Accelerated evolution of constraint elements for hematophagic adaptation in mosquitoes

Ming-Shan WANG1,2, Adeniyi C. ADEOLA1,2, Yan LI1, Ya-Ping ZHANG1,3,*, Dong-Dong WU1,*

1 State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China
2 University of Chinese Academy of Sciences, Beijing 100049, China
3 Laboratory for Conservation and Utilization of Bio-resources, Yunnan University, Kunming 650091, China

ABSTRACT

Comparative genomics is a powerful approach that comprehensively interprets the genome. Herein, we performed whole genome comparative analysis of 16 Diptera genomes, including four mosquitoes and 12 Drosophilae. We found more than 540,000 constraint elements (CEs) in the Diptera genome, with the majority found in the intergenic, coding and intronic regions. Accelerated elements (AEs) identified in mosquitoes were mostly in the protein-coding regions (>93%), which differs from vertebrates in genomic distribution. Some genes functionally enriched in blood digestion, body temperature regulation and insecticide resistance showed rapid evolution not only in the lineage of the recent common ancestor of mosquitoes (RCAM), but also in some mosquito lineages. This may be associated with lineage-specific traits and/or adaptations in comparison with other insects. Our findings revealed that although universally fast evolution acted on biological systems in RCAM, such as hematophagy, same adaptations also appear to have occurred through distinct degrees of evolution in different mosquito species, enabling them to be successful blood feeders in different environments.

Keywords: Mosquitoes; Constraint elements; Accelerated elements; Hematophagy

INTRODUCTION

Recent advances in DNA sequencing technology and comparative genomic analysis have unlocked many mysteries of genetic variation in phenotype determination and have greatly changed the landscape of biological research (Chen, 2015; Kim et al, 2011; Li et al, 2010). With the application of next-generation sequencing (NGS), whole genome sequencing of organisms of interest has become less time and cost consumptive. Draft genomes of many non-model species have been completed in the past five years. Comparative analysis of related whole genomes is a useful approach in genome interpretation (Chen, 2015; Zeng et al, 2013). This has facilitated the discovery of functionally important genomic elements responsible for specific phenotypes and identification of adaptive genomic imprints that occurred during evolution (Lindblad-Toh et al, 2011).

Mosquitoes are found within the dipteran flies (Culicidae), and feed on vertebrate blood (such as human, bird, rat and livestock). Many mosquito-borne pathogens, such as malaria, lymphatic filariasis and dengue fever, can be transmitted between humans and other vertebrates during blood feeding, causing a large number of deaths worldwide annually, particularly in Africa (Tsuji et al, 1990a). This demonstrates the extent of the threat on global health imposed by mosquitoes. Blood nutrients such as proteins and iron are essential for female mosquitoes to lay eggs and complete the reproductive process, even though there are variations in reproductive strategies between autogenous and anautogenous mosquitoes (Tsuji et al, 1990b). Mosquitoes are capable of ingesting a large volume of blood in excess of their body weight during a single meal (Friend et al, 1965), exposing them to high temperature and osmotic pressure and oxygen toxicity from heme, iron and ROS, which came from the processes of blood digestion and mitochondrial metabolism.

In addition, due to their long divergence time and various habitats, different mosquitoes have evolved variation in phenotype, feeding behavior and host preference. A good understanding of the genome components and their role in mosquito species[0] may help determine the mechanisms of
specific traits (i.e., hematophagy), and provide a potential starting point to ascertain the driving forces of disease transmission and develop measures to effectively control mosquito-borne diseases.

