to get information on the mapped position of small RNA on tRNA genes. We extracted all map positions where the small RNA aligned from its first base to the last base with the tRNA sequence allowing either one or no mismatch. Since ‘CCA’ is added at the 3’ end of tRNA by tRNA nucleotidyltransferase during maturation of tRNA [50], we allowed a special exception for the small RNA mapping to the 3’ ends of tRNAs in the tRNAdb allowing a terminal mismatch of < =3 bases. To remove any false positives, the small RNAs that mapped on to the ‘tRNAdb’ were again searched against the whole genome using blast search

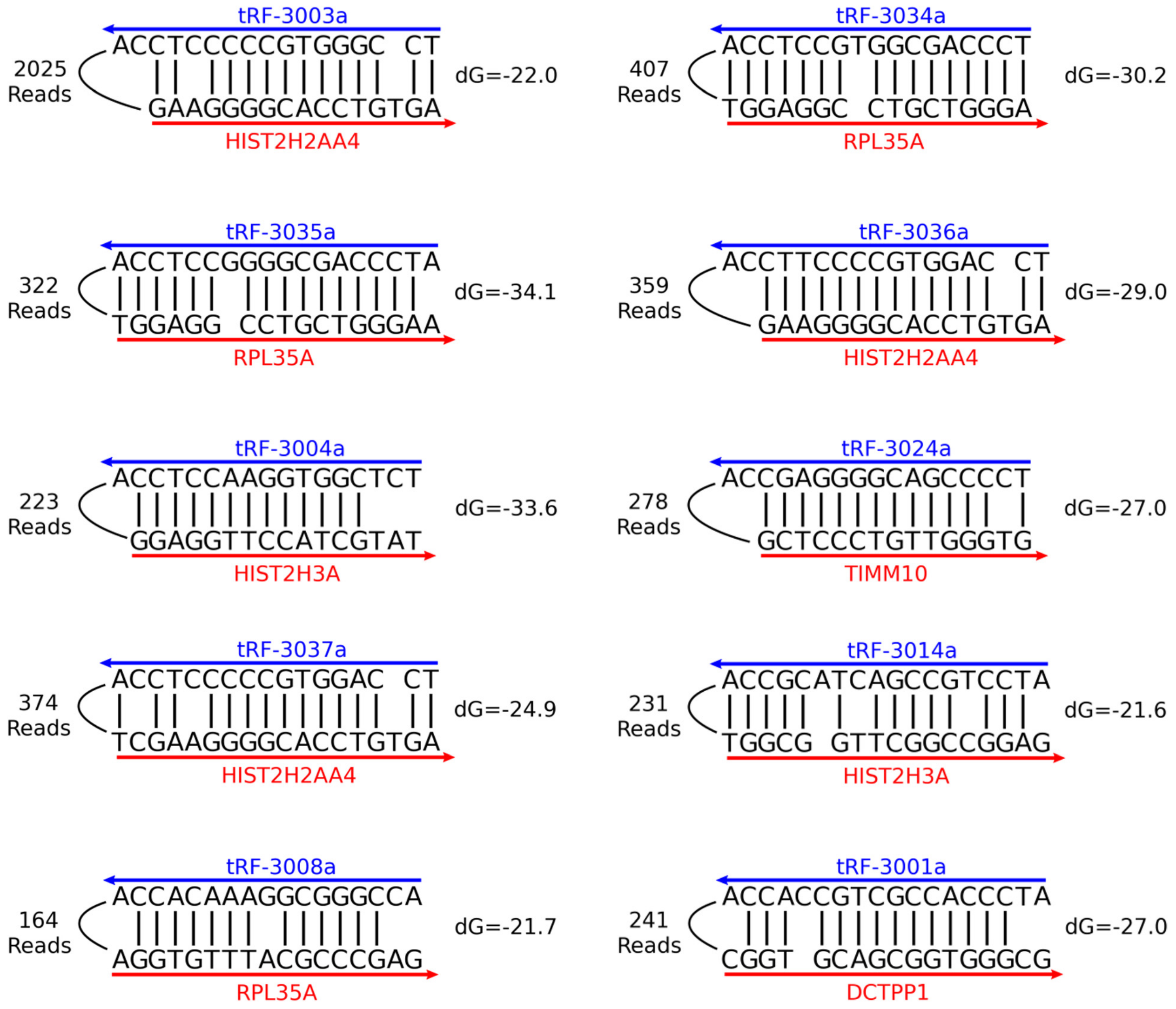


Figure 7 Examples of tRF-3-mRNA CLASH chimeras. The most abundant read of 10 of the most prevalent tRF-3-mRNA interactions found in our analysis were analyzed with mfold's RNA Folding Form using default settings. The output of mfold is represented with the tRF depicted 3′ to 5′ and the mRNA sequence 5′ to 3′. For each interaction the number of reads supporting the interaction is shown to the left, and the delta G from mfold is shown to the right. CLASH, cross-linking, ligation, and sequencing of hybrids.

excluding the tRNA loci. Only those small RNAs were qualified as tRFs that mapped exclusively on tRNAdb.

PAR-CLIP data analysis

We included the mature miRNA (miRNA:miRBase v20; genome-build-id: GRCh37.p5) and mRNA sequences in our previously built human specific tRNAdb that was used to query the expression level of tRFs and miRNA. We investigated tRF and miRNA expression with human Argonautes by analyzing the human Ago1 (GEO ID = GSM545212), 2 (GEO ID = GSM545213), 3 (GEO ID = GSM545214) and 4 (GEO ID = GSM545215) PAR-CLIP data isolated form HEK293 cell lines [37]. Data of all four small RNA libraries (AGO1 to 4) were combined together to examine the T to C mutation position and its frequency compared to wild type small RNA (miRNAs and tRFs). Sequence reads either mapped perfectly on miRNA or tRFs or mapped with one base mismatch were considered for T to C mutation analysis. Mismatched base and its position relative to the 5′ end of small RNA were collected for final analysis.

CLASH data analysis

The miRNA, tRF-5, tRF-3 and tRF-1 analyses were performed separately. For the miRNA analysis, reads were found that started with a mature miRNA allowing no mismatches, giving preference to longer miRNAs. For the tRF-5 analysis, reads were found that started with a sequence that mapped to the first 14 to 33 nucleotides of a tRNA, allowing no mismatches, giving preference to longer tRF-5s. For the tRF-3 analysis, reads were found that started with a sequence that mapped to the last 17 to 23 nucleotides of a mature tRNA, allowing