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

Shan Gao[1$], Xiaoxuan Tian[2$], Yu Sun[1$], Zhenfeng Wu[3], Zhi Cheng[1],
Pengzhi dong[2], Bingjun He[1], Jishou Ruan[3*], Wenjun Bu[1*]

1 College of Life Sciences, Nankai University, Tianjin, Tianjin 300071, P.R.China.
2 Tianjin State Key Laboratory of Modern Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin 300193, P.R.China.
3 School of Mathematical Sciences, Nankai University, Tianjin 300071, P.R.China.
4. Institute of Statistics, Nankai University, Tianjin 300071, P.R.China.

[$] These authors contributed equally to this paper.
* The corresponding authors.
JR Email: jsruan@nankai.edu.cn
Tel: +86-022-23501449

Abstract

In this study, we introduced a general framework to use PacBio full-length transcriptome sequencing for the investigation of the fundamental problems in mitochondrial biology, *e.g.* genome arrangement, heteroplasmy, RNA processing and the regulation of transcription or replication. As a result, we produced the first full-length human mitochondrial transcriptome from the MCF7 cell line based on the PacBio platform and characterized the human mitochondrial transcriptome with more comprehensive and accurate information. The most important finding is two novel genes hsa-MDL1 and hsa-MDL1AS, which are encoded by the mitochondrial D-loop regions. We propose hsa-MDL1 and hsa-MDL1AS, as the precursors of transcription initiation RNAs (tiRNAs), belong to a novel class of long non-coding RNAs (lnRNAs), which is named as long tiRNAs (ltiRNAs). Based on the mitochondrial RNA processing model, the primary tiRNAs, precursors and mature tiRNAs could be discovered to completely reveal tiRNAs from their origins to functions. The ltiRNA/tiRNA system and their regulation mechanisms could exist ubiquitously in eukaryotes and prokaryotes. These findings will enrich the basic concepts in the field of mitochondrial biology, lnRNA functions and regulation of gene expression.

Keywords: mitochondrial transcriptome; full-length transcriptome; D-loop; MDL1; ltiRNA

Background

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 [1]. The replication and transcription of human mtDNA is 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 modifications. These mature RNAs are pivotal for cellular Adenosine Triphosphate (ATP) production, numerous metabolic regulatory processes, the programming of cell death and other cell functions [2]. However, the exact mechanisms of mitochondrial gene transcription and its regulation are still not well understood [3].

In the year of 1981, the complete human mtDNA sequence was determined and characterized by the encoded genes [4]. With the Next Generation Sequencing (NGS) technologies, Mercer *et al*. tried to demonstrate wide variation in mitochondrial transcript abundance and resolve transcript processing and maturation events [5]. 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 [6] by sequencing the full-length insect (*Erthesina fullo* Thunberg) transcriptome [7] on the PacBio platform and built a straightforward and concise methodology to investigate mitochondrial gene transcription. Most of the results in the study of the full-length insect mitochondrial transcriptome were consistent with the 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 of tRNA punctuation *et al*. Using the full-length insect mitochondrial transcriptome data, Gao *et al*. have proved the high-content polycistronic transcripts are mRNA or rRNA precursors and the analysis of these precursors facilitates the investigation of RNA processing, maturation and their mechanisms.

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 transcriptome with more comprehensive and accurate information. Our research objectives included: 1) to introduce 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 further studies; 3) to study basic 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

questions, *e.g.* mtDNA heteroplasmy, RNA processing and the regulation of mtDNA transcription or replication.

Results and discussion

The quantitative transcription map of human mitochondrial genome

The full-length transcriptome dataset of the human MCF-7 cell line was downloaded to identify mitochondrial transcripts (**Materials and Methods**). This dataset had been acquired by sequencing five groups of cDNA libraries with the size 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, raw reads were split into 9,192,530 high-quality (Accuracy >= 75%) subreads with the average size of 2,668.85 bp. All the subreads were processed into Circular Consensus Sequencing (CCS) reads to further improve the data quality. Finally, 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 draft transcripts could be continuously mapped to the reconstructed human mitochondrial reference genome (**Materials and Methods**) 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 7.2% (21,689/301,149) using only data from nine of 28 libraries with the size 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 [6] (**Figure 1A**). 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 the experimental reasons (**Materials and Methods**). In this study, the first quantitative transcription map of human mitochondrial genome (**Figure 1B**) was constructed to show the positions, strands and transcriptional levels of all the identified transcripts and their full-length sequences were confirmed by at least 10 full-path CCS reads for their fidelity (**Supplementary file 1**).

