Production of extrachromosomal microDNAs is linked to mismatch repair pathways and transcriptional activity

Laura W. Dillon¹,⁵, Pankaj Kumar¹,⁵, Yoshiyuki Shibata¹,⁵, Yuh-Hwa Wang¹, Smaranda Willcox², Jack D. Griffith², Yves Pommier³, Shunichi Takeda⁴, and Anindya Dutta¹,*

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

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 reveal that homologous recombination and nonhomologous 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

*Correspondence: ad8q@virginia.edu.
¹Co-first author

Publisher’s Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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. and J.D.G performed electron microscopy of microDNA from mouse tissues. 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.

ACCESSION NUMBERS
Sequencing data was submitted to Gene Expression Omnibus under accession number GSE68644.

SUPPLEMENTAL INFORMATION
Supplemental Information includes Supplemental Experimental Procedures and four figures.
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 single-nucleotide polymorphisms 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 exhibit 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 microDNA 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 retro-transposon 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. Due to 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 are generated by replication slippage and the MMR pathway. In summary, the unexpectedly ubiquitous and abundant microDNAs are the products of break repair or mismatch repair following DNA damage associated with transcription or splicing.

RESULTS

Characterization of microDNA across a panel of adult mouse tissues

To determine if 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 to a random model (Figure 1D). Furthermore, the sources of microDNA are highly enriched in genic regions, especially 5’-untranslated regions (5’-UTRs) of genes, exons and CpG islands (Figure 1E). Within genes, microDNA 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.

**MicroDNA 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, 1996-2010). Approximately 40-50% of microDNA 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) (Figure 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 (Figure 2B and S2). Previously, full-length L1 transcripts and L1-encoded proteins have been detected in pre-pubertal 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 (Figure 2C and D). MicroDNAs from other mouse 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 “hot-spots” of microDNA generation can be identified that overlap between all the tissue types (Figure 3B). In agreement with the analyses in Figure 3A, these “hot-spots” correlate with regions of high GC content (Figure 3C) and gene density (Figure 3D).

Due to 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 (Guénol et al., 2008). Furthermore, microDNA 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 if microDNAs are selectively generated from sites that are expressed differentially between the two lineages. Tens to hundreds of thousands of unique microDNA 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 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**

Due to 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 homologous recombination (HR), NBS1 (Tauchi et al., 2002) involved in both HR and NHEJ and MSH3 involved in DNA mismatch repair (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, microDNA 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 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 to WT. After observing this alteration in the genomic location of microDNAs from MSH3−/− cells we examined if 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 to wild type (Figure 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 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 (Figure 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 microDNA are a widespread phenomenon found across numerous vertebrate species, 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 by-product 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., 2013; Ginno et al., 2012; Roy and Lieber, 2009; Skourti-Stathaki et al., 2011). Like R-loops, we often observed microRNA at CpG islands and the 5’- and 3’-ends of genes (Figure 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). Mismatch repair pathways excise these DNA loops (Schofield and Hsieh, 2003), but ligation of the excised product could form a 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 a 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 suggests 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 hot-spots 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 supplemental experimental procedures for further details.

**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 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 Multiple Displacement Amplification 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 was 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, microDNA 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.

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**ACKNOWLEDGEMENTS**

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 Dutta 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.).

REFERENCES

Adachi N, Ishino T, Ishii Y, Takeda S, Koyama H. DNA ligase IV-deficient cells are more resistant to ionizing radiation in the absence of Ku70: Implications for DNA double-strand break repair. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98:12109–12113. [PubMed: 11593023]

Beckmann JS, Estivill X, Antonarakis SE. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nature reviews. Genetics. 2007; 8:639–646.

Bezzubova O, Silbergleit A, Yamaguchi-Iwai Y, Takeda S, Buerstedde JM. Reduced X-ray resistance and homologous recombination frequencies in a RAD54−/− mutant of the chicken DT40 cell line. Cell. 1997; 89:185–193. [PubMed: 9108474]

Branciforte D, Martin SL. Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. Molecular and cellular biology. 1994; 14:2584–2592. [PubMed: 8139560]

Cohen S, Segal D. Extrachromosomal circular DNA in eukaryotes: possible involvement in the plasticity of tandem repeats. Cytogenetic and genome research. 2009; 124:327–338. [PubMed: 19556784]

Flores M, Morales L, Gonzaga-Jauregui C, Dominguez-Vidana R, Zepeda C, Yanez O, Gutierrez M, Lemus T, Valle D, Avila MC, et al. Recurrent DNA inversion rearrangements in the human genome. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104:6099–6106. [PubMed: 17389356]

Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nature reviews. Genetics. 2009; 10:241–251.