Constraint elements (CEs) are a group of genomic regions that evolved under restriction across different species over a long period of evolutionary time, and have been widely studied in vertebrate, insect, worm and yeast genomes (Lindblad-Toh et al., 2011; Siepel et al., 2005; Tian et al., 2009; Woolfe et al., 2005). Constraint in a genome sequence implies that it is conserved and plays an important role in biological function. Recent and rapid evolution in a group of CEs in a lineage (hereafter called accelerated elements (AEs)) implies that adaptive evolution may have occurred (Amemiya et al., 2013; Holloway et al., 2008; Lindblad-Toh et al., 2011). Comprehensive investigation of CEs and AEs at the genomic-wide scale has emerged as a useful research tool in evolutionary study and has provided directions to test hypotheses and design experiments (Amemiya et al., 2013; Lindblad-Toh et al., 2011). With the development of whole genome sequencing technologies, the genomes of four blood-feeding insects (Aedes aegypti (Ag) (Holt et al., 2002) and Culex quinquefasciatus (Cq) (Arensburger et al., 2010), along with 12 previously published (Clark et al., 2007) Drosophila genomes (Drosophila melanogaster, D. pseudoobscura, D. ananassae, D. erecta, D. grimshawi, D. mojavensis, D. persimilis, D. sechellia, D. simulans, D. virilis, D. willistoni, and D. yakuba) were retrieved from the UCSC Genome Brower and the Ensembl Metazoa database (version 20). Whole genome comparative analysis of these genomes provides the opportunity to comprehensively interpret the genomic evolution of Diptera and the molecular mechanism of hematophagy adaptation in mosquitoes. We carried out a comprehensive search for CEs through genome-wide multiple alignments of the Diptera genomes (four mosquitoes and 12 Drosophilae) to predict the AE divergence of constraint genomic regions in specific mosquito lineages.

MATERIALS AND METHODS

Whole genome alignments
Genomes of 12 Drosophila, including D. melanogaster (dm3), D. pseudoobscura (dp4), D. ananassae (droAna3), D. erecta (droEre2), D. grimshawi (droGri2), D. mojavensis (droMoj3), D. persimilis (droPer1), D. sechellia (droSec1), D. simulans (droSim1), D. virilis (droVir3), D. willistoni (droWil1), and D. yakuba (droYak2), and four mosquito species (Aedes aegypti, Anopheles darling, Anopheles gambia and Culex quinquefasciatus) were download from the UCSC Genome Brower and Ensembl Metazoa database (version 20). With the exception of Aedes aegypti, Anopheles darling, Anopheles gambia, Culex quinquefasciatus and D. melanogaster, reference base pair wise alignments for all species were download from the UCSC Genome Brower. Using the D. melanogaster (BDGP assembly release 5) genome as a reference, pair wise alignments were performed using the LASTZ (http://www.bx.psu.edu/miller_lab/) program with parameters set to “E=30 O=400 K=3000 L=2200 M=50 --format=axt”. To reduce single coverage with respect to the reference genome and obtain long-linked and longer syntenic contiguous alignments, the alignment blocks were passed through “chaining” and “netting” processes using axtChain, ChainNet and netSyntenic from UCSC tools, as described previously (Kent et al., 2003). Sixteen D. melanogaster-centric whole genome alignments were generated using MULTIZ (Blanchette et al., 2004) v012109/roast and graded by the phylogenic topology tree in Figure 1A.

Evolutionary constraint elements detection
To measure conservation tracks of alignment, four-fold degenerate sites were extracted from chromosomes (2R, 2L, 3R, 3L) of MUTIZ alignments using msa_view software from the PHAST (Siepel et al., 2005) package based on a gff format gene annotation file, and a neutral model (non-conserved model) was then generated using PhyloFit. Conservation scores and conserved sequences were predicted using PhastCons. Running parameters were set with an average conserved sequence length of 45 bp (“--expected-length 45”), target coverage of input alignments was set to 0.3 (“--target-coverage 0.3”), and scaling factor for non-conserved model was set to 0.3 (“--r0 0.3”). Elements with length <30 bp were discarded. The non-conserved phylogenies were estimated by PhyloFit using four-fold degenerate sites, as shown in Figure 1A.

Lineage-specific accelerated elements identification
Lineage-specific accelerated elements were identified by first defining candidate genomic regions across all Diptera, excluding the lineage of interest. Conserved regions were identified using the PhastCons program, as mentioned above, and the pipeline parameters were set to “--expected-length 45 --target-coverage 0.3 --rho 0.3”. [0] Lineage-specific accelerated elements were identified by likelihood ratio tests (Holloway et al., 2008) implemented using PhyloP software from the PHAST package by comparing conserved region substitution rate scores between lineages of interest and the rest of the subtree. A neutral model for inference was created by the above mentioned methods using PhyloFit. PhyloP was run with the scoring system set to “CONACC”. Lineages used in accelerated element detection (lineage of interest) included Aedes aegypti, Anopheles darling, Anopheles gambia, Culex quinquefasciatus and the common ancestor of mosquitoes (RCAM). The element with “alt_subscale >1” showed the region that evolved faster than the remaining part of the tree, thereby providing evidence of acceleration. A P-value of <0.01 was used as the cutoff (FDR-adjusted).