Comparison of the human and insect mitochondrial transcriptome

The quantitative transcription map of insect mitochondrial genome (**Figure 1A**) has indicated eight mRNA transcripts (ND2, COI, COII, ATP8/ATP6, COIII, ND3, ND6 and Cytb) are encoded by the major coding strand J(+), also known as the Heavy strand (H-strand) [6], while three mRNA transcripts (ND4L/ND4, ND5 and ND1) and two rRNA transcripts (16S rRNA and 12S rRNA) are encoded by the minor coding strand N(-), also known as the Light strand (L-strand). From **figure 1B**, it can be seen the all the abundant transcripts (in red color) of the human mtDNA were encoded by J(+), while their antisense transcripts (in blue color) at comparatively low levels were encoded by N(-). The further analysis of all the

polycistronic transcripts supported J(+) and N(-) were transcribed as two long primary transcripts respectively. The results in this study also showed although the ND6 transcript encoded by L-strand is responsible for the protein coding, the ND6 antisense transcript encoded by H-strand (ND6AS) is still more abundant. This finding was different from all the existing human mitochondrial transcription models from the previous studies, in which the ND6 transcript encoded by L-strand had been identified as the abundant transcript [8] [4]. Using the NGS data, the ND6 transcript and ND5 antisense transcript (ND5AS) encoded by L-strand were identified as the abundant transcripts in the year of 2011 [5]. The main cause of the differences is the PCR or NGS data do not contain the RNA polyA tails to determine the 3' ends of the mapped reads on the mitochondrial genomes. In the full-length human mitochondrial transcriptome, we did not find any full-length transcript of ND5 or ND6AS mature RNAs. This suggested ND5/ND6AS could be one of the fusion genes in the human mitochondrial genome, besides two other fusion genes (ATP8/ATP6 and ND4/ND4L) which had also been found in the insect mitochondrial genome. 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 the expression profile of the insect mitochondrial transcriptome due to the tissue specificity and other experimental reasons (**Materials and Methods**), 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 (**Figure 1AB**). In addition, we found eight high- and medium-level mRNA transcripts (COII, COI, COIII, Cytb, ATP8, ATP6, ND4L and ND4) in human mitochondrial genome had the preference to use "ATG" as the start codons, while the other low-level mRNA transcripts had the preference to "ATA" or "ATT" except ND6. This suggested the expression of human mitochondrial genes could be regulated coordinately on both of transcriptional and translational levels to ensure the high efficiency. Therefore, the regulation on the transcriptional level could have a major influence on the variation of gene expression levels, compared to post-transcriptional levels on the downstream.

By comparing the human and insect quantitative transcription map, we identified four Locally Collinear Blocks (LCBs) [9]. They were LCB1 (ND2, COI, COII, ATP8/ATP6, COIII and ND3), LCB2 (ND5 and ND4/ND4L), LCB3 (ND6 and Cytb) and LCB4 (ND1, 16S rRNA and 12S rRNA). From **figure 1AB**, it can be seen that LCB2 and LCB4 appears to be chromosomal inversions (reversed end to end) between the insects and human. The inversion of LCB2 from the insect to human mitochondrial genome could form the fusion of the gene ND5 and ND6AS. The transcriptional levels of ND5 are much higher than the levels of ND6 in insects but are close to the levels of ND6AS in human possibly due to their gene fusion. Since a multitude of RNA precusors contain the ND5/ND6AS junction regions, to determine the gene fusion of ND5/ND6AS requires high-depth and full-length transcriptome sequencing of the human mitochondrial extract.

Reconstruction of the human mitochondrial reference genome

Since the mitochondrial D-loop region had been deemed as a non-transcriptional genomic region before this study, the human mitochondrial reference genome (Genbank: NC_012920.1) contains an interruptted D-loop region. We had to reconstruct the human mitochondrial reference genome for the convenience of annotation and visualization (**Materials and Methods**). As a result, the new reference genome (**Supplementary file 2**) has been annotated with corrections using the full-length transcriptome data (**Table 2**). The important gonomic features have been added to include the origin of H-strand replication (O_H), the origin of L-strand replication (O_L), the initiation of the first transcription site on H-strand (IT_{H1}), the initiation of the second transcription site on H-strand (IT_{H2}) and the initiation of transcription site on L-strand (IT_L). O_L (mtDNA: 1632), O_H (mtDNA: 12617), IT_L (mtDNA: 12714) and IT_{H1} (mtDNA: 12868) have been determined, which validated predictions from the previous studies, but IT_{H2} (mtDNA: 12945) has not been determined in this study. O_H and O_L were identified using Single Nucleotide Polymorphisms (SNPs) in the human MCF-7 cell line. Among a total of 17 SNPs (**Supplementary file 3**), 13 single base substitutions were A507G (ND2), T2514C (COI), A4598G (ATP8), G5704A (COIII), T8998C (ND5), T10057C (ND6), A11064G (Cytb), A11118G (Cytb), C11886T (D-loop), T12257C (D-loop), A12570G (D-loop), A13057G (12S rRNA) and A13745G (12S rRNA), while four insertion-deletion (InDel) sites included O_L, O_H (D-loop), the position 3265 ($tRNA^{Asp}$) and 14924 (16S rRNA). These results suggested the position 14924 (16S rRNA) and O_L contributed most to the mtDNA heteroplasmy.