no mismatches, giving preference to longer tRF-3s. For the tRF-1 analysis, reads were found that started with a sequence that mapped to the first 14 to 33 nucleotides of a tRNA trailer, allowing no mismatches, giving preference to longer tRF-1s. For the miRNA analysis, the reads were confirmed to not be pre-miRNAs by running blastn, word size 7, default scoring matrix, against a database composed of miRNA hairpins from miRBase. For the tRF-5 analysis, the reads were confirmed to not be longer tRNA fragments or full length tRNAs by running blastn, word size 7, default scoring matrix, against a database composed of mature tRNA sequences. All reads were checked to not be RNA fragments by performing a blastn search against the human Ref-Seq database using blastn, word size 7, default scoring matrix, 20 maximum hits. Reads that had a hit which either overlapped 6 or more bases of the small RNA, or 6 or more bases of an 18 base minimal small RNA sequence for the longer small RNAs, and had an e-value less than or equal to .001, were discarded. For all analyses the portion of the read following the small RNA sequence was searched against the human RNA Ref-Seq database using blastn, word size 7, default scoring matrix, 20 maximum hits. Because blastn is a local aligner, a conservative approach was taken to adaptor removal. Adaptor was removed if a perfect match was found for 12 or more bases of the adaptor, or 1 mismatch for 21 or more bases of adaptor, or 2 mismatches for 26 or more bases of adaptor. Reads were considered chimeras if a hit was found within four nucleotides of the end of the small RNA and had an e-value less than or equal to .01. Because the search space can affect the e-value, reads still possibly containing adaptor sequence and having a borderline e-value underwent adaptor removal with Biopython's pairwise2 local aligner and were blast searched again to get an updated e-value. Many reads matched more than one transcript in Ref-Seq. To identify the most likely transcript, every read of the CLASH data was searched against the human Ref-Seq database using blastn, word size 7, default scoring matrix, 20 maximum hits. All hits with an e-value less than .1 were tabulated. For the chimeric reads, the most likely transcript was deemed to be the transcript which was most abundant in the data. If a tie still occurred, NM transcripts were given preference to XM transcripts, XM given preference to NR, and NR given preference to XR. All chimeras whose most likely transcript was either NM or XM were deemed to be a small RNA-mRNA chimera. All such tRF chimeras are reported in Additional file 2, along with up to 19 other possible transcripts sorted by likelihood.

