Available online at www.sciencedirect.com
ScienceDirect
Genomics 92 (2008) 85–93
GENOMICS
www.elsevier.com/locate/ygeno

Organization and evolution of mitochondrial gene clusters in human

Sunjin Moon [a], Seoae Cho [b], Heebal Kim [a,b,*]

[a] Laboratory of Bioinformatics and Population Genetics, Department of Agricultural Biotechnology, Seoul National University, Seoul 151-742, Korea
[b] Graduate Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea

Received 25 October 2007; accepted 8 January 2008
Available online 17 June 2008

Abstract

Currently, the spatial patterns of mitochondrial genes and how the genomic localization of (pseudo)genes originated from mitochondrial DNA remain largely unexplained. The aim of this study was to elucidate the organization of mitochondrial (pseudo)genes given their evolutionary origin. We used a keyword finding method and a bootstrapping method to estimate parameter values that represent the distribution pattern of mitochondrial genes in the nuclear genome. Almost half of mitochondrial genes showing physical clusters were located in the pericentromeric and subtelomeric regions of the chromosome. Most interestingly, the size of these clusters ranged from 0.085 to 3.2 Mb (average ± SD 1.3 ± 0.73 Mb), which coincides with the size of the evolutionary pocket, or the average size of evolutionary breakpoint regions. Our findings imply that the localization of mitochondrial genes in the human genome is determined independent of adaptation.

Keywords: Endosymbiosis; Evolutionary pocket; Gene cluster; Nuclear-encoded mitochondrial protein; Spatial distribution

The endosymbiosis theory states that mitochondrial gene sequences must be successfully integrated into the nuclear genome to maintain the metabolism of the eukaryotic cell [1]. This is thought to be important for increasing genomic complexity [2–4]. Mitochondrial biogenesis is currently proposed to involve signal cross talk between the nucleus and the mitochondria, leading to the coordinated regulation of gene expression from the nuclear genome rather than the mitochondrial genome. It has been reported that the human genome contains more than 1000 genes encoding nuclear-encoded mitochondrial proteins (NEMPs; [5,6]).

The origin and evolution of human mitochondrial (pseudo) gene sequences have been studied in detail [7–11]. The genomic location of newly retroposed gene copies plays an important role in determining their prevalence in the population's gene pool. Integration events often have deleterious effects, leading to complex diseases that may hinder the fixation of mitochondrial gene copies [12]. In addition, the integration of new genetic material may disrupt proper translation. Thus, relatively few integration events may contribute to important functional roles via the co-option of regulatory elements or through insertion into untranslated regions as "molecular passengers" [13,14].

Recent studies have suggested that genomic location might affect the gene expression of transposed mitochondrial sequences [15,16]. In most cases, transposed mitochondrial sequences lack functional promoters and regulatory elements, and therefore they gradually become nonfunctional pseudogenes and contribute to the accumulation of mutations. Given that the mutation rate is higher in the parental mtDNA than in the relocated genomic copy [17,18], the mitochondrial pseudogenes are regarded as "molecular fossils" that could help to unravel the history and mechanism of mitochondrial integration [19–21].

The evolutionary mechanism of mitochondrial integrations has been well studied in plants [22–24]. Some small metabolic gene clusters are thought to have formed in bacteria and in the yeast genome [25,26]. However, the processes underlying the genomic

Abbreviations: NEMP, nuclear-encoded mitochondrial protein; ppNEMP, processed pseudogene of NEMP; mtDNA, mitochondrial DNA; NHEJ, nonhomologous end-joining; DSB, double-strand breaks.

* Corresponding author. Laboratory of Bioinformatics and Population Genetics, Department of Agricultural Biotechnology, Seoul National University, Seoul 151-742, Korea. Fax: +82 2 883 8812.
E-mail address: heebal@snu.ac.kr (H. Kim).

0888-7543/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.01.004

evolution of these lineages are notably different from those of the human genome. For example, the vigorous activity of mobile elements in the mammalian genome has markedly increased the size of the human genome [27]. A number of particular genomic regions such as HSA19q12–q13.2 have been identified as candidate regions for metabolic syndromes [28–30]. However, there is currently no published comprehensive analysis of mitochondrial gene clusters in the human genome. In addition, thus far, no clear explanation exists for how mitochondrial gene clusters can be defined on a genomic scale. To understand the origin of these clusters, we need to consider a variety of parameters and quantify the impact of this phenomenon on genome organization.