Using polycistronic transcripts, IT_L and IT_{H1} were identified and confirmed starting at the exact position 12714 (D-loop) and 12868 (D-loop) on the human mitochondrial genome [10]. The 15-nt promoter element CAAACCCC**A**AAGACA (mtDNA: 12860-12874) surrounding IT_{H1} (underlined) was also consistent with the finding in the previous study [11]. However, we did not find any full-length transcript to validate H-strand transcription starting at the exact position 12945 ($tRNA^{Phe}$), 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 [10]. Therefore, to determine IT_{H2} requires high-depth and full-length transcriptome sequencing of the human mitochondrial extract.

Novel genes and mechanisms of gene transcription regulation

In the previous study of the insect mitochondrial transcriptome, we discovered some transcripts aligned forwardly to the D-loop region. The further analysis has indicated these transcripts are encoded by a novel gene Mitochondrial D-loop 1 (MDL1) of *Erthesina fullo* Thunberg (eft) with the length of 1,024 bp. MDL1 (eft-mtDNA: 15210-16233) covers 63.64% (1,024/1,609) of the insect D-loop region (eft-mtDNA: 14625-16233) and contains 11 or 12 repeated units (eft-mtDNA: 15260-16233). This gene with high diversity has two high-frequency mutation sites in each repeated unit (**Figure 2A**). In this study, we discovered some

polycistronic transcripts (**Supplementary file 1**) aligned forwardly to the human D-loop region. The most abundant one of these transcripts ($tRNA^{Pro}$/D-loop) covered the gene $tRNA^{Pro}$ (mtDNA: 11694-11761) and 100% of the human D-loop region (mtDNA: 11762-12883). 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 (mtDNA: 11762-12883), respectively. The analysis of all the RNA precursors of hsa-MDL1 supported the mature hsa-MDL1 transcript was cleaved from the long H-strand primary transcript (**Figure 2B**), as the other mRNA and rRNA transcripts were processed. This long H-strand primary transcript started at the position 12868 and ended at the position 12883, including a 20-nt palindrome sequence AAAGACACCC|CCCACAGTTT (mtDNA: 12868-12887) overlapping the 15-nt promoter CAAACCCCAAAGACA (mtDNA: 12860-12874) predicted in the previous study. We also discovered a few antisense transcripts overlapping hsa-MDL1 and named their gene as hsa-MDL1AS (mtDNA: 11762-12714). However, we had not found any antisense transcript overlapping eft-MDL1 in the previous study. Although the full-length transcripts of MDL1 gene were obtained from both of human and insects, we still validated the hsa-MDL1 transcript using qPCR with Sanger sequencing to rule out the DNA contamination or RNA-DNA chimeric sequences (**Supplementary file 3**) and the hsa-MDL1/hsa-MDL1AS sequence was submitted to the Genbank database with ID KX859178.

Since the detected transcriptional level of hsa-MDL1AS was much lower than the level of hsa-MDL1, we suspected 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 and obtained reversely aligned small RNAs enriched in the D-loop region (**Figure 1CD**). Particularly, the 5' ends of the most abundant small RNA AAAGATAAAATTTGAAAT (mtDNA: 12697-12714) with the length of 18 nt were precisely aligned to IT_L (mtDNA: 12714). This sequence was validated to be exist in 931 runs of human small RNA-seq data [12] and named as hsa-tir-MDL1AS-1. The fowardly aligned small RNAs in the human D-loop region were less than 5% of the reversely aligned small RNAs. Most of them were enriched at the 5' ends or 3' ends of the hsa-MDL1 gene. These results suggested small RNAs mapped to the D-loop region were the transcription initial RNAs (tiRNAs) [13] and hsa-MDL1/hsa-MDL1AS transcripts were their precursors. Since hsa-MDL1 and hsa-MDL1AS are encoded by the transcription initial region and have lengths ~1 Kbp, they are defined as long transcription initial RNAs (ltiRNAs) in this study, which is a novel class of long non-coding RNAs (lnRNAs). We propose that ltiRNAs and tiRNAs consititue a regulation system (**Figure 2B**) to ensure the expression levels of the mitochondrial genes are able to fluctuate in a large dynamic range. This ltiRNA/tiRNA regulation system have both positive and negative feedback mechanisms to control the expression levels of the mitochondrial genes. Positive feedback can initialize the transcription of all the 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 interactions. In addition, the further analysis of the small RNA-seq data showed the transcriptional levels of the tiRNA hsa-tir-MDL1AS-1 in normal tissues

were significantly higher than the levels of hsa-tir-MDL1AS-1 in Hepatocellular Carcinoma (HCC) tissues (**Figure 1CD**). This suggested the ltiRNA/tiRNA system could have a loss of balance control in cancer cells. Since the previous studies demonstrated that the replication of mammalian mtDNA are intimately linked with mitochondrial transcription, the ltiRNA/tiRNA system could also be involved in mtDNA replication. However, the relationship between the mtDNA transcription and replication is still not clear.