Element genomic location and functional category enriched analysis
All CEs and AEs were annotated with D. melanogaster ENSEMBLE gene annotations using ANNOVAR (Wang et al., 2010), and the genomic regions were classified as down- and up-stream, exonic, intergenic, untranslated and intronic regions. Functional enrichment and pathway annotation clustering were performed using the DAVID (Database for Annotation Visualization and Integrated Discovery) tool (Huang et al., 2008).
RESULTS

Evolutionary constraint elements (CEs) in Diptera
Genome-wide CEs in Diptera (four mosquitoes and 12 Drosophila) were retrieved using PhastCons (see Materials and Methods). A total of 547,726 CEs with length ranging from 30 bp to 6,500 bp (mean size 125.6 bp) were identified (Figure 1B), covering 40.8% of the D. melanogaster genome. About 38% of CEs were located in the protein-coding gene regions, which included 23.10% in the coding region, 8.27% in the untranslated regions (3'UTR and 5'UTRs) and 6.80% within 1 kb of the upstream and downstream regions. A large proportion of CEs were found in the intronic and intergenic regions (39.30% and 21.47%, respectively) (Figure 1C). The distribution pattern in the dipteran genome was similar to that of vertebrate genomes, indicating the potentially important role of non-coding regions in both vertebrates and dipterans. However, the proportion of the dipteran genome covered by CEs was higher than that in mammals (~4.5% of mammal genome) (Lindblad-Toh et al., 2011), which may relate to the different properties of the insect genome compared with the mammal genome, such as higher gene density and more compact gene distribution (Bird et al., 2007; Holloway et al., 2008; Lindblad-Toh et al., 2011). The constraint element results in our study were in accordance with previous identification in the four insects (Anopheles gambia, D. melanogaster, D. yakuba and D. pseudoobscura), except for a slight difference in the proportion of constraint element composition, which may be due to an increased number of higher quality genome sequences used in whole genome alignment (Siepel et al., 2005).

Accelerated conserved elements in mosquitoes
Identification of lineage-specific accelerated elements by likelihood ratio tests
We carried out a series of likelihood ratio tests to retrieve elements showing rapid evolution of the constraint regions in each mosquito lineage using PhyloP software. Constraint genomic regions used for acceleration detection were obtained using the same method as described above by excluding the lineage of interest across all 16 dipteran genome alignments (see Materials and Methods). We retrieved elements showing a divergence of constraint regions in RCAM, Aa, Ad, Ag and Cq lineages, respectively.

In total, 1,463, 199, 954, 660 and 779 AEs in RCAM, Aa, Ad, Ag and Cq lineages were identified, respectively (Figure 2A). These AEs showed a similar length distribution pattern (Figure 2B). The PhyloP scores for AEs were negatively correlated with their lengths ($P<0.01$) (Figure 2C). The majority of AEs were assigned to the protein-coding genomic region, but a few were found in the non-coding region (Figure 2D). This distribution pattern was similar to the AEs identified in D. melanogaster, but different to human and primate lineages with a large number of
Figure 2 Landscape of CEs, AEs and AEGs detected in mosquitoes
A: Number of CEs and AEGs detected in Aa, Ad, Ag, Cq and RCAM. B: Length distribution of accelerated elements in Aa, Ad, Ag, Cq and RCAM. C: Correlation between Phylop score and length of AEs. D: Distribution of AEs in specific genomic regions. E: Venn diagram shows overlap between AEGs detected in Aa, Ad, Ag, Cq and RCAM.

elements in the non-coding regions (Lindblad-Toh et al, 2011). These findings may be the result of the higher density and compactness of gene distribution in insect genomes compared with that of vertebrate genomes (Bird et al, 2007; Holloway et al, 2008; Lindblad-Toh et al, 2011).