We analyzed the genome-wide distribution of mitochondrial genes and their pseudogenes in the nuclear genome to trace the evolutionary progression from their original state. A probability-based approach for mitochondrial localization was used to find the parameters that provided the best description of the distribution patterns. Genome-wide scans for mitochondrial genes were then performed to identify regions that showed a bias toward higher scores of clustering. Finally, we analyzed spatial distribution in the context of function by calculating the probability of distance among mitochondrial genes. We used this approach to investigate their physical locations in the genome and compare their distributions to those of other known genes. Given our results, we suggest an evolutionary model of genomic distribution by examining the clustering of NEMP genes, which have evolved under the neutral processes of genome rearrangement (gene shuffling) by natural selection.

Results

Characterization of NEMP genes and pseudogene sequences

NEMP sequences were obtained from the MitoProteome database, which provides sequences for nuclear-encoded mitochondrial proteins. Each sequence was extensively annotated with data extracted from external databases [31]. Although the majority of the entries have been experimentally validated, some entries in the database were incomplete. Therefore, for quality assurance, the best hits of the alignment results using the BLASTP and BLASTX similarity search programs were recorded [32]. The final best hit in the curator-reviewed RefSeq was defined as a "reviewed" NEMP sequence. We obtained 668 reviewed NEMP sequences with accession numbers beginning with NP and NM for protein and mRNA, respectively, that met our validity criteria. In addition, 297 of 668 (16.2%) NEMPs were determined to be homologues with fairly stringent selection thresholds to a pool of protein sequences from the genomes of eight α-proteobacterial species that are considered to be precursors of eukaryotic mitochondrial genes [33,34].

To avoid overestimation of the number of processed pseudogenes of NEMPs (ppNEMPs), we modified our previous pseudogene identification method [14,21,35] (see Materials and methods for details) and identified 531 candidate ppNEMPs. These candidate sequences were examined to identify whether any ppNEMPs originated from the duplication of NEMP gene sequences. Among them, 24 candidate ppNEMPs were determined to be duplicated pseudogenes based on exon–intron structures identical to those of the parental NEMP genes. These

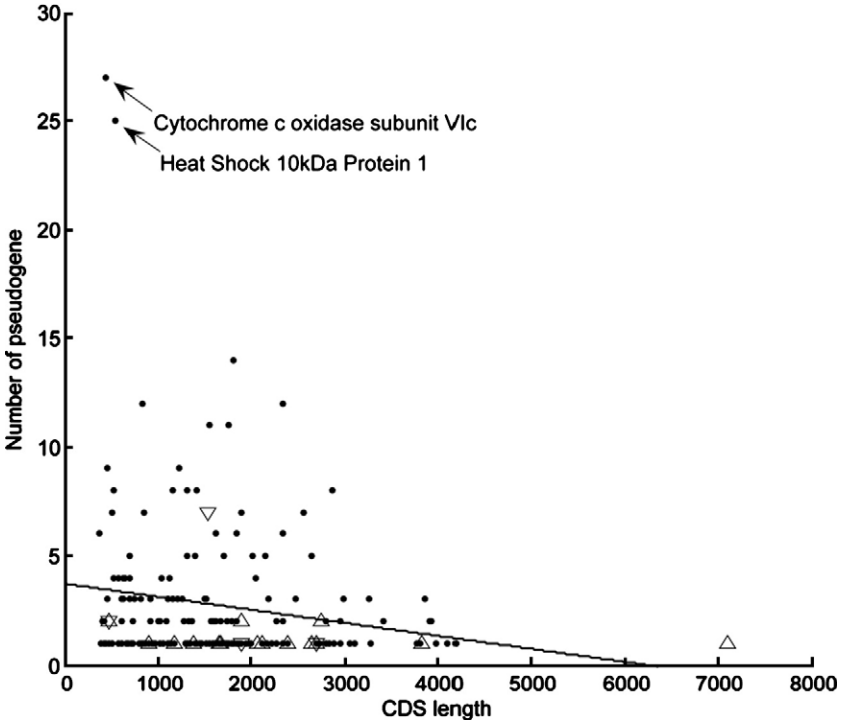

Fig. 1. Relationship between the number of pseudogenes and the protein coding sequence (CDS) length of the parental NEMP gene based on 197 genes having one or more pseudogenes; Spearman's rank correlation, $r=-0.16$; $p=0.02$. Triangles indicate ppNEMPs for highly expressed NEMP genes in germ-line tissues including testis (upward-pointing) and ovary (downward-pointing).