Additional files

Additional file 1: Figure S1. Non-random mapping of small RNA (tRFs) on tRNA genes in HEK293 cell lines. tRNA gene co-ordinates were collapsed to 1–73 bases long mature tRNA. The scale 1 to 73 on the x-axis is the 1st to 73rd base of mature tRNA gene. The 5′ and 3′ ends of tRFs mapped on tRNA were recorded. The number of tRF ends that map to a specific base of tRNA locus is shown. The dotted lines predict the three types of tRFs. **Figure S2.** Non-random mapping of small RNA (tRFs) on tRNA genes in other species. The axes and other details are same as given in **Figure S1** legend. The number of tRF ends (5′ or 3′) mapped at each base given as reads per million in: mouse embryonic stem cells, mouse cell line NIH3T3, *D. melanogaster*, *C. elegans*, *S. cerevisiae* and *S. pombe*. **Figure S3.** Predicted length distribution of tRNA trailer sequences in different organisms. The computational prediction of length distribution of tRNA trailer sequences (potential tRF-3 s) in human, mouse, Drosophila, *C. elegans*, *S. cerevisiae* and *S. pombe*. **Figure S4.** Matches of canonical and noncanonical seeds of the 50 most abundant tRF-1s seen in the AGO1 dataset to the 17,319 CCRs reported in Hafner *et al.* [37].

Additional file 2: Excel file containing all tRF-mRNA chimeric reads observed in the CLASH data. There are separate workbooks for tRF-1s, tRF-3s, and tRF-5s. The most likely to least likely mapping of the mRNA portion of the read is listed left to right with up to 20 transcripts listed.

Abbreviations

CCR: crosslink-centered region; CLASH: cross-linking, ligation, and sequencing of hybrids; miRNA: microRNA; PAR-CLIP: photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation; RISC: RNA-induced silencing complex; RPM: reads per million; siRNA: small interfering RNA; tiRNA: tRNA-derived stress-induced fragments; tRF: tRNA-derived RNA fragment.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PK was responsible for Figures 1, 2, 3 and 4 and parts of Figure 5. JA was responsible for Figures 6 and 7 and part of Figure 5; PK, JA and AD wrote the paper. All authors read and approved the final manuscript.

Authors' information

Pankaj Kumar and Jordan Anaya are joint first authors.

Acknowledgments

We thank all members of the Dutta Lab for helpful discussions. This work was supported by P01CA104106 and R01GM84465 to AD. Pankaj Kumar is supported by Department of Defense prostate cancer fellowship (W81XWH-13-1-0088). Jordan Anaya is supported by the UVA Medical Scientist Training Program.

Data access

The GEO or SRA number of the data used in this study is given in Table 1 and all the data used in this study are publicly available on the GEO database [45].

Author details

[1]Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, VA 22901, USA. [2]Department of Computer Science and Engineering, Grandhi Varalakshmi Venkatarao Institute of Technology (GVIT), Bhimavaram, Andhra Pradesh 534207, India. [3]PO Box 800733, 1340 Jefferson Park, Ave Jordan Hall Room 1232, Charlottesville, VA, USA.

Received: 21 August 2014 Accepted: 18 September 2014
Published online: 01 October 2014

References

1. Chu CY, Rana TM: **Small RNAs: regulators and guardians of the genome.** *J Cell Physiol* 2007, **213:**412–419.
2. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes.** *Science* 2007, **315:**1709–1712.

