Mitochondrion xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Mitochondrion

journal homepage: www.elsevier.com/locate/mito

Two novel lncRNAs discovered in human mitochondrial DNA using PacBio full-length transcriptome data

Shan Gao[a,b,1], Xiaoxuan Tian[c,1], Hong Chang[a], Yu Sun[a], Zhenfeng Wu[d], Zhi Cheng[a], Pengzhi Dong[c], Qiang Zhao[a], Jishou Ruan[d,*], Wenjun Bu[a,*]

[a] College of Life Sciences, Nankai University, Tianjin 300071, PR China
[b] Institute of Statistics, Nankai University, Tianjin 300071, PR China
[c] Tianjin State Key Laboratory of Modern Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin 300193, PR China
[d] School of Mathematical Sciences, Nankai University, Tianjin 300071, PR China

ARTICLE INFO

Keywords:
Mitochondrial transcriptome
Full-length transcriptome
lncRNA
psRNA
ltiRNA

ABSTRACT

In this study, we established a general framework to use PacBio full-length transcriptome sequencing for the investigation of mitochondrial RNAs. As a result, we produced the first full-length human mitochondrial transcriptome using public PacBio data and characterized the human mitochondrial genome with more comprehensive and accurate information. Other results included determination of the H-strand primary transcript, identification of the ND5/ND6AS/tRNA^{Glu}AS transcript, discovery of palindrome small RNAs (psRNAs) and construction of the “mitochondrial cleavage” model, *etc*. These results reported for the first time in this study fundamentally changed annotations of human mitochondrial genome and enriched knowledge in the field of animal mitochondrial studies. The most important finding was two novel long non-coding RNAs (lncRNAs) of MDL1 and MDL1AS exist ubiquitously in animal mitochondrial genomes.

1. Introduction

Animal mitochondrial DNA (mtDNA) is a small, circular and extrachromosomal genome, typically about 16 Kbp in length (16,569 bp for human) and contains almost the same 37 genes: two for rRNAs, 13 for mRNAs and 22 for tRNAs (Boore, 1999), although a few other genes (*e.g.* ATP9) have been discovered. Both replication and transcription of human mtDNA are initiated from a noncoding region named Displacement-loop (D-loop). Mitochondrial RNAs are transcribed as primary transcripts, then processed into polycistronic precursors and finally mature RNAs by enzyme cleavage and specific nucleotide modification. These mature RNAs are pivotal for Adenosine Triphosphate (ATP) production, programming of cell death and other cell functions (Shiota et al., 2015). However, the mechanisms of mitochondrial gene transcription and its regulation are still not well understood (Stewart & Beckenbach, 2009).

In 1981, the complete human mtDNA sequence was determined and characterized by it's 37 encoded genes (Anderson et al., 1981). With Next Generation Sequencing (NGS) technologies, Mercer et al. tried to demonstrate wide variation in mitochondrial transcript abundance and resolve RNA processing and maturation events (Mercer et al., 2011). However, the NGS short reads resulted in incompletely assembled mitochondrial transcripts which limited the use of transcriptome data. Based on the Single-molecule Real-time (SMRT) sequencing technology, the PacBio platform can provide longer and even full-length transcripts that originate from observations of single molecules without assembly. Gao et al. constructed the first quantitative transcription map of animal mitochondrial genomes (Gao et al., 2016) by sequencing the full-length insect (*Erthesina fullo* Thunberg) transcriptome (Ren et al., 2016) on the PacBio platform and established a straightforward and concise methodology to investigate mitochondrial gene transcription and RNA processing. Most of results in the study of the full-length insect mitochondrial transcriptome were consistent with those in previous studies, while new findings included an unexpectedly high level of mitochondrial gene transcription, a high-content of polycistronic transcripts, genome-wide antisense transcripts and a new model to describe

Abbreviations: tiRNA, transcription initiation RNA; ltiRNA, long tiRNA; lncRNA, long non-coding RNA; mtDNA, mitochondrial DNA; D-loop, Displacement-loop; ATP, Adenosine Triphosphate; NGS, Next Generation Sequencing; SMRT, Single-molecule Real-time; CCS, Circular Consensus Sequencing; H-strand, Heavy strand; L-strand, Light strand; O_H, Origin of H-strand replication; O_L, Origin of L-strand replication; IT_{H1}, the first Transcription Initiation Site on H-strand; IT_{H2}, the second Transcription Initiation Site on H-strand; IT_L, the Transcription Initiation Site on L-strand; SNP, Single Nucleotide Polymorphism; eft, *Erthesina fullo* Thunberg; HCC, Hepatocellular Carcinoma; TIS, Transcription Initiation Site

[*] Corresponding authors.
E-mail addresses: jsruan@nankai.edu.cn (J. Ruan), wenjunbu@nankai.edu.cn (W. Bu).

[1] These authors contributed equally to this paper.

http://dx.doi.org/10.1016/j.mito.2017.08.002
Received 22 April 2017; Received in revised form 11 June 2017; Accepted 7 August 2017
1567-7249/ © 2017 Published by Elsevier B.V.

mitochondrial RNA processing, *etc*. Using the full-length insect mitochondrial transcriptome data, Gao et al. have proven that the high-content polycistronic transcripts are mRNA or rRNA precursors and the analysis of these precursors facilitates the investigation of mitochondrial gene transcription, RNA processing and other relevant topics.

In this study, we used a public PacBio full-length transcriptome dataset to produce the full-length human mitochondrial transcriptome. By the identification and further analysis of full-length transcripts, we were able to characterize the human mitochondrial genome with more comprehensive and accurate information. Our research objectives included: 1) to establish a general framework for investigating mitochondrial gene transcription using PacBio full-length transcriptome sequencing; 2) to provide accurate annotation of the human mitochondrial reference genome for future studies; 3) to study biological similarities and differences between the human and insect quantitative transcription map constructed using their full-length transcriptome data; 4) to reveal molecular mechanisms underlying some fundamental problems in mitochondrial biology, *e.g.* mtDNA transcription, RNA processing and regulation of mtDNA transcription.