Ginno PA, Lott PL, Korf I, Chedin F. GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. Genome research. 2013; 23:1590–1600. [PubMed: 23868195]

Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Molecular cell. 2012; 45:814–825. [PubMed: 22387027]

Guellen L, Page L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008; 453:948–951. [PubMed: 18463634]

Hatanaka A, Yamazoe M, Sale JE, Takata M, Yamamoto K, Kitao H, Sonoda E, Kikuchi K, Yonetani Y, Takeda S. 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. Molecular and cellular biology. 2005; 25:1124–1134. [PubMed: 15657438]

Li X, Manley JL. Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. Cell. 2005; 122:365–378. [PubMed: 16096057]

Lupski JR. New mutations and intellectual function. Nature genetics. 2010; 42:1036–1038. [PubMed: 21102619]

Martin RW, Orelli BJ, Yamazoe M, Minn AJ, Takeda S, Bishop DK. RAD51 up-regulation bypasses BRCA1 function and is a common feature of BRCA1-deficient breast tumors. Cancer research. 2007; 67:9658–9665. [PubMed: 17942895]

Nakamura K, Kogame T, Oshiumi H, Shinohara A, Sumitomo Y, Agama K, Pommier Y, Tsutsui KM, Tsutsui K, Hartsuiker E, et al. Collaborative action of Brca1 and CtIP in elimination of covalent modifications from double-strand breaks to facilitate subsequent break repair. PLoS genetics. 2010; 6:e1000828. [PubMed: 20107609]
Ostertag EM, Kazazian HH Jr. Biology of mammalian L1 retrotransposons. Annual review of genetics. 2001; 35:501–538.

Penzkofer T, Dandekar T, Zemojtel T. L1Base: from functional annotation to prediction of active LINE-1 elements. Nucleic acids research. 2005; 33:D498–500. [PubMed: 15608246]

Roy D, Lieber MR. G clustering is important for the initiation of transcription-induced R-loops in vitro, whereas high G density without clustering is sufficient thereafter. Molecular and cellular biology. 2009; 29:3124–3133. [PubMed: 19307304]

Schofield MJ, Hsieh P. DNA mismatch repair: molecular mechanisms and biological function. Annual review of microbiology. 2003; 57:579–608.

Shibata Y, Kumar P, Layer R, Willcox S, Gagan JR, Griffith JD, Dutta A. Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. Science. 2012; 336:82–86. [PubMed: 22403181]

Skourtif-Stathaki K, Proudfoot NJ. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. Genes & development. 2014; 28:1384–1396. [PubMed: 24990962]

Skourtif-Stathaki K, Proudfoot NJ, Gromak N. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. Molecular cell. 2011; 42:794–805. [PubMed: 21700224]

Smit AH, R, Green P. RepeatMasker Open-3.0. 1996-2010

Sollier J, Stork CT, Garcia-Rubio ML, Paulsen RD, Aguilera A, Cimprich KA. Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. Molecular cell. 2014; 56:777–785. [PubMed: 25435140]

Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annual review of medicine. 2010; 61:437–455.

Takata M, Sasaki MS, Sonoda E, Morrison C, Hashimoto M, Utsumi H, Yamaguchi-Iwai Y, Shinohara A, Takeda S. 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. The EMBO journal. 1998; 17:5497–5508. [PubMed: 9736627]

Tauchi H, Kobayashi J, Morishima K, van Gent DC, Shiraishi T, Verkaik NS, vanHeems D, Ito E, Nakamura A, Sonoda E, et al. Nbs1 is essential for DNA repair by homologous recombination in higher vertebrate cells. Nature. 2002; 420:93–98. [PubMed: 12422221]
**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
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 = 100 nm (B) Length distribution of microDNAs identified in adult mouse tissues. (C) 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 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 to 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.
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 (grey). See also Figure S2.
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 standard deviation. (B) 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 (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 (LAD) 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 standard deviation. *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 calculated.
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 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 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 up- 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 3 independent experiments. Error bars represent standard deviation. *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 3 independent experiments. Results from each individual replicate are given in Figure S4B. See also Figure S3 and S4.
Figure 6. Models for the generation of ss or ds microDNA. Left
Polymerase slippage during DNA replication at succeeding short direct repeats (blue arrow heads) 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 a 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 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.
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 Paids | Read Sequences | Aligned Sequences | Unique Alignment | Unique microDNA (complexity) |
|-------------------------|------------------|---------------|-----------------------|----------------|-------------------|------------------|-------------------------------|
| **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