Using RNA precursors to study mitochondrial gene transcription

Our previous study have already proved the high-content of polycistronic transcripts in the full-length transcriptome are mRNA or rRNA precursors and the analysis of these precursors is indispensable for the studies of RNA processing, maturation and their mechanisms. This study has proved the analysis of RNA precursors plays an important role in the investigation of mitochondrial gene transcription and its regulation. Since the PCR or NGS methods cannot differentiate RNA precursors from their mature transcripts, researchers may consider to use PacBio full-length transcriptome sequencing in the future studies. In the full-length insect mitochondrial transcriptome, we found two main types of RNA precursors, which ended with mRNAs or rRNAs. In this study, we found a new type of precursors, which ended with reverse encoded tRNAs. For example, the COI/tRNA^{Pro} transcript included COI encoded by H-strand and tRNA^{Pro} encoded by L-strand. All three types of RNA precursors supported the “forward cleavage” model which had been proposed in our previous study [6]. In this model, the cleavage of tRNAs at 5' ends precedes 3' ends. Since the current protocol can only capture RNAs with 3' polyA tails in the full-length transcriptome sequencing, this suggested the 3' polyadenylation of mature RNAs could be tightly coupled with the tRNA cleavage at the 5' end. By the analysis of RNA precursors, the insect mtDNA was identified to have four primary transcripts with several antisense transcripts. The human mtDNA was identified to have two long primary transcripts. In this study, we used the analysis of RNA precursors to identify the abundant ND5/ND6AS transcripts, the initiation of transcription sites and particularly two novel genes hsa-MDL1 and hsa-MDL1AS *et al*. With a multitude of RNA precursors, the hsa-MDL1 and hsa-MDL1AS transcript have finally be determined to be cleaved from the long primary transcripts.

Conclusions

In this study, we introduced a general framework to use PacBio full-length transcriptome sequencing for the investigation of the fundamental problems in mitochondrial biology, *e.g.* genome arrangement, heteroplasmy, RNA processing and the regulation of transcription or replication. As a result, we produced the first full-length human mitochondrial transcriptome from the MCF7 cell line based on the PacBio platform and characterized the human mitochondrial transcriptome with more comprehensive and accurate information. The new findings corrected some concepts of mitochondrial gene transcription from the previous studies using qPCR [8] or the Illumina NGS technology [5]. The most important finding was two

novel genes MDL1 and MDL1AS, which were encoded by the mitochondrial D-loop regions. The hsa-MDL1 and hsa-MDL1AS transcript were identified as the precursors of tiRNAs, the most abundant of which was hsa-tir-MDL1AS-1 (AAAGATAAAATTTGAAAT). We propose that hsa-MDL1 and hsa-MDL1AS belong to a novel class of long non-coding RNAs (lncRNAs) and are named as long transcription initial RNAs (ltiRNAs). ltiRNAs and tiRNAs consititue a system to regulate gene expression with new mechanisams.

The phenomenon of tiRNAs was observed but neglected in the NGS data analysis from the beginning. 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 Transcription Initiation Sites (TISs) in human, chicken and *Drosophila*. The previous studies also reported tiRNAs could be on the same strand as the TISs and preferentially associated with GC-rich promoters. The characteristics of tiRNAs discovered using the full-length transcriptom data with the small RNA-seq data in this study were consistent with predominantly 18 nt in length but not consistent with the mapped genome range (-60 to +120 nt of TISs), high G+C content and transcription strand preferences discovered in the previous studies. Although tiRNAs were associated to some biological functions or human deseases by several experiments, the theroies and research models of tiRNAs have not be built without the knowledge of their precursors. In this study, we associated tiRNAs to their precursors ltiRNAs and proposed the ltiRNAs/tiRNAs regulation system. Based on the mitochondrial RNA processing model, the primary tiRNAs, precursors and mature tiRNAs could be discovered to completely reveal tiRNAs from their origins to functions. Since the mitochondrial genome originated from a eubacterial (specifically alpha-proteobacterial) ancestor [14], the ltiRNA/tiRNA system exist ubiquitously in eukaryotes and prokaryotes.

Methods

Full-length human transcriptome dataset

MCF-7 is a cancer cell line that was first isolated in 1970 from the breast tissue of a 69-year old Caucasian woman. 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 (released in 2013) using the P4/C2 sequencing reagents and a new dataset (released in 2015) using the P5/C3 sequencing reagents. Since P5/C3 reagents increased the sequencing length, we only used the new dataset for this study. To produce this new dataset, the cDNA libraries of MCF-7 had been prepared using SMARTer PCR cDNA Synthesis Kit (Clontech, USA). Size selection had been implemented using the SageELF device (Sage Science, USA) with every two lanes binned together to create libraries with the size 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.

To conduct 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 in the study of the *E. fullo* mitochondrial transcriptome. In this study, the human mitochondrial cDNAs had been amplified with size selection to reduce the bias caused by the SMART Cell loading for sequencing. However, there are many other experimental reasons which could induce bias into the sequencing data. These experimental reasons include the amplification efficiency of cDNAs with different sizes and the 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.