2. Results and discussion

2.1. The quantitative transcription map of human mitochondrial genome

The full-length transcriptome dataset of the human MCF-7 cell line was chosen to construct the first quantitative transcription map of human mitochondrial genome due to its highest data quality (Materials and methods). This dataset had been acquired by sequencing five groups of cDNA libraries with sizes of 0–1 Kbp, 1–2 Kbp, 2–3 Kbp, 3–5 Kbp and 5–7 Kbp (Table 1). The raw data contained 1,984,154 raw reads with the average size of 16,262 bp. After removing low quality regions and SMRTbell adapters, the raw reads were split into 9,192,530 high-quality (Accuracy ≥ 75%) subreads with the average size of 2668.85 bp. The subreads were processed into Circular Consensus Sequencing (CCS) reads with improved data quality. Finally, the CCS reads were used to produce 744,469 draft transcripts with sequence redundancy. Using adjusted parameters, at least 3.07% (22,853/744,469) of the draft transcripts could be continuously mapped to the reconstructed human mitochondrial reference genome (Supplementary file 1) to produce the full-length human mitochondrial transcriptome. Since the average length of mitochondrial transcripts is about 1–2 Kbp, the mapping rate was estimated to be 7.2% (21,689/301,149) using data from only nine of 28 libraries with sizes of 0–1 Kbp or 1–2 Kbp. This mapping rate was still significantly lower than the mapping rate 37.65% of the insect (*Erthesina fullo* Thunberg) mitochondrial transcripts (Gao et al., 2016). One cause of the difference between these mapping rates could be from the tissue specificity of materials for sequencing. The insect and human total RNA had been extracted from insect whole bodies and human cancer cells, respectively. Other causes could be from experiment processing (Materials and methods).

The quantitative transcription map of insect mitochondrial genome (Fig. 1A) has indicated that eight mRNA transcripts (ND2, COI, COII, ATP8/ATP6, COIII, ND3, ND6 and Cytb) are encoded by the Heavy strand (H-strand), also known as the major coding strand J(+) for insects (Gao et al., 2016), while three mRNA transcripts (ND4L/ND4, ND5 and ND1) and two rRNA transcripts (16S rRNA and 12S rRNA) are encoded by the Light strand (L-strand), also known as the minor coding strand N(−) for insects. From Fig. 1B, it can be seen that all the abundant transcripts (in red color) of human mtDNA are encoded by H-strand, while their antisense transcripts (in blue color) at comparatively low levels are encoded by L-strand. It can also be seen that although the ND6 transcript encoded by L-strand is responsible for protein coding, the ND5 transcript and the ND6 antisense (ND6AS) transcript encoded by H-strand is still more abundant than the ND5 antisense (ND5AS) transcript and the ND6 transcript, respectively. This finding was different from those in all other existing human mitochondrial gene

Table 1
Full-length transcriptome data of human mitochondrial genome.

Cell ID	Library size	CCS/Draft	mtDNA	mtDNA (%)
m140731_222056_42161_c100698070630000001823143403261500	0–1 Kbp	32,886	3072	9.34%
m140801_014337_42161_c100698070630000001823143403261501	0–1 Kbp	31,994	2849	8.90%
m140801_050734_42161_c100698070630000001823143403261502	0–1 Kbp	35,335	3757	10.63%
m140801_082952_42161_c100698070630000001823143403261503	0–1 Kbp	33,350	3425	10.27%
m140801_115238_42161_c100698070630000001823143403261504	1–2 Kbp	42,673	2425	5.68%
m140801_151509_42161_c100698070630000001823143403261505	1–2 Kbp	42,070	2281	5.42%
m140804_215651_42141_c100700040630000001823139203261500	1–2 Kbp	36,564	2028	5.55%
m140805_011552_42141_c100700040630000001823139203261501	1–2 Kbp	24,136	968	4.01%
m140805_044156_42141_c100700040630000001823139203261502	1–2 Kbp	22,141	884	3.99%
m140805_080756_42141_c100700040630000001823139203261503	2–3 Kbp	15,294	107	0.70%
m140805_113513_42141_c100700040630000001823139203261504	2–3 Kbp	16,208	116	0.72%
m140805_150259_42141_c100700040630000001823139203261505	2–3 Kbp	38,075	330	0.87%
m140808_221025_42161_c100696951270000001823138003261560	2–3 Kbp	34,586	276	0.80%
m140809_012938_42161_c100696951270000001823138003261561	2–3 Kbp	38,528	271	0.70%
m140809_044851_42161_c100696951270000001823138003261562	3–5 Kbp	26,329	3	0.01%
m140809_080804_42161_c100696951270000001823138003261563	3–5 Kbp	35,326	3	0.01%
m140809_112717_42161_c100696951270000001823138003261564	3–5 Kbp	33,676	4	0.01%
m140809_144702_42161_c100696951270000001823138003261565	3–5 Kbp	32,952	3	0.01%
m140809_181052_42161_c100696951270000001823138003261566	3–5 Kbp	34,567	6	0.02%
m140809_213120_42161_c100696870310000001823138003261570	3–5 Kbp	17,352	1	0.01%
m140810_025821_42161_c100696870310000001823138003261571	3–5 Kbp	33,556	2	0.01%
m140810_061610_42161_c100696870310000001823138003261572	5–7 Kbp	14,345	6	0.04%
m140810_093523_42161_c100696870310000001823138003261573	5–7 Kbp	15,631	7	0.04%
m140810_125744_42161_c100696870310000001823138003261574	5–7 Kbp	15,875	4	0.03%
m140810_161642_42161_c100693060150000001823146703241590	5–7 Kbp	11,850	9	0.08%
m140810_193644_42161_c100693060150000001823146703241591	5–7 Kbp	12,570	6	0.05%
m140810_225944_42161_c100693060150000001823146703241592	5–7 Kbp	6523	4	0.06%
m140811_021934_42161_c100693060150000001823146703241593	5–7 Kbp	10,077	6	0.06%
Sum of all CCS/Draft		744,469	22,853	3.07%
Sum of 0–2 Kbp CCS/Draft		301,149	21,689	7.20%

CCS/Draft represents the number of CCS reads and draft transcripts. mtDNA represents the number of draft transcripts aligned to the human mitochondrial genome (RefSeq: NC_012920.1).

