Infection, Genetics and Evolution 32 (2015) 330–337

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Elucidating evolutionary features and functional implications of orphan genes in *Leishmania major*

Sumit Mukherjee [a,b], Arup Panda [a], Tapash Chandra Ghosh [a,*]

[a] Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata 700 054, West Bengal, India
[b] Department of Physical Sciences, Indian Institute of Science Education and Research-Kolkata, Mohanpur 741246, Nadia, West Bengal, India

ARTICLE INFO

Article history:
Received 13 January 2015
Received in revised form 25 March 2015
Accepted 26 March 2015
Available online 2 April 2015

Keywords:
Orphan genes
Evolutionary rate
Protein disorder
Interaction and trafficking motifs
Host–parasite interaction
Lineage-specific adaptation

ABSTRACT

Orphan genes are protein coding genes that lack recognizable homologs in other organisms. These genes were reported to comprise a considerable fraction of coding regions in all sequenced genomes and thought to be allied with organism's lineage-specific traits. However, their evolutionary persistence and functional significance still remain elusive. Due to lack of homologs with the host genome and for their probable lineage-specific functional roles, orphan gene product of pathogenic protozoan might be considered as the possible therapeutic targets. *Leishmania major* is an important parasitic protozoan of the genus *Leishmania* that is associated with the disease cutaneous leishmaniasis. Therefore, evolutionary and functional characterization of orphan genes in this organism may help in understanding the factors prevailing pathogen evolution and parasitic adaptation. In this study, we systematically identified orphan genes of *L. major* and employed several in silico analyses for understanding their evolutionary and functional attributes. To trace the signatures of molecular evolution, we compared their evolutionary rate with non-orphan genes. In agreement with prior observations, here we noticed that orphan genes evolve at a higher rate as compared to non-orphan genes. Lower sequence conservation of orphan genes was previously attributed solely due to their younger gene age. However, here we observed that together with gene age, a number of genomic (like expression level, GC content, variation in codon usage) and proteomic factors (like protein length, intrinsic disorder content, hydropathicity) could independently modulate their evolutionary rate. We considered the interplay of all these factors and analyzed their relative contribution on protein evolutionary rate by regression analysis. On the functional level, we observed that orphan genes are associated with regulatory, growth factor and transport related processes. Moreover, these genes were found to be enriched with various types of interaction and trafficking motifs, implying their possible involvement in host–parasite interactions. Thus, our comprehensive analysis of *L. major* orphan genes provided evidence for their extensive roles in host–pathogen interactions and virulence.

1. Introduction

Orphan genes are protein coding genes that do not share detectable sequence similarity with the genomes of other organisms (Tautz and Domazet-Loso, 2011). Due to their phylogenetic restriction these genes are also called as lineage-specific or taxonomically restricted genes (Wilson et al., 2005). Orphan genes comprise a considerable fraction of genes in all domains of life including viruses (Khalturin et al., 2009; Wilson et al., 2005; Yin and Fischer, 2008). These genes can be broadly classified into two categories (i) taxon-specific orphan genes (TSOGs) that lack homology outside of a focal taxonomic group and (ii) species-specific orphan genes (SSOGs), a subset of SSOGs sharing no homology with any gene in any other species (Wissler et al., 2013). Several hypotheses have been put forward to explain the origin of orphan genes. For instance, gene duplication and rearrangement processes followed by rapid divergence were considered to be an important pathway for the emergence of orphan genes in primates, *Arabidopsis* and zebrafish (Donoghue et al., 2011; Toll-Riera et al., 2009; Yang et al., 2013). In primates, it has been found that the majority of orphan genes arise from frequent recruitment of transposable elements (Toll-Riera et al., 2009). Orphan genes may also arise de

Abbreviations: L. major, *Leishmania major*; BLAST, Basic Local Alignment Search Tool; GRAVY, grand average of hydropathy index; *Nc*, effective number of codon; CAI, Codon Adaptation Index; FPKM, Fragments Per Kilobase of exon per Million fragments mapped.

* Corresponding author. Tel.: +91 33 2355 6626; fax: +91 33 2355 3886.
E-mail address: tapash@jcbose.ac.in (T.C. Ghosh).

http://dx.doi.org/10.1016/j.meegid.2015.03.031
1567-1348/

novo from non-coding regions (Cai et al., 2008; Heinen et al., 2009; Knowles and McLysaght, 2009; Neme and Tautz, 2013; Wu et al., 2011; Xie et al., 2012; Yang and Huang, 2011). These genes were also found to emerge from overlapping of anti-sense reading frames and frameshift mutations in protein coding sequences (Wissler et al., 2013).

Orphan genes are emerging to play critical roles in lineage-specific adaptation of different species to a broad range of ecological conditions (Khalturin et al., 2009). These genes were reported to play substantial roles in response to a variety of abiotic stresses in plant genomes (Donoghue et al., 2011). Imperative roles of orphan genes were also evidenced in several development processes. For instance, orphan gene products were found to be crucial for human early brain development (Zhang et al., 2011) and also for regulation of tentacle formation in Hydra species (Khalturin et al., 2008). Lineage-specific putative surface antigen of *plasmodium* were shown to be involved in host–parasite interactions (Kuo and Kissinger, 2008). In 2010, Zhang et al. ectopically expressed 14 *Leishmania donovani*-specific genes in *Leishmania major* and observed that two of these genes could increase *L. major* survival in visceral organs (Zhang and Matlashewski, 2010).

