No evidence of paralogous loci or new *bona fide* microRNAs in telomere to telomere (T2T) genomic data

Arun H. Patil[1], Marc K. Halushka[1*], Bastian Fromm[2]

[1] Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, 21205 USA
[2] The Arctic University Museum of Norway, UiT- The Arctic University of Norway, Tromsø, Norway

* Corresponding author.
Email: mhalush1@jhmi.edu (Halushka MK).

Abstract

The telomere to telomere (T2T) genome project discovered and mapped ~240 million additional base pairs of primarily telomeric and centromeric reads. Much of this sequence was comprised of satellite sequences and large segmental duplications. We evaluated the extent to which human *bona fide* microRNAs (miRNAs) may be found in additional paralogous genomic loci or if previously undescribed microRNAs are present in these newly sequenced regions of the human genome. New genomic regions of the T2T project spanning ~240 million bp of sequence were obtained and evaluated by blastn for the human miRNAs contained in MirGeneDB2.0 (N=556) and miRBase (N = 1917) along with all species of MirGeneDB2.0 miRNAs (N=10,899). Additionally, bowtie was used to compare unmapped reads from >4,000 primary cell samples to the new T2T sequence. Based on sequence and structure, no *bona fide* miRNAs were identified. Ninety-seven miRNAs of questionable authenticity (frequently known repeat elements) were identified from the miRBase dataset across the newly described regions of the human genome. These 97 represent only 51 miRNA families due to paralogy of highly similar miRNAs such as 24 members of the hsa-mir-548 family. Altogether, this data strongly supports our having identified widely expressed *bona fide* miRNAs in the human genome and move us further toward the completion of human miRNA discovery.

Introduction

microRNAs (miRNAs) are a class of small regulatory RNAs that block protein translation. Since their discovery in 1993, a large amount of research has sought to determine the number of, and types of members of this RNA family (Lee et al. 1993; Lagos-Quintana et al. 2001; Halushka et al. 2018). This is a contentious issue in science, with a number of different estimates of miRNA family size depending on competing views of the criteria for designating small RNA sequences of on average 22 bp as miRNAs (Griffiths-Jones 2006; Fromm et al. 2015; Backes et al. 2018; Fromm et al. 2020).

While miRBase used to be a repository for published miRNA candidates, where authors could submit their published miRNA candidates (Griffiths-Jones 2004), MirGeneDB uses a uniform system for microRNA annotation & nomenclature, based on next generation sequencing detectable criteria of miRNA biogenesis to arrive at *bona fide* miRNA complements for metazoan species (Fromm et al. 2015; Fromm et al. 2020; Fromm et al. 2021).

Additionally, due to the ancient origins of miRNAs and genomic rearrangements across species, including segmental duplications, several miRNAs are expressed from multiple genomic loci (Peterson et al. 2021). For example, the miRNA hsa-let-7a is found on chromosomes 9, 11, and 22 (Hsa-Let-7-P2a1, Hsa-Let-7-P1a, Hsa-Let-7-P1D), with the -5p (mature) sequence being identical across all three loci. Other miRNAs have undergone minimal sequence changes through these duplication events and now exist as families such as the MIR-17 family, from which hsa-miR-18a (Hsa-Mir-17-P2a) on chromosome 13 and hsa-miR-18b (Hsa-Mir-17-P2c) on the X chromosome differ only by the 20th base of their -5p (dominant) sequence.

A significant amount of our understanding of miRNA expression patterns comes from small RNA-sequencing datasets that capture miRNAs, tRNA halves and fragments, rRNA fragments and other RNA species (de Rie et al. 2017; McCall et al. 2017; Lorenzi et al. 2021; Patil

and Halushka 2021). A common curiosity of small RNA-seq data is a significant number of unmapped reads in these datasets. A number of rationales have been ascribed to this material. Some explanations for this extra data include poor sequence quality reads, contamination, infectious organisms, and repeat elements. Another explanation could be that the sequence aligns to the unmapped and uncharacterized regions of the human genome.