Functional enrichment analysis of AEs
The AEs were annotated with D. melanogaster ENSEMBLE gene annotations using ANNOVAR (Wang et al, 2010), and functional enrichment analysis of protein-codon genes covered by AEs (here after referred to as AEGs) were performed using DAVID, including KEGG pathways and GO categories. We found that these AEGs were enriched in a wide variety of functional categories in different lineages. In total, 957, 126, 586, 483 and 545 AGEs were identified in RCAM, Aa, Ad, Ag and Cq lineages, respectively (Figure 2A). Of these, only 10 AEGs were shared by RCAM and the other lineages (Figure 2E).

We found that AEGs in RCAM showed divergence from CEs, and these AEGs were enriched in various functional categories. However, some were over-presented in metabolic-related GO categories, including iron transport/binding, metal ion transport/binding, lipid metabolic process, iron, metal ion and potassium ion channel activity (P<0.05) (Supplementary Table 1). This finding was not unexpected because compared with Drosophila, mosquitoes have the ability to handle high levels of iron, lipids and heme after ingesting a large volume of blood (Graca-Souza et al, 2006). In addition, some genes participated in exopeptidase, metalloexopeptidase, metallooxynzyme, metalloperoxidase and hydrolase activities (P<0.05) (Supplementary Table 1). For instance, exopeptidases, one of the two major classes of secreted proteases in the midgut, function as carboxypeptidases, while amino peptides degrade polypeptides from the terminal ends into exopeptides (Ise et al, 2009). Therefore, the rapid evolution of these genes in mosquitoes may explain their high efficiency in blood feeding and digestion. Imbibing a large volume of blood from a warm-blooded, vertebrate host not only quickly generates osmotic stress, which requires an efficient excretory system for rapid removal of excess water, but also increases the mosquito’s body temperature as blood enters the gut (Benoit et al, 2011). Interestingly, we found heat shock protein 26 (Hsp26) with rapid evolution in RCAM, which has...
been reported with increased expression when mosquitoes are exposed to high temperature (42°C) (Zhao et al, 2010). Due to the long-term and constant use of insecticides, mosquitoes have shown partial tolerance or resistance to insecticides such as DTT, permethrin and malathion. Certain genes or mutations may have contributed to this drug tolerance or resistance in mosquitoes (You et al, 2013). As expected, we found many genes showing evolutionary acceleration enrichment in detoxification and drug metabolism GO categories and pathways, such as response to toxic substances, response to arsenic-containing substances, and detoxification of arsenic-containing substances (P<0.05) (Supplementary Table 1). Three genes, GSTD1, GSTS1 and GSTD9, belong to the glutathione s-transferase (GST) family of proteins, which are involved in the detoxification of a wide range of xenobiotics and protection from oxidative damage, as well as the intracellular transport of hormones, endogenous metabolites and exogenous chemicals including insecticides (Sanil et al, 2014). Our findings also revealed that CYP4C3, CYP4D8 and CYP4F1, members of the cytochrome P450 monoxygenases (CYPs), showed rapid evolution. These genes play vital roles in insecticide resistance and enhance higher rates of insecticide metabolism (Chandor-Proust et al, 2013).

Via analysis of the AEGs detected in the studied lineages, we discovered many genes showed rapid evolution, which were functionally associated with various ion, ATP, lipid and glucose metabolic processes in Aa, Ad, Ag and Cq (Supplementary Table 2-5), and with hydrolase activity in Ad, Ag and [0]Cq (Supplementary Table 3-5). In Cq, more genes and GO categories related to blood feeding showed rapid evolution compared with the other lineages (Figure 2A), including ion, ATP, lipid hydrolase activity and glucose metabolism categories, exopeptidase, metalloexopeptidase, metallocarboxypeptidase and metallopeptidase activities, but few blood feeding related genes and terms were found in Aa, Ad and Ag (Supplementary Table 2-5). More genes showing CE acceleration in Cq may help explain how these diverse blood feeders effectively and successfully digest blood from multiple hosts, including birds, humans and livestock. In Ag, Ad and Cq, we also found several genes enriched in the functional categories of xenobiotic transporter activity, drug transporter activity and xenobiotic-transporting ATPase activity, and most were members of CYPs and GSTs, such as GSTD4, GSTD9, CYP4D8, and CYP6A9 (Supplementary Table 3-5).