Studies conducted on different eukaryotes demonstrated that orphan genes evolve faster than non-orphan genes (Cai et al., 2006; Domazet-Loso and Tautz, 2003; Donoghue et al., 2011; Kuo and Kissinger, 2008; Toll-Riera et al., 2009). An inverse relationship between gene age and protein evolutionary rate has been widely observed in a broad range of organisms including primates (Toll-Riera et al., 2009), mammals (Alba and Castresana, 2005), drosophila (Domazet-Loso and Tautz, 2003), *Plasmodium* (Kuo and Kissinger, 2008), fungi (Cai et al., 2006) and bacteria (Daubin and Ochman, 2004). Since, orphan genes are younger genes in a particular lineage it was hypothesized that these genes evolve faster mainly due to their recent evolutionary origin (Cai et al., 2006; Domazet-Loso and Tautz, 2003; Toll-Riera et al., 2009). Later it was found that protein evolutionary rate could not be determined by a single factor, rather protein's intrinsic properties as well as their evolutionary age independently modulate the rates of protein evolution (Toll-Riera et al., 2012). Protein evolutionary rate was shown to correlate with a number of gene level and protein level attributes, such as expression level (Drummond et al., 2005; Drummond et al., 2006; Pal et al., 2001), number of protein–protein interactions (Fraser et al., 2002), protein complex number (Chakraborty and Ghosh, 2013), its centrality in the protein interaction network (Hahn and Kern, 2005), protein dispensability (Hirsh and Fraser, 2001), sequence length (Marais and Duret, 2001), Codon Adaptation Index (CAI), effective number of codons (*Nc*) (Pal et al., 2001; Wall et al., 2005), protein disorder content (Chen et al., 2011; Podder and Ghosh, 2010), *etc*. In spite of all these findings, factors determining the evolutionary rate of orphan genes are still under debate and the relative contribution of different genomic and proteomic attributes on the evolutionary rates of orphan genes remains elusive.

With the availability of high-throughput genomic sequences together with expression data and bioinformatics prediction tools, it has now become easier to identify and characterize orphan genes in different species. *L. major* is one of the most important protozoan parasites of the genus *Leishmania*. It is associated with the disease cutaneous leishmaniasis, affecting more than 2 million people throughout the world every year (Ivens et al., 2005). In spite of multiple research endeavors, till date, there is no available vaccine for this disease. Because of their absence in the host genomes orphan gene products in pathogenic protozoan were considered to be possible therapeutic targets (Kuo and Kissinger, 2008). Therefore, profiling orphan genes of *L. major* from the perception of protein evolutionary rates and comparing them with non-orphan genes along with understanding their functional roles will be helpful to recognize the molecular signature of parasitic adaptation. With this aim we carried out rigorous analysis to understand the functionality of orphan genes and investigated the evolutionary forces affecting orphan gene evolution. To evaluate the attributes of orphan genes in the evolutionary framework we performed a comprehensive analysis comparing orphan genes with the non-orphan genes. In this study our primary objective is to characterize all the possible determinants that may have shaped the evolutionary rate of orphan genes in *L. major*. One of the main obstacles to such a study is the limitation of required data on orphan genes. Therefore, in this study we consider several genomic and proteomic attributes that could be easily identified from coding sequences and analyzed their relative influence on the evolutionary rate heterogeneity between orphan and non-orphan genes.

Confirming earlier observations our study revealed that orphan genes evolve faster than non-orphan genes (Domazet-Loso and Tautz, 2003; Toll-Riera et al., 2009). However, in contrary to the suggestions of those studies, here, we found that gene age could account for a fraction of variation of their evolutionary rate. Instead, together with gene age, a number of factors like gene expression, codon bias, genic GC content, protein hydropathicity, protein disorder content and protein length were found to have substantial contribution on the evolutionary rate difference between orphan and non-orphan genes. On functional level, we found that sequences of orphan genes are endowed with host targeting motifs, prenylation motifs, heparin-binding consensus sequences, signal peptides and transmembrane domains, implying their possible roles in host–parasite interactions. Thus, our study on orphan genes of *L. major* shed light on the factors governing pathogen evolution and reveals their contribution in parasitic adaptations.

2. Materials and methods

2.1. Collection of dataset and gene expression data

We retrieved the protein coding sequences of *L. major* (strain Friedlin) from TriTrypDB version 7.0 (http://tritrypdb.org/tritrypdb/) (Aslett et al., 2010). CDS sequences containing internal stop codons and partial codons were removed using CodonW (http://codonw.sourceforge.net). Signal peptide, transmembrane domain, epitope, paralogs and pathway information's of all *L. major* genes were downloaded from TriTrypDB version 7.0. To compute gene expression level, we retrieved high-throughput RNA-seq expression profile data of *L. major* promastigote stage from the dataset of Rastrojo et al. (2013). We searched for protein domains via InterProScan (Zdobnov and Apweiler, 2001).

2.2. Identification of orphan genes

To identify orphan gene models which are restricted to the *Leishmania* genus, we used a systematic way based on homology search. First, BLASTP followed by TBLASTN filtering approach ($E < 10^{-5}$ and use of low-complexity filters) was used against NCBI nr databases. Additionally, to further screen for similarity between sequences we employed Position-Specific Iterated BLAST (PSI-BLAST) (Altschul et al., 1997) that can detect weaker homologous relationships that would otherwise be missed by the standard BLAST algorithms.

2.3. Calculation of nucleotide substitution rate

The ratio of the rate of non-synonymous substitutions (dN) to the rate of synonymous substitutions (dS) was widely used as an