HCC small RNA-seq dataset

The public small RNA-seq dataset used to identify tiRNAs was downloaded from the NCBI SRA database with ID SRP002272 (**Supplementary file 3**). This dataset included 15 clinical samples including 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 [12].

Data analysis

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 [15]. Fastq_clean is a Perl based pipeline to clean DNA-seq [16], RNA-seq [17] and small RNA-seq data [12] with quality control and had included some tools to process Pacbio data in the version 2.0 (https://github.com/gaoshanT/Fastq_clean). The reconstruction of the human mitochondrial reference genome included two steps. The "N" nucleotide was removed from the common-used human mitochondrial reference genome (Genbank: NC_012920.1) and the resulted sequence shifted 4,263 bp counter-clockwise (**Supplementary file 2**). 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 coverage less than 90% or identities less than 90% were removed to filter out alignments with poor quality. Statistics and plotting were conducted using the software R v2.15.3 with the Bioconductor packages [18]. All the identified transcripts (**Supplementary file 1**) were observed and curated using the software Tablet v1.15.09.01 [19].

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

LCB: Locally Collinear Block

O_H: the origin of H-strand replication

O_L: the origin of L-strand replication

IT_{H1}: the initiation of the first transcription site on H-strand

IT_{H2}: the initiation of the second transcription site on H-strand

IT_L: the initiation of transcription site on L-strand

SNP: Single Nucleotide Polymorphism

eft: *Erthesina fullo* Thunberg

HCC: Hepatocellular Carcinoma

TIS: Transcription Initiation Site

Declarations

Ethics (and consent to participate)

Not applicable

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, National Natural Science Foundation of China (31371974) to Bingjun He 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 processed the data. YS and BH curated the sequences and prepared all the figures, tables and additional files. 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 with accesstion number SRP002272.

Acknowledgements

We appreciate the help equally from the people listed below. They are the graduate student Xiaoran Niu, Hong Chang 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.

REFERENCES

1. Boore JL: **Animal mitochondrial genomes**. *Nucleic acids research* 1999, **27**(8):1767-1780.
2. Shiota T, Imai K, Qiu J, Hewitt VL, Tan K, Shen H-H, Sakiyama N, Fukasawa Y, Hayat S, Kamiya M: **Molecular architecture of the active mitochondrial protein gate**. *Science* 2015, **349**(6255):1544-1548.
3. Stewart JB, Beckenbach AT: **Characterization of mature mitochondrial transcripts in Drosophila, and the implications for the tRNA punctuation model in arthropods**. *Gene* 2009, **445**(1):49-57.
4. Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F *et al*: **Sequence and organization of the human mitochondrial genome**. *Nature* 1981, **290**(5806):457-465.
5. Mercer TR, Neph S, Dinger ME, Crawford J, Smith MA, Shearwood A-MJ, Haugen E, Bracken CP, Rackham O, Stamatoyannopoulos JA: **The human mitochondrial transcriptome**. *Cell* 2011, **146**(4):645-658.
6. Gao S, Ren Y, Sun Y, Wu Z, Ruan J, He B, Zhang T, Yu X, Tian X, Bu W: **PacBio Full-length transcriptome profiling of insect mitochondrial gene expression**. *RNA biology* 2016(just-accepted):00-00.

7. Ren Y, Jiaqing Z, Sun Y, Wu Z, Ruan J, He B, Liu G, Gao S, Bu W: **Full-length transcriptome sequencing on PacBio platform (in Chinese)**. *Chinese Sci Bull* 2016, **61**(11):1250-1254.
8. Sanchez MI, Mercer TR, Davies SM, Shearwood AM, Nygård KK, Richman TR, Mattick JS, Rackham O, Filipovska A: **RNA processing in human mitochondria**. *Cell Cycle* 2011, **10**(17):2904-2916.
9. Darling AC, Mau B, Blattner FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements**. *Genome Res* 2004, **14**(7):1394-1403.
10. Montoya J, Attardi G: **Identification of Initiation Sites for Heavy-Strand and Light-Strand Transcription in Human Mitochondrial DNA**. *Proceedings of the National Academy of Sciences of the United States of America* 1982, **79**(23):7195-7199.
11. Chang DD, Clayton DA: **Precise identification of individual promoters for transcription of each strand of human mitochondrial DNA**. *Cell* 1984, **36**(3):635-643.
12. Wang F, Sun Y, Ruan J, Chen R, Chen X, Chen C, Kreuze JF, Fei Z, Zhu X, Gao S: **Using small RNA deep sequencing to detect human viruses**. *BioMed Research International* 2016, **2016**(2016):2596782.
13. Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest AR, Grimmond SM, Schroder K: **Tiny RNAs associated with transcription start sites in animals**. *Nat Genet* 2009, **41**(5):572-578.
14. Gray MW, Burger G, Lang BF: **The origin and early evolution of mitochondria (Reviews). Genome Biol 2:1018**. *Genome Biol* 2001, **2**(6):181-200.
15. Zhang M, Sun H, Fei Z, Zhan F, Gong X, Gao S: **Fastq_clean: An optimized pipeline to clean the Illumina sequencing data with quality control**. In: *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on: 2014*: IEEE; 2014: 44-48.
16. Wang Y, Wang Z, Chen X, Zhang H, Guo F, Zhang K, Feng H, Gu W, Wu C, Ma L: **The Complete Genome of Brucella Suis 019 Provides Insights on Cross-Species Infection**. *Genes* 2016, **7**(2):7.
17. Cao Q, Li A, Chen J, Sun Y, Tang J, Zhang A, Zhou Z, Zhao D, Ma D, Gao S: **Transcriptome Sequencing of the Sweet Potato Progenitor (Ipomoea Trifida (HBK) G. Don.) and Discovery of Drought Tolerance Genes**. *Tropical Plant Biology* 2016:1-10.
18. Gao S, Ou J, Xiao K: **R language and Bioconductor in bioinformatics applications(Chinese Edition)**. Tianjin: Tianjin Science and Technology Translation and Publishing Co.; 2014.
19. Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D: **Using Tablet for visual exploration of second-generation sequencing data**. *Brief Bioinform* 2012:bbs012.