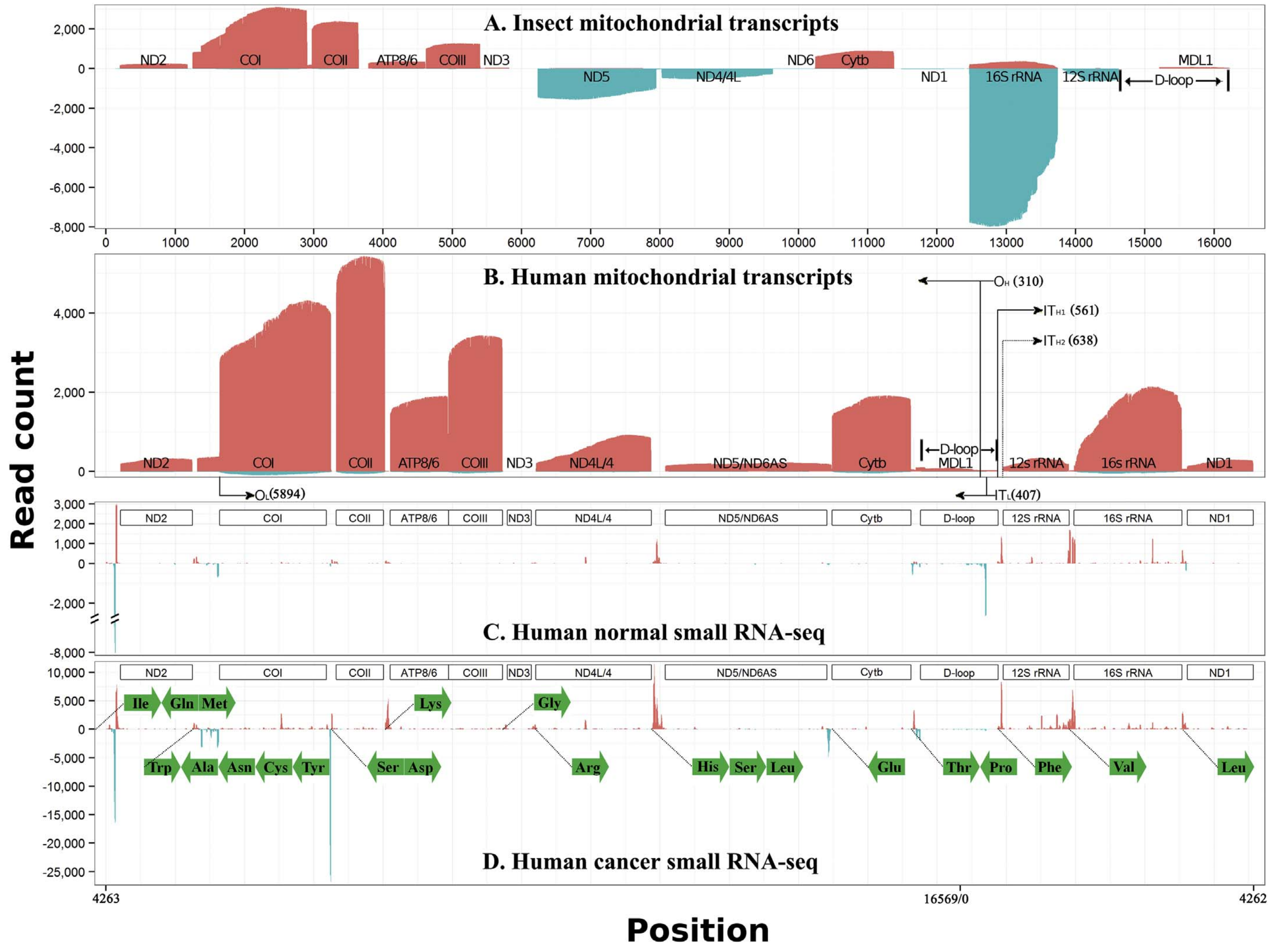

Fig. 1. The quantitative transcription map of human mitochondrial genome.

A. The quantitative transcription map of insect mitochondrial genome is from our previous study (Gao et al., 2016). Forward alignments of transcripts (in red color) are piled along the positive y-axis. Reverse alignments of transcripts (in blue color) are piled along the negative y-axis. B. The reconstruction of the human mitochondrial genome included two steps. The "N" nucleotide was removed from the complete human mitochondrial genome (RefSeq: NC_012920.1) and the resulted sequence shifted 4263 bp counter-clockwise (Supplementary file 1). C. Small RNA-seq data of the normal liver tissue (SRA: SRR039612) was aligned to the reconstructed human mitochondrial genome. D. Small RNA-seq data of the Hepatocellular Carcinoma (HCC) tissue (SRA: SRR039625) was aligned to the reconstructed human mitochondrial genome. The tRNA genes (in green color) are represented using their amino acids. The x-axis in figure D represents positions on the mitochondrial genome (RefSeq: NC_012920.1). The x-axes in figure A represents positions on the reconstructed human mitochondrial genome. O_H, O_L, IT_{H1}, IT_{H2} and IT_L are marked with their positions on the mitochondrial genome (RefSeq: NC_012920.1). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

expression profiles described in previous studies, in which the ND6 transcript had been identified as more abundant (Anderson et al., 1981; Rackham et al., 2011). Using NGS data, the ND5AS and the ND6 transcript were identified as more abundant in the year of 2011 (Mercer et al., 2011). The human mitochondrial transcripts sorted by transcriptional levels from the highest to the lowest were COII, COI, COIII, 16S rRNA, Cytb, ATP8/ATP6, ND4L/ND4, 12S rRNA, ND2, ND1, ND5/ND6AS and ND3. Although this expression profile had differences from that of the insect mitochondrial genes due to the tissue specificity or other experimental reasons, the expression profiles of animal mitochondrial mRNA and rRNA genes were still conservative and their transcripts can be roughly classified into high-, medium- and low-level groups according to their transcriptional levels (Fig. 1AB). In addition, we found that eight high- and medium-level mRNA genes (COII, COI, COIII, Cytb, ATP8, ATP6, ND4L and ND4) in human mitochondrial genome had the preference to use "ATG" as their start codons, while low-level mRNA genes (ND2, ND1, ND5 and ND3) had preferences to "ATT" or "ATA". Since "ATG" has higher efficiency than "ATT" or "ATA" to initial protein synthesis, this suggested that the expression of human mitochondrial genes could be regulated coordinately at both of transcriptional and translational levels to ensure the high efficiency.