Despite the declaration of completing the human genome in 2001, we have always known that large segments of pericentromeric and peritelomeric regions remained unmapped (Lander et al. 2001). Recently, the telomere to telomere (T2T) consortia has used long read sequencing and other methods to fully map a complete human genome (Nurk et al. 2021). This effort increased the amount of sequenced and aligned genome by ~8%. Most of this sequence is repetitive pericentromeric satellite structures (Altemose et al. 2021; Vollger et al. 2021). However, some segmental duplications and other gene-containing genomic structures exist in this area, which may harbor miRNA paralogs or even heretofore unmapped *bona fide* miRNAs (Nurk et al. 2021; Vollger et al. 2021).

We wondered if this newly characterized region of the human genome contained paralogous loci of known miRNAs and whether any yet described *bona fide* miRNAs may exist in these areas. To do this we surveyed ~240 million bp of T2T sequence using both known miRNAs and unmapped sequence from 4,150 mostly primary cell type datasets.

Methods

Data set data acquisition

The complete T2T genome was obtained from https://github.com/marbl/CHM13 (v1.0) retrieved 7/30/2021) (Nurk et al. 2021). It is also located at https://www.ncbi.nlm.nih.gov/assembly/GCA_009914755.2. Dr. Mitchell Vollger kindly provided a

list of non-syntenic regions of T2T relative to GRCh38, which were selected from the full T2T genome using getfasta (BEDTools)(Quinlan 2014). This 240 million bp dataset was subsequently used for all searching (hereafter, the T2T genome). A list of both homo sapiens and "All species" miRNA precursor sequences were obtained from MirGeneDB (https://mirgenedb.org/download, accessed 7/30/21) (Fromm et al. 2020). Separately, the hairpin.fa file from miRBase.org was obtained (https://www.mirbase.org/ftp.shtml accessed 7/30/21) and the homo sapiens (hsa) miRNAs were subsetted out (Griffiths-Jones 2006).

Unaligned small RNA-seq reads

From an ongoing project to characterize the cellular microRNAome, 4,150 primary cell, plasma, exosome, and cancer cell line small RNA-seq files were obtained from the Sequence Read Archive (SRA) and processed through miRge3.0 using parameters (miRge3.0 -s file_names -gff -bam -trf -lib /miRge3_Lib/ -on human -db mirbase -o <output folder> -mEC -ks 20 -ke 20 -a variable_adapters) (Patil and Halushka 2021). All unmapped reads from the 4,150 files were concatenated and collapsed into a list of unique reads with the read count denoted using a python (3.8.5) script.

Alignment to the non-syntenic regions of T2T

All reported miRNAs from MirGeneDB and miRBase were used in a local blastn (version: 2.6.0+) search with the non-syntenic T2T sequence. Parameters were: "blastn -query mirgenedb_allSps.fasta -db ../blastthis/jhu_chm13_v1.0_T2T.fasta -html > blast_output_preAllSps_miRGeneDB.html". Unmapped reads from small RNA-seq datasets of length 16-27 bp were aligned to the non-syntenic regions of T2T using Bowtie (v1.2.3) and command line: "bowtie <T2T index> -f all_unmappedReads.fasta -S <Sam Output> -p <threads> -n 0" (Langmead 2010).

Results

Search for paralogous miRNA loci in T2T regions

There were 126 separate non-syntenic T2T segments with a total read length of 240,044,315 bp. This was the search space for all further miRNA analyses. These segments represented ~8% new sequence compared to GRCh38. Both the homo sapiens (N=556) and all species (N=10,899) searches of MirGeneDB (*bona fide*) miRNA precursors yielded zero blastn alignments to the T2T data indicating no previously undiscovered paralogous miRNA loci.

A search of human miRBase miRNA hairpins (N=1917) identified 97 miRNA hairpins with alignments to the T2T data (Table 1). In total, 1794 separate alignments with a percent identity between 80.5 and 100% were detected. There were 445 alignments with a perfect match across a range of 27-179 bases, with roughly half (226), having a perfect alignment of only 27-30 bases. The average length of the hairpin sequence was 82 bp. The dominant miRNA found in this collection was the hsa-miR-548 class. This putative miRNA shares high sequence homology with the Made1 (Tc1/Mariner) repeat family (Piriyapongsa and Jordan 2007). miRbase lists 73 members of the miR-548 family. Twenty-four of these were listed among the miRNAs with alignment in the T2T sequences. Other miRNAs listed repeatedly, such as hsa-miR-1302 (MER53 element), hsa-miR-3118 (Line 1 element), and hsa-miR-5701 (Rep522 repeat), were generally overlapping known repeat elements. The miRNA hairpin with the highest sequence identity was hsa-miR-3648-1, a putative miRNA with a high G/C content (81%), which matched identically (180 bp) in four T2T locations: Two locations in Chr. 13 and one each on Chr. 14 and 21.