Figure legends

Figure 1 The quantitative transcription map of human mitochondrial genome

A. The quantitative transcription map of insect mitochondrial genome was from our previous study [9]. 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 (Genbank: NC_012920.1) and the resulted sequence shifted 4,263 bp counter-clockwise (Supplementary file 2). 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) were represented using their amino acids.

Figure 2 MDL1 and MDL1AS encoded by mitochondrial D-loop regions

A. The insect D-loop region (eft-mtDNA: 14625-16233) encodes the gene eft-MDL1 (eft-mtDNA: 15210-16233). The repeated region (eft-mtDNA: 15260-16233) contains 11 to 12 repeated units. Each unit has a length of 83 bp. The gene eft-MDL1 has two high-frequency mutation sites (in red columns). The genotypes for two mutations are "GTAG/***A" and "T/C". B. The human D-loop region (mtDNA: 11762-12883) encodes the gene hsa-MDL1 (mtDNA: 11762-12883) and its antisense gene hsa-MDL1AS (mtDNA: 11762-12714). The most abundant aligned small RNAs were marked by *.

Tables

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	3,072	9.34%
m140801_014337_42161_c100698070630000001823143403261501	0-1 Kbp	31,994	2,849	8.90%
m140801_050734_42161_c100698070630000001823143403261502	0-1 Kbp	35,335	3,757	10.63%
m140801_082952_42161_c100698070630000001823143403261503	0-1 Kbp	33,350	3,425	10.27%
m140801_115238_42161_c100698070630000001823143403261504	1-2 Kbp	42,673	2,425	5.68%
m140801_151509_42161_c100698070630000001823143403261505	1-2 Kbp	42,070	2,281	5.42%
m140804_215651_42141_c100700040630000001823139203261500	1-2 Kbp	36,564	2,028	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%