2.2. Correction of human mtDNA annotation

The human mitochondrial reference genome (RefSeq: NC_012920.1) was annotated using full-length mature mRNA and rRNA transcripts (Table 2). These full-length transcripts were confirmed by at least 10 full-path CCS reads for their fidelity (Supplementary file 2). In general, the identified mature RNAs in this study were longer than their corresponding mitochondrial genes annotated by their Coding Sequences (CDSs) (Anderson et al., 1981). COI, COII, ATP8/6 and ND1 were assigned new annotations, while 12S rRNA was annotated by its two 5′ ends of 648 and 649 on the reference genome (Table 2). We also found that the mature transcripts of 12S rRNA were cleaved more frequently at the position 649 than the position 648. In addition, the cleavage at the position 648 or 649 produced two types of mature $tRNA^{Phe}$, which ended with nudeotides CA or CAA, respectively. Based on the classical theory, CCA triplets are added to 3′ ends of tRNAs by enzyme after their

Table 2
Annotation of the human mitochondrial genome with corrections.

Transcript	Strand	Start	End	Startnew	Endnew	Length
ND2	H(+)	4470	5511	4470	5511	1042
COI	H(+)	5904	7445	5901[a]	7442[a]	1542
COII	H(+)	7586	8269	7586	8294[a]	709
ATP8/6	H(+)	8366	9207	8365[a]	9206[a]	842
COIII	H(+)	9207	9990	9207	9990	784
ND3	H(+)	10,059	10,404	10,059	10,404	346
ND4L/4	H(+)	10,470	12,137	10,470	12,137	1668
ND5/6AS	H(+)	12,337	14,148	12,337	14,742[a]	2410
Cytb	H(+)	14,747	15,887	14,747	15,887	1141
12S rRNA	H(+)	648	1601	648/649[a]	1601	954/955[a]
16S rRNA	H(+)	1671	3229	1671	3229	1559
ND1	H(+)	3307	4262	3305[a]	4262	958
MDL1	H(+)	–	–	15,954	576	1192
MDL1AS	L(−)	–	–	16,024	407	953
D-loop	–	16,024	576	16,024	576	1122
O$_L$	–	–	–	5894	–	–
O$_H$	–	–	–	310	–	–
IT$_L$	–	–	–	407	–	–
IT$_{H1}$	–	–	–	561	–	–

H(+) and L(−) represents the Heavy and Light strand of the mitochondrial genome, respectively. I$_{H2}$ is still not determined.

[a] Represents the corrected annotation using the full-length transcriptome data. 12S rRNA is annotated by its two 5′ ends of 648 and 649 on the reference genome. We also found that the mature transcripts of 12S rRNA were cleaved more frequently at the position 649 than the position 648.

cleavage. Our finding suggested that tRNAs could contain CCA without adding by enzyme after their cleavage.

In this study, we found a great number of ND5/ND6AS/tRNA^{Glu}AS (NC_012920: 12337-14746) transcripts but did not find any full-length ND5 or ND6AS transcript (Fig. 2A). This suggested that ND5/ND6AS/tRNA^{Glu}AS could be mature RNAs without further cleavage, as ATP8/ATP6 and ND4L/ND4. Then, we used two other larger PacBio datasets (Materials and methods) and strand-specific qPCR to validate this hypothesis. The results supported that ND5/ND6AS/tRNA^{Glu}AS was unlikely to be further cleaved into the mature ND5 and ND6AS transcript. One previous study reported the detection of ND6AS (NC_012920: 13993-14673) as lncRNA using Rapid Amplification of cDNA Ends (RACE) (Rackham et al., 2011) (Fig. 2A). But RACE is unable to differentiate full-length mature RNAs from their partial segments which could be produced by RNA degradation or random breaks during experimental processing. The PacBio full-length transcriptome sequencing produces hundreds to even thousands of sequences to support each full-length mature RNA for its identification. Therefore, our finding had higher confidence than that using RACE. From ND6AS identified in this study, two peptides were predicted as MPPSNLNYNMYTNKQCSTSNYY and MMMQSPRTNRILPNQPWPLSFMNYSASYTMKVYHNHHPIMLFHPQHQSYLHR.

The corrected genome annotation also included some important gonomic features that were the Origin of H-strand replication (O$_H$), the Origin of L-strand replication (O$_L$), the first Transcription Initiation Site (TIS) on H-strand (IT$_{H1}$), the second TIS on H-strand (IT$_{H2}$) and the TIS on L-strand (IT$_L$). In this study, O$_L$ (NC_012920: 5894), O$_H$ (NC_012920: 310), IT$_L$ (NC_012920: 407) and IT$_{H1}$ (NC_012920: 561) were determined, which validated the predictions from previous studies, but IT$_{H2}$ (NC_012920: 638) was not determined. Using full-length transcripts, IT$_L$ and IT$_{H1}$ were identified and confirmed starting at the position 407 (D-loop) and 561 (D-loop) on the human mitochondrial genome, respectively (Montoya & Attardi, 1982). A 15-nt promoter element CAAACCCCAAAGACA (NC_012920: 553-567) surrounding IT$_{H1}$ (underlined) was also consistent with the finding in the previous study (Chang & Clayton, 1984). However, we did not find any full-length transcript to validate H-strand transcription starting at the position 638 in the tRNA^{Phe} region, which had been predicted as IT$_{H2}$ in the previous study. According to the dual H-strand transcription model, transcription starts relatively highly frequent at IT$_{H1}$ and lowly frequent at IT$_{H2}$ for the synthesis of two rRNA transcripts and the entire H-strand, respectively (Montoya & Attardi, 1982). Using two other larger PacBio datasets, we still did not find any sequences to support the existence of IT$_{H2}$ and the dual H-strand transcription model.