**DISCUSSION**

We identified more than 540 000 genomic regions (>30 bp) that evolved under constraint over several million years in Diptera flies by comparative analysis of multiple alignments of 12 Drosophilae and four mosquito genomes. Most CEs were categorized into intergenic, intronic and exonic regions, similar in proportion with that of vertebrates and previous reports (Holloway et al, 2008; Siepel et al, 2005). A higher fraction of CEs (covering 40.8% of the *D. melanogaster* genome) were identified in the dipteran genome than that identified in mammals (~4.5% of the mammal genome) (Lindblad-Toh et al, 2011), which may be due to the higher density and compactness of gene distribution in the insect genome compared with that of the vertebrate genome (Bird et al, 2007; Holloway et al, 2008; Lindblad-Toh et al, 2011).

Being conserved in a genome sequence implies constraint in function. Genomic elements in a lineage that have been under constraint over long periods of evolutionary time, but which have experienced recent and rapid evolutionary expansion, suggest adaptive evolution in those lineages (Holloway et al, 2008). We identified AEs in RCAM, Aa, Ad, Ag and Cq lineages, respectively. Contrary to CEs, the majority of AEs were found in the protein-coding genomic region, with a small proportion located in the non-coding region. This distribution pattern is similar to the AEs identified in the *D. melanogaster* lineage reported previously (Holloway et al, 2008), but different from human and primate lineages in which a large number of elements are assigned into the non-coding regions (Lindblad-Toh et al, 2011). Due to the larger genome size and general biological complexity of vertebrates compared with insects, greater fractions of AEs are found in the non-coding genomic regions in vertebrates. This is likely a reflection of the important role of other non-coding sequences (i.e., IncRNA and circ-RNA) in eukaryotes, which serve as important regulatory factors to regulate gene expression (Lindblad-Toh et al, 2011; Siepel et al, 2005), as verified in recent studies (Iyer et al, 2015; Zhang et al, 2013).

Compared with Drosophila, blood feeding is an intrinsic characteristic of mosquitoes and nutrients in blood are essential for completing reproductive processes. Mosquitoes suffer osmotic and temperature stress after ingesting a large volume of blood, but can deal with it effectively. Comparison of the mosquito genome with that of Drosophila may provide insight into the factors responsible for the above traits. As expected, we discovered some genes related to the transportation and binding of iron, lipids and heme and the digestion of blood essential exopeptidase. We also found hydrolase-related genes, which showed divergence from the CEs in the mosquito lineages and which may contribute to the evolution of blood feeding habits after the divergence from Drosophilae. [0] In addition, some of the detoxification and insecticide resistance-related genes discovered (i.e., glutathione s-transferases (GSTs) and cytochrome P450s) showed rapid evolution in mosquitoes, which may be an adaptive change due to the constant use of chemical insecticides. Genes related to blood feeding adaptation and insecticide resistance also showed rapid evolution in RCAM. While this further evolved in certain mosquitoes (Ag, Aa, Ad and Cq lineages), few genes were shared by RCAM and the other lineages (Figure 2E). This implies that mosquitoes probably possess an independent mechanism and/or different ability to adapt to blood meals or drug resistance. Earlier evidence showed that mosquitoes exhibit differences in host preference (Takken & Verhulst, 2013). For instance, Cq is a diverse blood feeder that can effectively and successfully digest blood from multiple hosts, including birds, humans and livestock. In Cq, we discovered more genes that displayed rapid evolution (Figure 2) and more GO categories related to blood feeding adaptation than any other
lineage. This finding indicates that although the same biological systems were universally fast-evolving, adaptations occurred through somewhat distinct degrees of evolution in different mosquito clades (Lindblad-Toh et al., 2011).

Our study had a number of limitations. First, some of the genomes used in our comparison are recently drafted genomes with potential sequencing and assembly errors, which may lead to bias. In addition, multiple-species, whole genome comparisons are a complex and tedious task, and might led to alignment errors (incorrect alignment blocks or alignment failures) resulting from the program used. Conserved elements (6000): 86-88.

REFERENCES


Zoological Research 36(6): 320-327, 2015 325