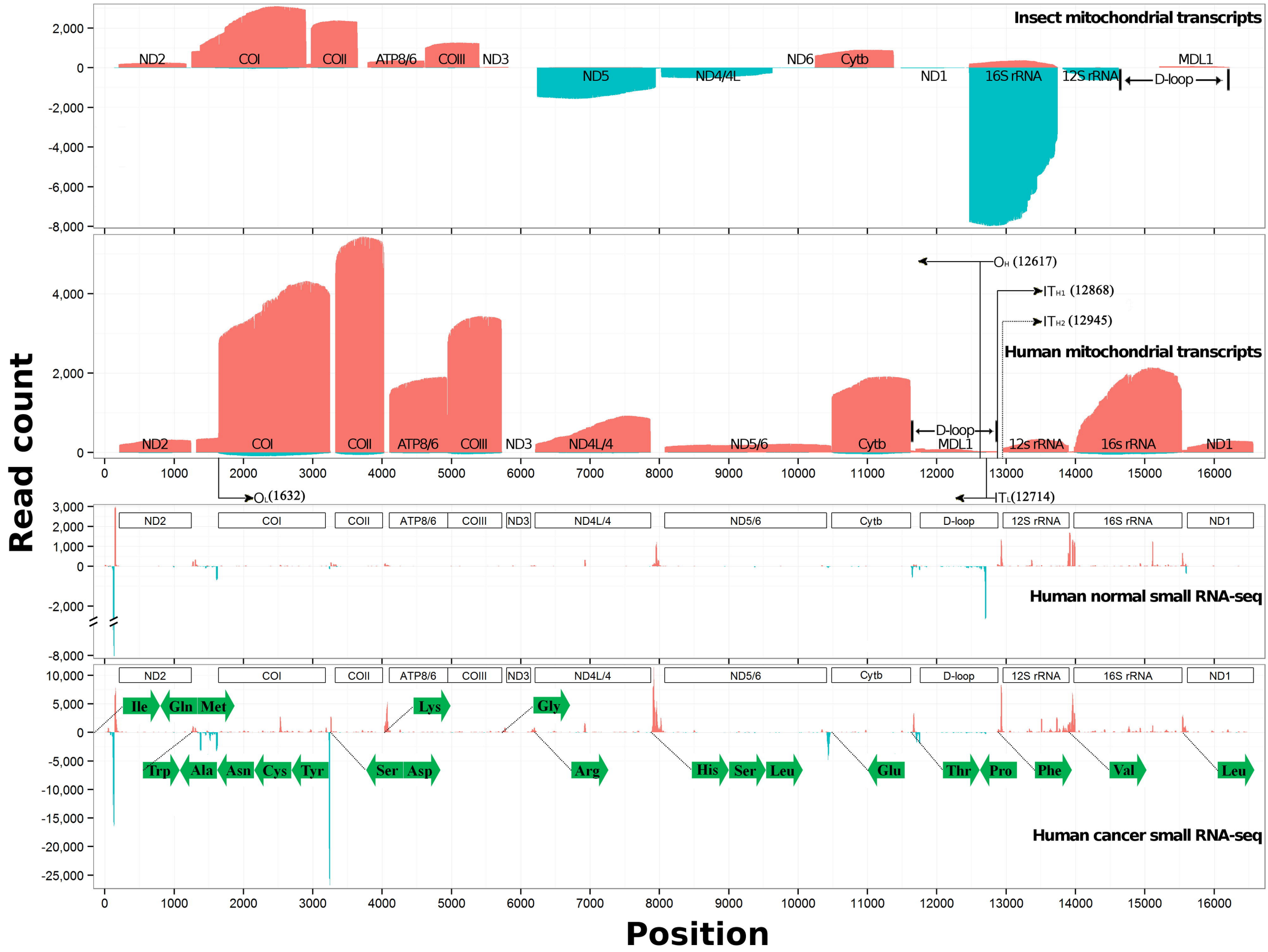
m140810_061610_42161_c100696870310000001823138003261572	5-7 Kbp	14,345	6	0.04%
m140810_093523_42161_c100696870310000001823138003261573	5-7 Kbp	15,631	7	0.04%
m140810_125744_42161_c100696870310000001823138003261574	5-7 Kbp	15,875	4	0.03%
m140810_161642_42161_c100693060150000001823146703241590	5-7 Kbp	11,850	9	0.08%
m140810_193644_42161_c100693060150000001823146703241591	5-7 Kbp	12,570	6	0.05%
m140810_225944_42161_c100693060150000001823146703241592	5-7 Kbp	6,523	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%

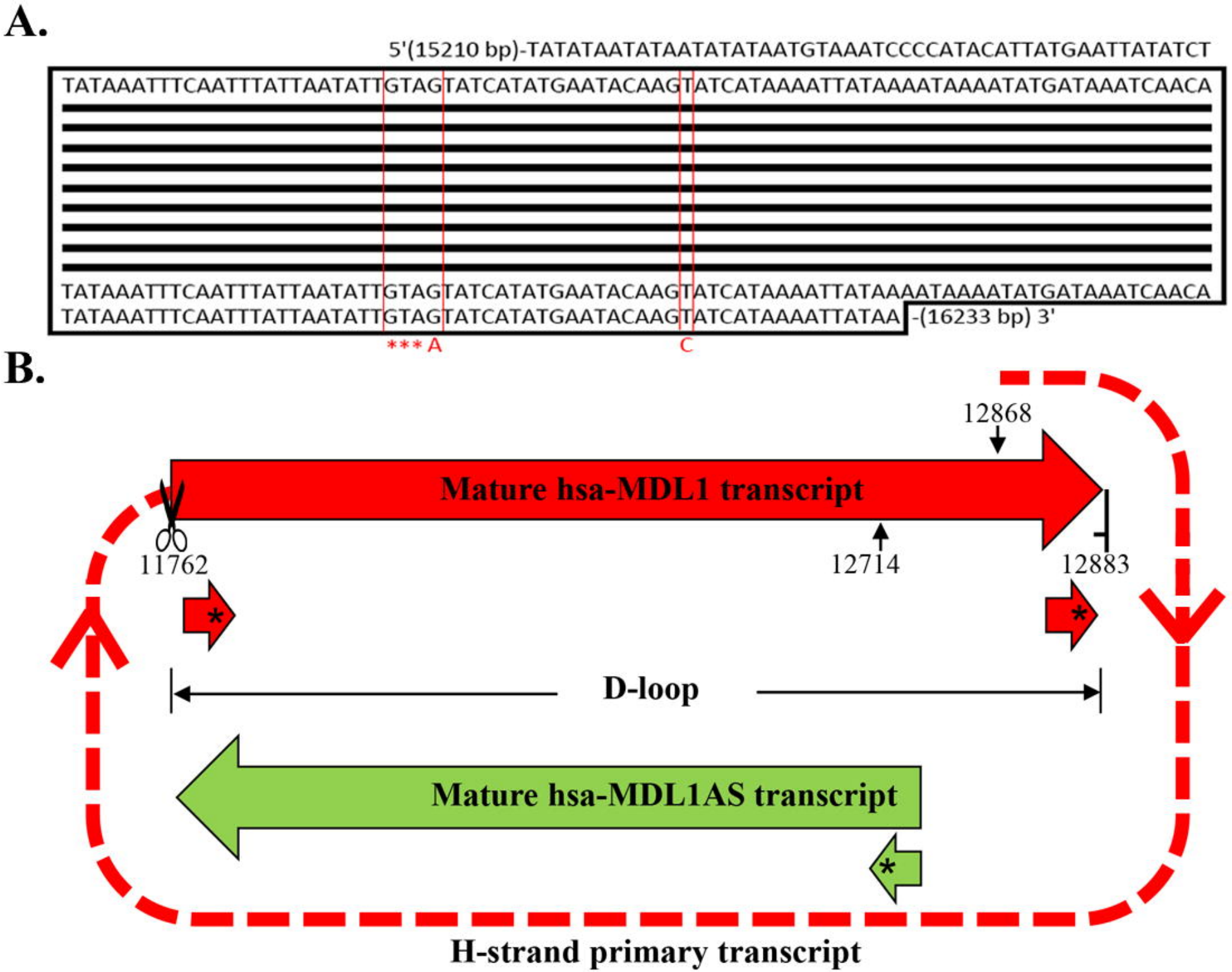
CCS/Draft represents the number of CCS reads and draft transcripts. mtDNA represents the number of draft transcripts mapped to the human mitochondrial genome.

