CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

The discovery of the double-stranded structure of DNA by Watson and Crick in 1953 has shaped current understanding of genetics. Moreover, the central dogma of molecular biology as proposed by Crick in 1958 has become the general interpretation in understanding the relationship between DNA, RNA, and protein. Subsequently, the field of genetic continuously advancing through the development of sequencing techniques with Sanger's 'chain-termination' or dideoxy method as the first-generation of DNA sequencing (reviewed in (Heather & Chain, 2016)).

The human genome encodes for thousands of proteins essential for human growth and development. The idea of the Human Genome Project (HGP) emerged in the early 1980s with the aim to create a global view of human genomes (International Human Genome Sequencing Consortium, 2001). This initiative is expected to drive the advancement of biomedical research by providing and analyzing the genetic information of the human in a comprehensive and unbiased manner. The completion of HGP in the early 2000s serves as a foundation for the genome-wide protein annotation. Through computational genome sequence analysis, 26,588 protein-encoding transcripts have been identified with substantial evidence from both protein families and domains prediction (Venter et al., 2001). However, validating such information require intensive research and an exceptional amount of resources.

Unlike organisms with small genomes, human genes are composed of short exons separated by long introns, hence creating a signal-to-noise problem (International Human Genome Sequencing Consortium, 2001). Therefore, the availability of cDNA sequences and the concept of orthology provide valuable information for annotating genes function (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). This approach is exclusively reasonable for highly conserved

genes, leaving behind those paralogous genes accountable for speciation, sex determination, and fertilization (International Human Genome Sequencing Consortium, 2001).

The completion of HGP also raises more questions regarding the human genome and its importance. Accordingly, Encyclopedia of DNA Elements (ENCODE), a continuation project is instigated to investigate functional elements of the human genome sequence and annotate all protein-coding genes including other non-coding parts of the genome (The ENCODE Project Consortium, 2011). This project has successfully generated the map of protein-coding regions with only 1.22% of the human genome is classified as exon, and transcribed as mature RNAs before translated as proteins (The ENCODE Project Consortium, 2012).

Different from human, the genome sequence of yeast (*Saccharomyces cerevisiae*) has been available since 1996 (Goffeau et al., 1996). The initial annotation of yeast genome uncovers the presence of 6,275 ORFs encoding proteins longer than 99 amino acids (Goffeau et al., 1996). At the same time, proteins consisting of less than 99 amino acids have also been discovered through genetic or biochemical methods. The term “small open reading frames (sORFs)” is coined to distinguish these small proteins with the evidence of more than 260,000 ORFs are having this discrete feature (Basrai, Hieter, & Boeke, 1997). However, the abundance these sORFs may also include the artefactual ORFs that do not possess any biological functions (previously discussed by (Das et al., 1997; Fickett, 1995)).

The development of computational biology and ribosome profiling have corroborated the growing realization of sORFs. Multiple studies have been able to accentuate the existence of sORF and sORF-encoded peptides (SEPs) in different organisms, including yeast, fruit fly, nematode, zebrafish, mouse, and human (Aspden et al., 2014; Bazzini et al., 2014; Crappé et al., 2015, 2013; Frith et al., 2006; Fritsch et al., 2012; Ingolia et al., 2014; Ingolia, Lareau, & Weissman, 2011; Kessler et al., 2003; Ladoukakis, Pereira, Magny, Eyre-Walker, & Couso, 2011; Mackowiak et al., 2015; B. Vanderperre, Lucier, & Roucou, 2012; Benoît Vanderperre et al., 2013). Furthermore, the non-canonical translation initiation with the start codon other than AUG has been observed in the mining of mammalian SEPs

(Ingolia et al., 2011; Slavoff et al., 2013). This highlights the uncharted biology of genetic information and unprecedented translation of essential endogenous peptides.

Mitochondria, as the powerhouse of the cell, are responsible for generating energy in the form of adenosine triphosphate (ATP). It also serves as the biosynthetic hubs of various biomolecules, including nucleotides, glucose, heme, fatty acids, cholesterols, and amino acids (reviewed in (Spinelli & Haigis, 2018)). Recent studies have characterized two novel mitochondrial peptides encoded by sORF, mitoregulin or MOXI and MIEF1 uORF microprotein. These peptides have been shown to control mitochondrial metabolic activities and mitochondrial ribosomes activity, respectively (Chugunova et al., 2019; Delcourt et al., 2018; Lin et al., 2019; Makarewich et al., 2018; Rathore et al., 2018; Stein et al., 2018). By considering the significant role of mitochondria, it is compelling to explore the existence of mitochondrial SEPs and their biological relevance.

In the direction of identifying novel mitochondrial SEPs, three-different strategies were employed to computationally predict the mitochondrial localization of SEPs. The first line of prediction relied on the enrichment of mitochondrial gene expression signature by Weighted Correlation Network Analysis (WCNA) and Gene Set Enrichment Analysis (GSEA). The second line of prediction emphasized the existence of distinct features of classical mitochondria targeting sequence (MTS), disulfide bonds, signal peptide, and transmembrane domain (TMD). These motifs were predicted by multiple bioinformatics resources, namely TargetP, MitoFates, MitoProt, SignalP 4.1, and SignalP 5. The last line of prediction was based on the empirical evidence for protein expression in mitochondria from mass spectrometric data. By combining this three-pronged approach, 173 plausible mitochondrial SEPs have been collated (unpublished data). Therefore, further experimental validation is necessary to identify the existence of peptides encoded by these predicted transcripts and its sub-mitochondrial localization. Additionally, functional characterization of these peptides will unravel its biological significance in modulating mitochondrial energy metabolism, specifically electron transport chains (ETCs) or oxidative phosphorylation (OXPHOS).

1.2 RESEARCH OBJECTIVES

- To experimentally validate the mitochondrial localization of the predicted mitochondrial SEPs;
- To develop a novel technique for assessing the sub-mitochondrial localization of mitochondrial SEPs;
- To functionally predict the importance of newly identified mitochondrial SEPs in regulating mitochondrial respiration.