3. Olovnikov I, Chan K, Sachidanandam R, Newman Dianne K, Aravin Alexei A: **Bacterial argonaute samples the transcriptome to identify foreign DNA.** *Mol Cell* 2013, **51:**594–605.
4. Tuck AC, Tollervey D: **RNA in pieces.** *Trends Genet* 2011, **27:**422–432.
5. Levitz R, Chapman D, Amitsur M, Green R, Snyder L, Kaufmann G: **The optional E.coli prr locus encodes a latent form of phage T4-induced anticodon nuclease.** *EMBO J* 1990, **9:**1383–1389.
6. Thompson DM, Parker R: **Stressing out over tRNA cleavage.** *Cell* 2009, **138:**215–219.
7. Ivanov P, Emara MM, Villen J, Gygi SP, Anderson P: **Angiogenin-induced tRNA fragments inhibit translation initiation.** *Mol Cell* 2011, **43:**613–623.
8. Gebetsberger J, Zywicki M, Kunzi A, Polacek N: **tRNA-derived fragments target the ribosome and function as regulatory non-coding RNA in Haloferax volcanii.** *Archaea* 2012, **2012:**260909.
9. Yuan Y, Hwang ES, Altman S: **Targeted cleavage of mRNA by human RNase P.** *Proc Natl Acad Sci U S A* 1992, **89:**8006–8010.
10. Elbarbary RA, Takaku H, Uchiumi N, Hiroko T, 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.
11. Lee YS, Shibata Y, Malhotra A, Dutta A: **A novel class of small RNAs: tRNA-derived RNA fragments (tRFs).** *Genes Dev* 2009, **23:**2639–2649.
12. Cole C, Sobala A, Lu C, Thatcher SR, Bowman A, Brown JW, Green PJ, Barton GJ, Hutvagner G: **Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs.** *RNA* 2009, **15:**2147–2160.
13. Haussecker D, Huang Y, Lau A, Parameswaran P, Fire AZ, Kay MA: **Human tRNA-derived small RNAs in the global regulation of RNA silencing.** *RNA* 2010, **16:**673–695.
14. Pederson T: **Regulatory RNAs derived from transfer RNA?** *RNA* 2010, **16:**1865–1869.
15. Garcia-Silva MR, Cabrera-Cabrera F, Güida MC, Cayota A: **Hints of tRNA-derived small RNAs role in RNA silencing mechanisms.** *Genes (Basel)* 2012, **3:**603–614.
16. Gebetsberger J, Polacek N: **Slicing tRNAs to boost functional ncRNA diversity.** *RNA Biol* 2013, **10:**1798–1806.
17. Li Z, Ender C, Meister G, Moore PS, Chang Y, John B: **Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs.** *Nucleic Acids Res* 2012, **40:**6787–6799.
18. Burroughs AM, Ando Y, de Hoon ML, Tomaru Y, Suzuki H, Hayashizaki Y, Daub CO: **Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin.** *RNA Biol* 2011, **8:**158–177.
19. Maute RL, Schneider C, Sumazin P, Holmes A, Califano A, Basso K, Dalla-Favera R: **tRNA-derived microRNA modulates proliferation and the DNA damage response and is down-regulated in B cell lymphoma.** *Proc Natl Acad Sci U S A* 2013, **110:**1404–1409.
20. 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.
21. Sobala A, Hutvagner G: **Small RNAs derived from the 5′ end of tRNA can inhibit protein translation in human cells.** *RNA Biol* 2013, **10:**553–563.
22. Saikia M, Fu Y, Pavon-Eternod M, He C, Pan T: **Genome-wide analysis of N1-methyl-adenosine modification in human tRNAs.** *RNA* 2010, **16:**1317–1327.
23. Schopman NCT, Heynen S, Haasnoot J, Berkhout B: **A miRNA-tRNA mix-up: tRNA origin of proposed miRNA.** *RNA Biol* 2010, **7:**573–576.
24. **tRFdb: A Relational Database of tRNA Related Fragments.** http://genome.bioch.virginia.edu/trfdb/.
25. Mayr C, Bartel DP: **Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.** *Cell* 2009, **138:**673–684.
26. Hagenbuchle O, Larson D, Hall GI, Sprague KU: **The primary transcription product of a silkworm alanine tRNA gene: identification of in vitro sites of initiation, termination and processing.** *Cell* 1979, **18:**1217–1229.
27. Koski RA, Clarkson SG: **Synthesis and maturation of Xenopus laevis methionine tRNA gene transcripts in homologous cell-free extracts.** *J Biol Chem* 1982, **257:**4514–4521.
28. Babiarz JE, Ruby JG, Wang Y, Bartel DP, Blelloch R: **Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs.** *Genes Dev* 2008, **22:**2773–2785.
