Insect Biochemistry and Molecular Biology 61 (2015) 25–33

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

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

Genome-wide analysis of homeobox genes from *Mesobuthus martensii* reveals Hox gene duplication in scorpions

Zhiyong Di [a, b], Yao Yu [c], Yingliang Wu [a], Pei Hao [d], Yawen He [a], Huabin Zhao [a], Yixue Li [d], Guoping Zhao [c], Xuan Li [d, *], Wenxin Li [a, *], Zhijian Cao [a, *]

[a] College of Life Sciences, Wuhan University, Wuhan, China
[b] College of Life Sciences, University of Science and Technology of China, Hefei, China
[c] Key Laboratory of Synthetic Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China
[d] Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

ARTICLE INFO

Article history:
Received 25 March 2014
Received in revised form
10 April 2015
Accepted 12 April 2015
Available online 21 April 2015

Keywords:
Homeobox genes
Scorpion
Genome-wide analysis
Hox cluster

ABSTRACT

Homeobox genes belong to a large gene group, which encodes the famous DNA-binding homeodomain that plays a key role in development and cellular differentiation during embryogenesis in animals. Here, one hundred forty-nine homeobox genes were identified from the Asian scorpion, *Mesobuthus martensii* (Chelicerata: Arachnida: Scorpiones: Buthidae) based on our newly assembled genome sequence with approximately 248 × coverage. The identified homeobox genes were categorized into eight classes including 82 families: 67 ANTP class genes, 33 PRD genes, 11 LIM genes, five POU genes, six SINE genes, 14 TALE genes, five CUT genes, two ZF genes and six unclassified genes. Transcriptome data confirmed that more than half of the genes were expressed in adults. The homeobox gene diversity of the eight classes is similar to the previously analyzed Mandibulata arthropods. Interestingly, it is hypothesized that the scorpion *M. martensii* may have two Hox clusters. The first complete genome-wide analysis of homeobox genes in Chelicerata not only reveals the repertoire of scorpion, arachnid and chelicerate homeobox genes, but also shows some insights into the evolution of arthropod homeobox genes.

1. Introduction

Homeobox genes were first discovered in the fruit fly (McGinnis et al., 1984; Scott and Weiner, 1984). They belong to a large gene group which includes 11 classes (Holland et al., 2007). Homeobox genes encode a highly conserved homeodomain. A homeobox is usually 180 base pairs (bp) long and encodes 60 amino acids that bind DNA (Bürglin, 1994, 2005). Homeobox genes play a key role in development and cellular differentiation during embryogenesis in simple and complex organisms.

The Hox cluster is a group of clustered homeobox genes, and also usually named the Hox genes that play important roles in pattern formation along the anterior-posterior body axis. They are assembled in two groups: the Bithorax complex (BX-C) and the Antennapedia complex (ANT-C) in *Drosophila melanogaster* (see the review in Heffer and Pick (2013)).

Recently, an increasing number of genomes of different species have been sequenced, which provides the data for analyzing homeobox genes on a genome-wide level. Genome-wide analysis of homeobox genes has been completed for several animals and indicates that 255 (plus 78 pseudogenes or unassigned) homeobox genes were identified in human, 279 (plus 45 pseudogenes) in mouse, 133 in amphioxus, and 92 in nematode (http://homeodb.zoo.ox.ac.uk/).

In the phylum Arthropoda, the genome-wide analyses of the homeobox genes of several model insects have recently been completed. In the Mandibulata lineage of arthropods, the fruit fly *D. melanogaster* has 104, the beetle *Tribolium castaneum* has 105, and the honeybee *Apis mellifera* has 93 homeobox genes (http://homeodb.zoo.ox.ac.uk/). The silkworm *Bombyx mori* was reported to have 102 homeobox genes and was found to have a special group of 12 tandemly duplicated homeobox genes located between Bmpb and Bmzen, indicating that this Hox cluster had experienced a

* Corresponding authors. Tel.: +86 2768756746.
E-mail addresses: lixuan@sibs.ac.cn (X. Li), liwxlab@whu.edu.cn (W. Li), zjcao@whu.edu.cn (Z. Cao).

http://dx.doi.org/10.1016/j.ibmb.2015.04.002
0965-1748/

lineage-specific expansion in silkworms (Chai et al., 2008). Only two genome sequences have been determined in Chelicerates, *Tetranychus urticae* (2011) (Grbić et al., 2011) and *Ixodes scapularis* (2010, unpublished). Furthermore, a partial survey of homeobox genes was performed in *T. urticae*. The spider mite *T. urticae* was found to contain 8 of the canonical 10 Hox genes. Among them, the *ftz* gene is present in duplicate in two closely linked copies, but orthologues of Hox3 and abdominal A (*abdA*) were not found (Grbić et al., 2011). The morphological evolution of the segment loss in mites is speculated to correlate with the loss of a Hox gene (Grbić et al., 2011).

Known as "living fossils", scorpions are one of the oldest arthropods. Scorpiones is a basal lineage of Arachnida, and Arachnida forms a major branch of Chelicerata. Their body is divided into three parts (tagmata): the head (prosoma), the abdomen (mesosoma) and the tail (metasoma). They have eight legs, two remarkable claws (pedipalps), particular pectines, and a narrow segmented tail ending with a venomous telson. The body morphology of mites consists of an anterior prosoma and posterior opisthosoma and is further distinguished by an extremely reduced body plan. The body morphology and structure of *T. urticae* is distinctly simpler than the scorpion. The unique morphological structures of scorpions imply that they can be used as a model for investigating the patterns, evolution and functions of homeobox and Hox genes. Recently, Sharma et al. (2014) reported that the scorpion *Centruroides sculpturatus* has two paralogues of all Hox genes except Hox3, suggesting cluster and/or whole genome duplication in this arachnid order.