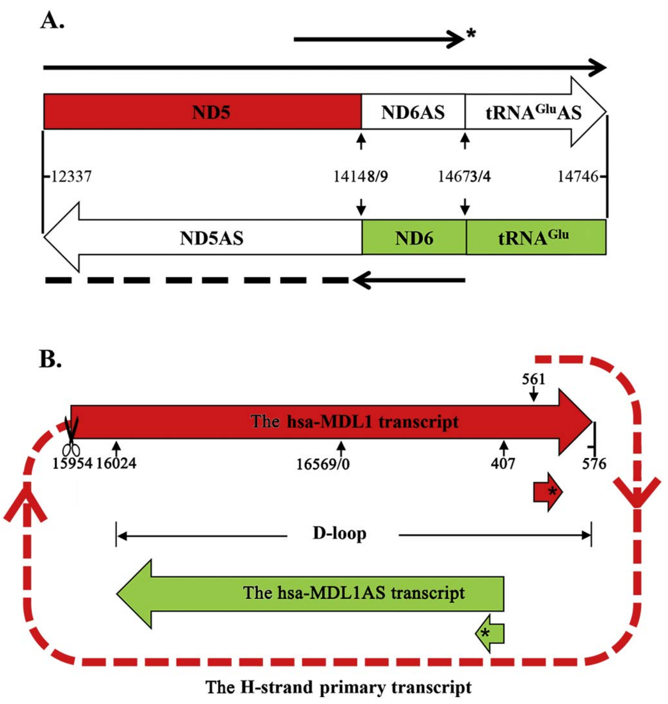

Fig. 2. Transcription of ND5, ND6, MDL1 and MDL1AS.
A. The annotated ND5, ND6, ND5AS and ND6AS genes are marked by their start and end positions on the human mitochondrial genome (RefSeq: NC_012920.1). The detected ND5/ND6AS/tRNAGlu AS transcript in this study is represented by an arrow with a solid line. The 3′ end of mature ND6 transcript is still not determined. *Represents the detected ND6AS transcript (NC_012920: 13993-14673) in the previous study. B. The reconstruction of the human mitochondrial genome included two steps. The "N" nucleotide was removed from the complete human mitochondrial genome (RefSeq: NC_012920.1) and the resulted sequence shifted 4263 bp counter-clockwise (Supplementary file 1). The H-strand primary transcript was precisely determined to start at the position 561 and end at the position 576. The mature hsa-MDL1 (NC_012920: 15956-576) and hsa-MDL1AS (NC_012920: 16024-407) were identified as two novel lncRNAs.*tiRNAs were aligned to Transcription Initiation Sites (TISs).

2.3. Two novel lncRNAs from the D-loop region

In the previous study of the insect mitochondrial transcriptome, we discovered some transcripts aligned forwardly to the insect D-loop region. Further analysis indicated that these transcripts were encoded by a novel gene Mitochondrial D-loop 1 (MDL1) (Fig. 1A). In this study, we discovered some polycistronic transcripts (Supplementary file 2) aligned forwardly to the human D-loop region. This novel gene covered the tRNA^{Pro} antisense gene (NC_012920: 15954-16023) and 100% of the human D-loop region (NC_012920: 16024-576). Therefore, we named the gene from the insect D-loop region and the gene from the human D-loop region as eft-MDL1 and hsa-MDL1 (NC_012920: 15956-576), respectively. The analysis of all the RNA precursors of hsa-MDL1 supported that the mature hsa-MDL1 transcript was cleaved from the long H-strand primary transcript (Fig. 2B), as other mRNA and rRNA transcripts were processed. In this study, we also discovered a few hsa-MDL1 antisense (hsa-MDL1AS) transcripts (NC_012920: 16024-407). However, we had not found any eft-MDL1 antisense transcript in the previous study. Although the full-length transcripts of MDL1 gene were discovered in human, rats, mice and insects to prove its ubiquitous existence, we still validated the hsa-MDL1 transcript using qPCR with Sanger sequencing to rule out the possible DNA contamination or RNA-

DNA chimeric sequences (Supplementary file 3). The hsa-MDL1 and hsa-MDL1AS sequences were submitted to the GenBank database under the accession number KX859178. In addition, the rno-MDL1 and mmu-MDL1 sequence for rat and mouse were submitted to the GenBank database under the accession number MF133497 and MF133498, respectively.

Since the detected transcriptional level of hsa-MDL1AS was much lower than that of hsa-MDL1, we suspected that some hsa-MDL1AS transcripts had been processed into small RNAs for specific biological roles. To validate this hypothesis, we aligned the public small RNA-seq data (Materials and methods) to the human mitochondrial genome. Theoretically, the enriched small RNAs should have only appeared as peaks in the tRNA or rRNA regions, but we obtained a great number of reversely aligned small RNAs enriched in the D-loop region (Fig. 1CD). The fowardly aligned small RNAs in the human D-loop region were less than 5% of the reversely aligned one. Particularly, 5′ ends of the most abundant small RNA **A**AAGATAAAATTTGAAAT (NC_012920: 407-390) with the length of 18 nt were precisely aligned to IT$_L$ (NC_012920: 407). This sequence was validated to be exist in 931 runs of human small RNA-seq data (Wang et al., 2016a). Since this small RNA overlapped the L-strand promoter CATACCGCCA**A**AAGATA (NC_012920: 417-401), it could belong to the transcription initial RNAs (tiRNAs) (Taft et al., 2009) and the hsa-MDL1AS transcript was its precursor. Therefore, we named this small RNA as hsa-tir-MDL1AS-18. Further analysis showed that hsa-tir-MDL1AS included a series of small RNAs with lengths from 18 to 25 nt. We also found a 20-nt palindrome small RNA **A**AAGACACCC|CCCAC**A**GTTT (NC_012920: 561-580) containing the TIS and the Transcription Termination Site (TTS) of H-strand (underlined). Then, we named this palindrome small RNA (psRNA) as hsa-tir-MDL1-20.