29. Ameres SL, Horwich MD, Hung JH, Xu J, Ghildiyal M, Weng Z, Zamore PD: **Target RNA-directed trimming and tailing of small silencing RNAs.** *Science* 2010, **328:**1534–1539.
30. de Lencastre A, Pincus Z, Zhou K, Kato M, Lee SS, Slack FJ: **MicroRNAs both promote and antagonize longevity in C. elegans.** *Curr Biol* 2010, **20:**2159–2168.
31. Barraud P, Emmerth S, Shimada Y, Hotz HR, Allain FH, Buhler M: **An extended dsRBD with a novel zinc-binding motif mediates nuclear retention of fission yeast Dicer.** *EMBO J* 2011, **30:**4223–4235.
32. Drinnenberg IA, Fink GR, Bartel DP: **Compatibility with killer explains the rise of RNAi-deficient fungi.** *Science* 2011, **333:**1592.
33. Chiang HR, Schoenfeld LW, Ruby JG, Auyeung VC, Spies N, Baek D, Johnston WK, Russ C, Luo S, Babiarz JE, Blelloch R, Schroth GP, Nusbaum C, Bartel DP: **Mammalian microRNAs: experimental evaluation of novel and previously annotated genes.** *Genes Dev* 2010, **24:**992–1009.
34. Zhou R, Czech B, Brennecke J, Sachidanandam R, Wohlschlegel JA, Perrimon N, Hannon GJ: **Processing of Drosophila endo-siRNAs depends on a specific Loquacious isoform.** *RNA* 2009, **15:**1886–1895.
35. Ghildiyal M, Xu J, Seitz H, Weng Z, Zamore PD: **Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway.** *RNA* 2010, **16:**43–56.
36. Valen E, Preker P, Andersen PR, Zhao X, Chen Y, Ender C, Dueck A, Meister G, Sandelin A, Jensen TH: **Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes.** *Nat Struct Mol Biol* 2011, **18:**1075–1082.
37. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T: **Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.** *Cell* 2010, **141:**129–141.
38. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, Loewer A, Ziebold U, Landthaler M, Kocks C, le Noble F, Rajewsky N: **Circular RNAs are a large class of animal RNAs with regulatory potency.** *Nature* 2013, **495:**333–338.
39. Leung AK, Young AG, Bhutkar A, Zheng GX, Bosson AD, Nielsen CB, Sharp PA: **Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs.** *Nat Struct Mol Biol* 2011, **18:**237–244.
40. Helwak A, Kudla G, Dudnakova T, Tollervey D: **Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding.** *Cell* 2013, **153:**654–665.
41. 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.
42. Yang JH, Li JH, Shao P, Zhou H, Chen YQ, Qu LH: **starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data.** *Nucleic Acids Res* 2011, **39:**D202–D209.
43. Ruggero K, Guffanti A, Sharma VK, De Bellis G, Corti G, Grassi A, Zanovello P, Bronte V, Ciminale V, D'Agostino DM: **Small noncoding RNAs in cells transformed by human T-cell leukemia virus type 1: a role for a tRNA fragment as a primer for reverse transcriptase.** *J Virol* 2014, **88:**3612–3622.
44. Huang V, Li LC: **Demystifying the nuclear function of Argonaute proteins.** *RNA Biol* 2014, **11:**18–24.
45. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A: **NCBI GEO: archive for functional genomics data sets–update.** *Nucleic Acids Res* 2013, **41:**D991–D995.
46. Leinonen R, Sugawara H, Shumway M: **The sequence read archive.** *Nucleic Acids Res* 2011, **39:**D19–D21.
47. Marcel M: **Cutadapt removes adaptor sequences from high-throughput sequencing reads.** *EMBnet. journal* 2011, **17:**10–12.
48. Chan PP, Lowe TM: **GtRNAdb: A database of transfer RNA genes detected in genomic sequence.** *Nucleic Acids Res* 2009, **37:**D93–D97.
49. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389–3402.
50. Xiong Y, Steitz TA: **A story with a good ending: tRNA 3′-end maturation by CCA-adding enzymes.** *Curr Opin Struct Biol* 2006, **16:**12–17.

doi:10.1186/s12915-014-0078-0
Cite this article as: Kumar *et al.*: Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Biology* 2014 **12**:78.