In this study, we characterized homeobox genes from the *Mesobuthus martensii* genome. Transcriptome data revealed the expression profiles of the analyzed homeobox genes. In particular, the structure of the Hox cluster was analyzed in detail in *M. martensii*. *M. martensii* was hypothesized to have two Hox clusters (Hox A and Hox B). In total, we identified 149 homeobox genes, which were grouped into eight classes, including 82 families. This is the first complete genome-wide analysis of homeobox genes in Chelicerata. The results not only show that the Chelicerata lineage possibly has several patterns of Hox clusters, but also provide new insight into the structure and evolution of Hox cluster genes in arthropods.

2. Materials and methods

2.1. Identification and classification of homeobox genes

The new version of the scorpion *M. martensii* genome sequence and predicted protein database (version 1.0 gene models: http://lifecenter.sgst.cn/main/en/Scorpion-Suppl/gene-models-v1.0.gff.zip) was used to search for homeobox genes, as in the previously reported methods (Cao et al., 2013). The whole-genome sequencing project and genome assemblies of *M. martensii* are deposited in the GenBank database under BioProject PRJNA171479. The complete homeodomain protein sequences of 10 species, including human (*Homo sapiens*), mouse (*Mus musculus*), chicken (*Gallus gallus*), frog (*Xenopus (Silurana) tropicalis*), zebrafish (*Danio rerio*), amphioxus (*Branchiostoma floridae*), nematode (*Caenorhabditis elegans*), fruit fly (*D. melanogaster*), beetle (*T. castaneum*), and honeybee (*A. mellifera*) were downloaded from HomeoDB (Homeobox Database) (Supplementary file 1) (http://homeodb.zoo.ox.ac.uk/; Zhong and Holland, 2011) and the HGR (National Human Genome Research Institute) (NCBI). Homeodomain sequences from these ten organisms (HomeoDB) were collected and used as queries for BLASTP searches in the predicted protein database of the scorpion *M. martensii* genome. Considering that some homeodomain sequences cannot be classified because of low identity with the reported homeobox family members, we analyzed all blast results with an identity value greater than 30%.

All of the retrieved candidate homeobox genes were further validated using a program in the SMART database to identify whether the protein sequences encoded by the candidate genes contain homeodomains (same method as Chai et al. (2008)). Candidate homeobox genes were further subjected to a detailed manual survey, particularly those with truncated genome annotations and atypical homeoboxes (NCBI, HomeoDB). The classification scheme and nomenclature for the scorpion homeobox genes was primarily based on the published descriptions (Holland et al., 2007). All of the identified homeobox genes encode proteins containing complete or partial homeodomain sequences. The pseudogenes were not analyzed. Furthermore, we performed a scaffold location analysis on homeobox genes in the scorpion *M. martensii* genome to investigate the Hox clusters.

2.2. Sequence alignment and phylogenetic analysis

The 156 homeodomain sequences from 149 homeobox genes resulting from BLASTP analysis of HomeoDB were aligned by ClustalX 1.83. The alignment file was produced by DNAMAN (highlight homology level $\geq$ 33%). A phylogenetic tree of the scorpion homeodomain family was reconstructed using the maximum likelihood (ML) algorithms implemented in Phyml 3.0 with the 100-fold bootstrap test after the best models were estimated by the ProtTest 2.4 server (http://darwin.uvigo.es/software/prottest2_server.html). The tree diagram was generated and edited using FigTree (Version 1.40) and Mega (5.1).

2.3. Gene expression analysis

To determine the cDNA/ESTs available for each identified homeobox gene, a local TBLASTN search was performed on transcriptome data from the scorpion *M. martensii* using the 149 putative scorpion homeobox genes as queries. Those consistent in the reciprocal TBLAST searches (identity value of homeodomain regions > 95%) were considered cDNA/EST evidence.

3. Results and discussion

3.1. Identification and classification of homeobox genes from the scorpion genome

We identified 149 homeobox genes from the scorpion genome, which represents *M. martensii* homeobox genes (Fig. 1, and Supplementary files 2&3). This is the first survey of scorpion Hox genes. All of the identified homeobox genes were distributed on 134 contigs (Supplementary file 4). The gene expression of the identified scorpion homeobox genes was analyzed using whole transcriptome data from the scorpion *M. martensii*. Of the 149 putative scorpion homeobox genes, 121 have EST evidence. More than 80% (121/149) of the putative homeobox genes are transcribed in the adult scorpion (Supplementary file 5).

Using the recently reported classification scheme as a reference (Holland et al., 2007; Chai et al., 2008), we classified 143 of the 149 identified scorpion homeobox genes into eight classes that cover 82 families. The remaining six scorpion homeobox genes, MMa08570, MMa03106, MMa18441, MMa32551, MMa11725 and MMa49726, could not be assigned to a corresponding family based on the current classification scheme. Phylogenetic relationships were analyzed between the scorpion homeobox genes and their homologs from beetle *T. castaneum*, honeybee *A. mellifera*, and human *H. sapiens* genomes. Most of the identified scorpion homeobox genes could form a monophyletic group with their homologs supported