Since hsa-MDL1 and hsa-MDL1AS were precursors of hsa-tir-MDL1-20 and hsa-tir-MDL1AS-18, they were defined as long transcription initial RNAs (ltiRNAs) constituting a novel class of long non-coding RNAs (lncRNAs). We proposed that ltiRNAs and tiRNAs could constitute a regulation system (Fig. 2B) to ensure that the expression levels of human mitochondrial genes were able to fluctuate in a large dynamic range. This ltiRNA/tiRNA regulation system had both positive and negative feedback mechanisms to control the expression levels of mitochondrial genes as a whole. Positive feedback can increase the transcription of all mitochondrial genes to ensure a high productive efficiency. Negative feedback could be used to maintain the expression of mitochondrial genes at normal levels by sense-antisense RNA interactions. Further analysis of the small RNA-seq data showed the transcriptional levels of hsa-tir-MDL1AS-18 in normal tissues were significantly higher than those in Hepatocellular Carcinoma (HCC) tissues (Fig. 1CD). This suggested the ltiRNA/tiRNA system could have a loss of balanced control in cancer cells.

2.4. H-strand primary transcript, RNA precursor and mature RNA

Our previous study has already proven that a high-content of polycistronic transcripts in the full-length transcriptome are mRNA or rRNA precursors and the analysis of RNA precursors facilitates the investigation of mitochondrial gene transcription and its regulation. By the analysis of RNA precursors, the H-strand primary transcript was precisely determined to start at the TIS 561 and end at the TTS 576 (Fig. 2B). Since the TTS was 15 (576–561) nt after the TIS, we were curious if the RNA polymerase had abilities to read through the TTS in the D-loop region after the H-strand primary transcript had been completely synthesized. Surprisingly, we found two long sequences from two other larger PacBio datasets to support this "read through" model. One 3853-nt sequence (NC_012920: 14628-1911) spanned the regions of ND6AS, tRNA^{Glu}AS, Cytb, tRNA^{Thr}, tRNA^{Pro}AS, D-loop, tRNA^{Phe}, 12S rRNA, tRNA^{Val} and 16S rRNA. Another 1727-nt sequence (NC_012920: 16202-1260) spanned the regions of D-loop, tRNA^{Phe} and 12S rRNA. However, we did not find any mutation around the TTS to explain this "read through" phenomenon.

Our previous study has also proven that the analysis of mitochondrial RNA precursors is indispensable for the studies of RNA processing, maturation and their mechanisms. In the classical 'tRNA punctuation' model, mitochondrial RNAs are transcribed as primary transcripts, then processed into polycistronic precursors and finally mature RNAs by enzyme cleavage. RNA cleavage is processed at 5′ and 3′ ends of tRNAs. By the analysis of RNA precursors, we constructed a "mitochondrial cleavage" model to improve the 'tRNA punctuation' model. In the new model, RNA cleavage can be processed: 1) at 5′ and 3′ ends of tRNAs, 2) between mRNAs and mRNAs (*e.g.* ATP8/6 and COIII) and 3) between mRNAs and antisense tRNAs (*e.g.* COI and tRNA^{Ser}AS), but can not be processed: 1) between mRNAs and lncRNAs (*e.g.* ND5 and ND6AS), 2) between antisense tRNAs and lncRNAs (*e.g.* tRNA^{Pro}AS and D-loop), 3) between lncRNAs and antisense tRNAs (*e.g.* ND6AS and tRNA^{Glu}AS) and 4) between antisense tRNAs and antisense tRNAs (*e.g.* tRNA^{Ala}AS/tRNA^{Asn}AS/tRNA^{Cys}AS/tRNA^{Tyr}AS). Particularly, tRNA^{Ala}AS-tRNA^{Tyr}AS (NC_012920: 1318-1638) could not be further cleaved, which was against a hypothesis from one previous study (Seligmann, 2010). That hypothesis was both tRNAs and antisense tRNAs could be recognized by enzyme for cleavage. However, our "cleavage" model supports that most of antisense tRNAs in mtDNA can not be further cleaved into single antisense tRNAs. Our model can also be used to explain the unexpected low level of ND6. The reason could be most of mature ND6 transcripts containing the downstream non-coding sequences (Fig. 2A) were not amplified for PacBio sequencing due to overlength. Based on our model, the H-strand primary transcript should be cleaved into 30 transcripts: tRNA^{Ile}, tRNA^{Gln}AS, tRNA^{Met}, ND2, tRNA^{Trp}, tRNA^{Ala}AS-tRNA^{Tyr}AS, COI, tRNA^{Ser}AS, tRNA^{Asp}, COII, tRNA^{Lys}, ATP8/6, COIII, tRNA^{Gly}, ND3, tRNA^{Arg}, ND4L/4, tRNA^{His}, tRNA^{Ser}, tRNA^{Leu}, ND5/ND6AS/tRNA^{Glu}AS, Cytb, tRNA^{Thr}, MDL1(tRNA^{Pro}AS/D-loop), tRNA^{Phe}, 12S rRNA, tRNA^{Val}, 16S rRNA, tRNA^{Leu} and ND1. Presumably, the L-strand primary transcript should be cleaved into 14 transcripts: MDL1, tRNA^{Pro}, tRNA^{Thr}AS-Cytb, tRNA^{Glu}, ND6-tRNA^{Asp}AS, tRNA^{Ser}, COIAS, tRNA^{Tyr}, tRNA^{Cys}, tRNA^{Asn}, tRNA^{Ala}, tRNA^{Trp}AS-tRNA^{Met}AS, tRNA^{Gln} and tRNA^{Ile}AS.

3. Conclusions

In this study, we established a general framework to use PacBio full-length transcriptome sequencing for the investigation of mitochondrial RNAs. The main results included determination of the H-strand primary transcript, identification of ND5/ND6AS/tRNA^{Glu}AS and tRNA^{Ala}AS/tRNA^{Asn}AS/tRNA^{Cys}AS/tRNA^{Tyr}AS, discovery of palindrome small RNAs and construction of the "cleavage" model. The most important finding was two novel lncRNAs of MDL1 and MDL1AS from mitochondrial D-loop regions. The existence of MDL1 and MDL1AS in human, rats, mice and insects was confirmed using big data and qPCR with high confidence, but ND5/ND6AS/tRNA^{Glu}AS and tRNA^{Ala}AS/tRNA^{Asn}AS/tRNA^{Cys}AS/tRNA^{Tyr}AS need be further confirmed in more species or animal tissues.