Table 1: miRBase putative miRNAs aligned to T2T sequence		
hsa-mir-10396a	hsa-mir-3680-2	hsa-mir-548h-3
hsa-mir-10396b	hsa-mir-3908	hsa-mir-548h-4
hsa-mir-10401	hsa-mir-3929	hsa-mir-548i-1
hsa-mir-1233-1	hsa-mir-4252	hsa-mir-548i-2

hsa-mir-1233-2	hsa-mir-4258	hsa-mir-548i-3
hsa-mir-1255b-2	hsa-mir-4267	hsa-mir-548i-4
hsa-mir-1267	hsa-mir-4273	hsa-mir-548o-2
hsa-mir-1268a	hsa-mir-4436a	hsa-mir-548t
hsa-mir-1273h	hsa-mir-4436b-1	hsa-mir-548u
hsa-mir-1285-1	hsa-mir-4436b-2	hsa-mir-548w
hsa-mir-1299	hsa-mir-4448	hsa-mir-548z
hsa-mir-1302-10*	hsa-mir-4456	hsa-mir-570
hsa-mir-1302-11*	hsa-mir-4472-2	hsa-mir-5701-1*
hsa-mir-1302-2*	hsa-mir-4477a	hsa-mir-5701-2*
hsa-mir-1302-3*	hsa-mir-4477b	hsa-mir-5701-3*
hsa-mir-1302-9*	hsa-mir-4502	hsa-mir-5708
hsa-mir-1303	hsa-mir-4509-1	hsa-mir-619
hsa-mir-1324	hsa-mir-4509-2	hsa-mir-663a
hsa-mir-1972-1	hsa-mir-4509-3	hsa-mir-663b
hsa-mir-1972-2	hsa-mir-548aa-1	hsa-mir-6724-1
hsa-mir-3118-1*	hsa-mir-548aa-2	hsa-mir-6724-2
hsa-mir-3118-2*	hsa-mir-548ad	hsa-mir-6724-3
hsa-mir-3118-3*	hsa-mir-548ae-2	hsa-mir-6724-4
hsa-mir-3118-4*	hsa-mir-548ai	hsa-mir-6829
hsa-mir-3135a	hsa-mir-548aj-2	hsa-mir-6859-1
hsa-mir-3135b	hsa-mir-548am	hsa-mir-6859-2
hsa-mir-3159	hsa-mir-548ar	hsa-mir-6859-3
hsa-mir-3648-1	hsa-mir-548ay	hsa-mir-6859-4
hsa-mir-3648-2	hsa-mir-548ba	hsa-mir-7705
hsa-mir-3674	hsa-mir-548c	hsa-mir-7851
hsa-mir-3675	hsa-mir-548d-1	hsa-mir-8078
hsa-mir-3680-1	hsa-mir-548d-2	hsa-mir-8086
		hsa-mir-9901
* Indicates known overlap of a putative miRNA to a repeat element. No MirGeneDB *bona fide* miRNAs are in this table.		

Novel miRNA discovery

From 4,150 small RNA-seq datasets of primarily primary cell sequencing that had been processed through miRge3.0, 2,872,614,004 unmapped reads were collapsed into a single search space. Using Bowtie alignment without mismatches, 296,475 unique 16-27bp length reads aligned, representing 383,109,405 total reads mapping to the T2T sequences. These

sequences were reviewed for the likelihood they may represent sequences from a miRNA hairpin loop (multiple sequences aligning with the same 5p edge (±1bp) separated from at least one additional sequence by 10-25 bp loop), yielding 8 potential hits. RNA folding of these 8 sequences did not generate a viable stem loop structure. Thus, no novel miRNAs were described.

Conclusions

A thorough investigation of 240 million bases of new T2T sequences did not identify any new *bona fide* miRNAs or paralogous genomic regions containing known *bona fide* miRNAs. 97 miRNAs of questionable provenance from miRBase were aligning to the T2T sequence, but these most likely represent repeat elements that have been misassigned as miRNAs.