Table 2. Annotation of the human mitochondrial genome

Transcript	Strand	Start	End	Length
ND2	J(+)	208	1,249	1,042
COI	J(+)	1,639	3,183	1,545
COII	J(+)	3,324	4,032	709
ATP8/6	J(+)	4,103	4,945	843
COIII	J(+)	4,945	5,728	784
ND3	J(+)	5,797	6,142	346
ND4L/4	J(+)	6,208	7,875	1,668
Cytb	J(+)	10,485	11,625	1,141
12S rRNA	J(+)	12,955	13,908	954
16S rRNA	J(+)	13,978	15,535	1,558
ND1	J(+)	15,611	16,568	958
Ala/Asn/Cys/Tyr	J(+)	1,319	1,639	321
Ala/Asn/Cys/Tyr/COI/Ser	J(+)	1,319	3,252	1,934
COI/Ser	J(+)	1,639	3,252	1,614
ATP8/6/COIII	J(+)	4,104	5,728	1,625
ND5/ND6/Glu	J(+)	8,075	10,480	2,406
ND5/ND6/Glu/Cytb	J(+)	8,075	11,625	3,551
Pro/Loop	J(+)	11,694	12,883	1,190
Phe/12SrRNA	J(+)	12,884	13,908	1,025
Val/16SrRNA	J(+)	13,909	15,535	1,627
O_L	N(-)	1,632	-	-
O_H	J(+)	12,617	-	-
I_L	N(-)	12,714	-	-
I_{H1}	J(+)	12,868	-	-
I_{H2}	J(+)	12945*	-	-

J(+) and N(-) represents the major and minor coding strand of the mitochondrial genome, respectively.
* I_{H2} is still not well determined. The tRNA genes were represented using their amino acids.


Insect mitochondrial transcripts
ND2
COI
COII
ATP8/6
COIII
ND3
ND5
ND4/4L
ND6
Cytb
ND1
16S rRNA
12S rRNA
D-loop
MDL1
Human mitochondrial transcripts
OH (12617)
ITH1 (12868)
ITH2 (12945)
ND4L/4
ND5/6
12s rRNA
16s rRNA
OL(1632)
TL(12714)
Human normal small RNA-seq
Human cancer small RNA-seq
Ile
Gln
Met
Lys
Gly
Trp
Ala
Asn
Cys
Tyr
Ser
Asp
Arg
His
Ser
Leu
Glu
Thr
Pro
Phe
Val
Leu
Read count
Position


A.
5'(15210 bp)-TATATAATATAATATATAATGTAAATCCCCATACATTATGAATTATATCT
TATAAATTTCAATTTATTAATATTGTAGTATCATATGAATACAAGTATCATAAAATTATAAAATAAAATATGATAAATCAACA
TATAAATTTCAATTTATTAATATTGTAGTATCATATGAATACAAGTATCATAAAATTATAAAATAAAATATGATAAATCAACA
TATAAATTTCAATTTATTAATATTGTAGTATCATATGAATACAAGTATCATAAAATTATAA -(16233 bp) 3'
*** A
C
B.
12868
Mature hsa-MDL1 transcript
11762
12714
12883
*
*
D-loop
Mature hsa-MDL1AS transcript
*
H-strand primary transcript