The hsa-MDL1 and hsa-MDL1AS transcripts were predicted as the precursors of tiRNAs, the most abundant of which was hsa-tir-MDL1AS-18 (AAAGATAAAATTTGAAAT). We proposed that hsa-MDL1 and hsa-MDL1AS constituted a novel class of long non-coding RNAs (lncRNAs), which were named as long transcription initial RNAs (ltiRNAs). The phenomenon of tiRNAs was observed but was not highly regarded because of their unknown origin. Then, tiRNAs were considered as a new class of small RNAs that were predominantly 18 nt in length and mapped within −60 to +120 nt of TISs in human, chicken and *Drosophila*. Previous studies also reported that tiRNAs could be on the same strand as TISs and preferentially associated with GC-rich promoters. Although tiRNAs were associated to some biological functions or human diseases by several experiments, the theories and research models of tiRNAs have not been built without knowledge of their precursors. In this study, we associated tiRNAs to ltiRNAs and proposed that ltiRNAs and tiRNAs could constitute a system to regulate gene expression with new mechanisms. Since mtDNAs originated from a eubacterial (specifically alpha-proteobacterial) ancestor (Gray et al., 2001), the ltiRNA/tiRNA system maybe exist ubiquitously in eukaryotes and prokaryotes.

4. Materials and methods

4.1. Full-length human transcriptome datasets

MCF-7 is a cancer cell line that was first isolated from breast tissues of a 69-year old Caucasian woman in 1970. Full-length transcriptome data of the human MCF-7 cell line were downloaded from the amazon website (http://datasets.pacb.com.s3.amazonaws.com), including an old dataset (MCF-7 2013) using P4/C2 sequencing reagents and a new dataset (MCF-7 2015) using P5/C3 sequencing reagents. Since P5/C3 reagents increased the sequencing length, we used the MCF-7 2015 dataset for identification of RNA precursors and mature RNAs. To produce the MCF-7 2015 dataset, the cDNA libraries of MCF-7 had been prepared using SMARTer PCR cDNA Synthesis Kit (Clontech, USA). Size selection had been implemented using a SageELF device (Sage Science, USA) with every two lanes binned together to create libraries with sizes of 0–1 Kbp, 1–2 Kbp, 2–3 Kbp, 3–5 Kbp and 5–7 Kbp. These libraries had been sequenced on the Pacific Biosciences RS II sequencer using 28 SMART Cells in the year of 2014. The MCF-7 2013 dataset containing data from 119 SMART Cells was used to confirm the results from the MCF-7 2015 dataset. Another larger full-length human transcriptome dataset had been obtained by sequencing of normal brain, heart and liver tissues using 115 SMART Cells. This dataset was downloaded from http://www.pacb.com/blog/data-release-whole-human-transcriptome and also used to confirm the results from the MCF-7 2015 dataset.

To perform transcriptome comparison, we also used the full-length *E. fullo* mitochondrial transcriptome data from our previous study. Since the lengths of mitochondrial rRNA and mRNA transcripts distribute in the range of 1–2 Kbp, the insect mitochondrial cDNAs had been amplified without size selection for sequencing. To produce the MCF-7 2015 dataset, the human mitochondrial cDNAs had been amplified with size selection to reduce the bias caused by SMART Cell loading before sequencing. However, there are many other experimental reasons which could induce bias into sequencing data. These reasons include amplification efficiency of cDNAs with different sizes and data yield of different SMART Cells et al. Therefore, cDNA amplification with size selection cannot ensure a more accurate quantification of the full-length transcriptome.

4.2. HCC small RNA-seq dataset

The public small RNA-seq dataset used to identify tiRNAs was downloaded from the NCBI SRA database under the project accession number SRP002272 (Supplementary file 3). This dataset included 15 clinical samples, which were three normal liver tissues, one HBV-infected liver tissue, one severe chronic Hepatitis B liver tissue, two Hepatitis B Virus (HBV) positive Hepatocellular Carcinoma (HCC) tissues, one Hepatitis C Virus (HCV) positive HCC tissue, one HCC tissue without HBV or HCV and six controls. The data analysis (*e.g.* alignment) was conducted following the procedure used in our previous study (Wang et al., 2016a).

4.3. Data analysis

The human mitochondrial reference genome (RefSeq: NC_012920.1) was downloaded from the NCBI RefSeq database, which has the same sequence as UCSC hg18. Since the mitochondrial D-loop region had been deemed as a non-transcriptional genomic region before this study, the human mitochondrial reference genome contains an interrupted D-loop region. We had to reconstruct the human mitochondrial reference genome for the convenience of annotation and visualization by two steps. The “N” nucleotide was removed from the reference genome (RefSeq: NC_012920.1) and the resulted sequence shifted 4263 bp counter-clockwise (Supplementary file 1).

The IsoSeq™ protocol (Pacific Biosciences, USA) was used to process the sequenced reads to Circular Consensus Sequencing (CCS) reads with parameters (Minimum Full Passes = 1, Minimum Predicted Accuracy = 75), then to produce draft transcripts with parameters (Minimum Sequence Length = 300) by removing the 5' end cDNA primers, 3′ end cDNA primers and 3′ ployA sequences, which had been identified by the pipeline Fastq_clean (Zhang et al., 2014). Fastq_clean is a Perl based pipeline to clean DNA-seq (Wang et al., 2016b), RNA-seq (Cao et al., 2016) and small RNA-seq data (Wang et al., 2016a) with quality control and has included tools to process PacBio data in the version 2.0 (https://github.com/gaoshanT/Fastq_clean). The software BWA v0.7.12 was used to align draft transcripts to the reconstructed human mitochondrial genome. Alignment quality control and filtering were performed using in-house Perl programs to remove errors in draft transcripts from the IsoSeq™ protocol. Aligned transcripts with query coverages less than 90% or identities less than 90% were removed to filter out alignments with poor quality. Statistics computation and plotting were performed using the software R v2.15.3 with the Bioconductor packages (Gao et al., 2014). All the identified transcripts (Supplementary file 2) were observed and curated using the software Tablet v1.15.09.01 (Milne et al., 2012).