Although every explorer looks for new worlds to conquer, there is some satisfaction in knowing we are nearing the end of the discovery of widely expressed human *bona fide* miRNAs. While it is possible that rare human cell types may harbor miRNAs yet found in RNA-seq, there is no reason to think they will be found in pericentromeric or peritelomeric regions. One last area of potential discovery would be larger genomic regions variable between ethnic populations. It is possible that groups with lower requirements for miRNA discovery may claim findings in the new T2T space, but one should be cautious, if those reports appear.

Acknowledgements

The authors thank Mitchell Vollger and Evan Eichler for their assistance and helpful comments.

Funding

A.H.P. and M.K.H. were supported by grants R01HL137811 and R01GM130564. B.F. is supported by the Tromsø forskningsstiftelse (TFS) [20_SG_BF ‘MIRevolution’].

References

Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ, Uralsky L, Ryabov FD, Shew CJ et al. 2021. Complete genomic and epigenetic maps of human centromeres. *bioRxiv* doi:10.1101/2021.07.12.452052: 2021.2007.2012.452052.

Backes C, Fehlmann T, Kern F, Kehl T, Lenhof HP, Meese E, Keller A. 2018. miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res* **46**: D160-D167.

de Rie D, Abugessaisa I, Alam T, Arner E, Arner P, Ashoor H, Astrom G, Babina M, Bertin N, Burroughs AM et al. 2017. An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nature biotechnology* **35**: 872-878.

Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, Newcomb JM, Sempere LF, Flatmark K, Hovig E et al. 2015. A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annual review of genetics* **49**: 213-242.

Fromm B, Domanska D, Hoye E, Ovchinnikov V, Kang W, Aparicio-Puerta E, Johansen M, Flatmark K, Mathelier A, Hovig E et al. 2020. MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res* **48**: D132-D141.

Fromm B, Hoye E, Domanska D, Zhong X, Aparicio-Puerta E, Ovchinnikov V, Umu SU, Chabot PJ, Kang W, Aslanzadeh M et al. 2021. MirGeneDB 2.1: toward a complete sampling of all major animal phyla. *Nucleic Acids Res* doi:10.1093/nar/gkab1101.

Griffiths-Jones S. 2004. The microRNA Registry. *Nucleic Acids Res* **32**: D109-111.

Griffiths-Jones S. 2006. miRBase: the microRNA sequence database. *Methods Mol Biol* **342**: 129-138.

Halushka MK, Fromm B, Peterson KJ, McCall MN. 2018. Big Strides in Cellular MicroRNA Expression. *Trends in genetics : TIG* **34**: 165-167.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294**: 853-858.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Langmead B. 2010. Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics* **Chapter 11**: Unit 11 17.

Lee RC, Feinbaum RL, Ambros V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**: 843-854.

Lorenzi L, Chiu HS, Avila Cobos F, Gross S, Volders PJ, Cannoodt R, Nuytens J, Vanderheyden K, Anckaert J, Lefever S et al. 2021. The RNA Atlas expands the catalog of human non-coding RNAs. *Nature biotechnology* **39**: 1453-1465.

McCall MN, Kim MS, Adil M, Patil AH, Lu Y, Mitchell CJ, Leal-Rojas P, Xu J, Kumar M, Dawson VL et al. 2017. Toward the human cellular microRNAome. *Genome Res* **27**: 1769-1781.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A et al. 2021. The complete sequence of a human genome. *bioRxiv* doi:10.1101/2021.05.26.445798: 2021.2005.2026.445798.

Patil AH, Halushka MK. 2021. miRge3.0: a comprehensive microRNA and tRF sequencing analysis pipeline. *NAR Genom Bioinform* **3**: lqab068.

Peterson KJ, Beavan A, Chabot PJ, McPeek MA, Pisani D, Fromm B, Simakov O. 2021. microRNAs as Indicators into the Causes and Consequences of Whole Genome Duplication Events. *Mol Biol Evol* **Online ahead of print,**.

Piriyapongsa J, Jordan IK. 2007. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS One* **2**: e203.

Quinlan AR. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current protocols in bioinformatics* **47**: 11 12 11-34.

Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, Diekhans M, Sulovari A, Munson KM, Lewis AM et al. 2021. Segmental duplications and their variation in a complete human genome. *bioRxiv* doi:10.1101/2021.05.26.445678: 2021.2005.2026.445678.