Consent to publish

Not applicable.

Competing interests

No potential conflicts of interest were disclosed.

Funding

This work was supported by grants from Fundamental Research Funds for the Central Universities (for Nankai University) to Shan Gao, Internationalization Training of Excellent Postdoc Project from Tianjin Government to Shan Gao and National Key Technology R & D Program (2012BAI29B01) to Xiumei Gao.

Authors' contributions

SG and WB conceived and supervised this project. SG, XT and ZW analyzed the data. YS and HC curated the sequences and prepared all the figures, tables and additional files. ZC and QZ conducted the qPCR experiments. SG drafted the main manuscript. JR and PD revised the manuscript. All authors have read and approved the manuscript.

Availability of data and materials

Full-length transcriptome data of the human MCF-7 cell line can be obtained from the amazon website (http://datasets.pacb.com.s3.amazonaws.com). The public small RNA-seq dataset used to identify tiRNAs can be obtained from the NCBI SRA database under the project accession number SRP002272.

Acknowledgements

We appreciate the help equally from the people listed below. They are the graduate student Ekta Singh, Xiaoran Niu, and Professor Guoqing Liu from College of Life Sciences, Nankai University. Dr. Gloria Sheynkman from Dana-Farber Cancer Institute and Dr. Roberto Lleras from Pacific Biosciences kindly provided the experiment information of the full-length human transcriptome. The data analysis in this study was supported by National Scientific Data Sharing Platform for Population and Health Translational Cancer Medicine Specials.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.mito.2017.08.002.

References

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., et al., 1981. Sequence and organization of the human mitochondrial genome. Nature 290 (5806), 457–465.

Boore, J.L., 1999. Animal mitochondrial genomes. Nucleic Acids Res. 27 (8), 1767–1780.

Cao, Q., Li, A., Chen, J., Sun, Y., Tang, J., Zhang, A., Zhou, Z., Zhao, D., Ma, D., Gao, S., 2016. Transcriptome sequencing of the sweet potato progenitor (ipomoea trifida (HBK) G. Don.) and discovery of drought tolerance genes. Trop. Plant Biol. 1–10.

Chang, D.D., Clayton, D.A., 1984. Precise identification of individual promoters for transcription of each strand of human mitochondrial DNA. Cell 36 (3), 635–643.

Gao, S., Ou, J., Xiao, K., 2014. R Language and Bioconductor in Bioinformatics Applications, Chinese Edition. Tianjin Science and Technology Translation Publishing Co. Ltd., Tianjin.

Gao, S., Ren, Y., Sun, Y., Wu, Z., Ruan, J., He, B., Zhang, T., Yu, X., Tian, X., Bu, W., 2016. PacBio Full-length transcriptome profiling of insect mitochondrial gene expression. RNA Biol. 13 (9), 820–825.

Gray, M.W., Burger, G., Lang, B.F., 2001. The origin and early evolution of mitochondria (reviews). Genome Biol. 2 (6), 181–200 (Genome Biol 2:1018).

Mercer, T.R., Neph, S., Dinger, M.E., Crawford, J., Smith, M.A., Shearwood, A.-M.J., Haugen, E., Bracken, C.P., Rackham, O., Stamatoyannopoulos, J.A., 2011. The human mitochondrial transcriptome. Cell 146 (4), 645–658.

Milne, I., Stephen, G., Bayer, M., Cock, P.J., Pritchard, L., Cardle, L., Shaw, P.D., Marshall, D., 2012. Using Tablet for visual exploration of second-generation sequencing data. Brief. Bioinform. bbs012.

Montoya, J., Attardi, G., 1982. Identification of initiation sites for heavy-strand and light-strand transcription in human mitochondrial DNA. Proc. Natl. Acad. Sci. U. S. A. 79 (23), 7195–7199.

Rackham, O., Shearwood, A.M., Mercer, T.R., Davies, S.M., Mattick, J.S., Filipovska, A., 2011. Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins. Rna-a Publ. Rna Soc. 17 (12), 2085.

Ren, Y., Jiaqing, Z., Sun, Y., Wu, Z., Ruan, J., He, B., Liu, G., Gao, S., Bu, W., 2016. Full-length transcriptome sequencing on PacBio platform. Chin. Sci. Bull. 61 (11), 1250–1254 (in Chinese).

Seligmann, H., 2010. Undetected antisense tRNAs in mitochondrial genomes? Biol. Direct 5 (1), 1–25.

Shiota, T., Imai, K., Qiu, J., Hewitt, V.L., Tan, K., Shen, H.-H., Sakiyama, N., Fukasawa, Y., Hayat, S., Kamiya, M., 2015. Molecular architecture of the active mitochondrial protein gate. Science 349 (6255), 1544–1548.

Stewart, J.B., Beckenbach, A.T., 2009. Characterization of mature mitochondrial transcripts in *Drosophila*, and the implications for the tRNA punctuation model in arthropods. Gene 445 (1), 49–57.

Taft, R.J., Glazov, E.A., Cloonan, N., Simons, C., Stephen, S., Faulkner, G.J., Lassmann, T., Forrest, A.R., Grimmond, S.M., Schroder, K., 2009. Tiny RNAs associated with transcription start sites in animals. Nat. Genet. 41 (5), 572–578.

Wang, F., Sun, Y., Ruan, J., Chen, R., Chen, X., Chen, C., Kreuze, J.F., Fei, Z., Zhu, X., Gao, S., 2016a. Using small RNA deep sequencing data to detect human viruses. Biomed. Res. Int. 2016 (2016), 2596782.

Wang, Y., Wang, Z., Chen, X., Zhang, H., Guo, F., Zhang, K., Feng, H., Gu, W., Wu, C., Ma, L., 2016b. The complete genome of Brucella suis 019 provides insights on cross-species infection. genes 7 (2), 7.

Zhang, M., Zhan, F., Sun, H., Gong, X., Fei, Z., Gao, S., 2014. Fastq_clean: an optimized pipeline to clean the Illumina sequencing data with quality control. In: Bioinformatics and Biomedicine (BIBM). IEEE International Conference on 2014 IEEE.