A Profound Role for the Expansion of Trypsin-Like Serine Protease Family in the Evolution of Hematophagy in Mosquito

Dong-Dong Wu,† Guo-Dong Wang,*† David M Irwin,‡§ and Ya-Ping Zhang*†‖*

*State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China; †Graduate School of the Chinese Academy of Sciences, Beijing, China; ‡Department of Laboratory Medicine and Pathobiology, University of Toronto, Ontario, Canada; §Banting and Best Diabetes Centre, University of Toronto, Ontario, Canada; and ‖Laboratory for Conservation and Utilization of Bio-resource, Yunnan University, Kunming, China

The trypsin-like serine protease (Tryp_SPc) family is ubiquitous in animals and plays diverse roles, especially in the digestive system, in different phyla. In the mosquito, some Tryp_SPc proteases make important contributions to the digestion of the blood meal. Here, we have defined the complete Tryp_SPc gene repertoire in the genome of the malaria mosquito, a repertoire that has expanded remarkably compared with that of *Drosophila*. Phylogenetic analysis also indicates that the large-scale lineage-specific expansion occurred leading to mosquitoes. Expression of Tryp_SPc genes elevates after a blood meal, and the expression level of genes that belong to subfamilies that specifically expanded on the mosquito lineage increased significantly more than genes that belong to subfamilies that did not expand in number, suggesting a profound role for the Tryp_SPc genes, especially the expanded subfamilies, in the hematophagous trait of the mosquito. The mosquito Tryp_SPc genes are mostly distributed in a tandem manner on chromosomes, suggesting a role for tandem duplication in the expansion of the subfamilies. Furthermore, evidence for positive selection was found for some genes. Structural modeling indicates that the positively selected sites locate to the surface that is conjugated by protein inhibitors. Our results suggest that the expansion and diversification of the Tryp_SPc domain family in mosquito was driven by positive selection and helps to explain the adaptive hematophagy of the mosquito.

Introduction

Mosquito is a scourge of humanity as the vector for parasitic and viral diseases such as malaria, filariasis, dengue, and yellow fever. Given the serious impact on human health, the genomes of three mosquito species, *Anopheles gambiae* (malaria mosquito), *Culex quinquefasciatus* (Southern house mosquito), and *Aedes aegypti* (yellow fever mosquito), have been sequenced to aid in research to design countermeasures to these organisms (Holt et al. 2002; Nene et al. 2007). A blood meal is essential for the female mosquito to complete each gonotrophic cycle, successfully lay eggs, and synthesize yolk proteins (Attardo et al. 2005); accordingly, the identification of genes associated with the adapted hematophagy of the mosquito is pivotal for developing strategies to control these mosquito-borne diseases. Previous comparative genomic studies have highlighted the roles of adaptive, dynamic, and diversifying evolution of genes in these species (Zdobnov et al. 2002; Waterhouse et al. 2007).

Serine proteases play crucial roles in many physiological processes by activating specific proteins by proteolytic cleavage (Rawlings and Barrett 1994). Trypsin-like serine proteases (Tryp_SPc), which contain a Tryp_SPc domain essential for catalysis, are ubiquitous in eukaryotic animals and are responsible for many essential functions, including processes of food digestion, hemostasis, immune defense response, and the nervous system (Wang et al. 2008). The roles of two types of Tryp_SPc containing serine proteases, trypsin and chymotrypsin, in the digestion of blood in the mosquito have been well documented (Dana et al. 2005). The *A. gambiae* genome contains seven trypsin genes that are clustered within an 11-kb region of chromosome 3R, which encode the five characterized functional proteins (Muller et al. 1995). Potentially, trypsin could activate a chitinase, which is encoded by the malaria parasite, facilitating the passage of the parasite through the peritrophic matrix surrounding the parasite-containing blood meal in the mosquito (Shahabuddin and Kaslow 1994). It has been suggested that significant parasite growth retardation should occur in mosquito larvae that are fed trypsin inhibitors (Casu et al. 1994). Three chymotrypsin genes were also identified in the malaria mosquito genome (Shen et al. 2000; Vizioli et al. 2001).

Presumably due to functional specialization forced by selective pressures, the Tryp_SPc family has evolved diversely both in size and in complexity among organisms. For example, proteome-wide studies in *A. gambiae*, *Drosophila melanogaster*, *Homo sapiens*, and *Fugu rubripes*, have identified more than 305, 206, 110, and 125 functional Tryp_SPc genes, respectively (Zdobnov et al. 2002). In contrast, the *Caenorhabditis elegans* proteome contains only ~13 members suggesting that this species has a significant narrower repertoire of targets (Zdobnov et al. 2002). Compared with invertebrates, vertebrates have more complex Tryp_SPc proteins, with genes that contain many additional domains in addition to the Tryp_SPc domain (Patthy 1999). In function, this family of enzymes plays diverse roles among organisms. In invertebrates, serine proteases are reported to have roles in digestion (Dana et al. 2005), hemolymph coagulation (Iwanaga et al. 1998), antimicrobial peptide synthesis (Hoffmann et al. 1997; Levashina et al. 1999), melanin synthesis (Tang et al. 2006), and the activation of a rapidly immune pathways in response to pathogen detection (Gorman and Paskewitz 2001).

The most profound role of the Tryp_SPc is played in the digestive system. To understand the contributions of the Tryp_SPc, particular to hematophagy, we have systematically defined the mosquito's Tryp_SPc gene repertoires and studied their complexity and evolution. The Tryp_SPc repertoire has expanded substantially in mosquito genomes,

Key words: Tryp_SPc, trypsin-like serine protease, gene family, mosquito, hematophagy, positive selection.

E-mail: zhangyp1@263.net.cn.

Mol. Biol. Evol. 26(10):2333–2341. 2009
doi:10.1093/molbev/msp139
Advance Access publication July 3, 2009

and elevated levels of gene expression occurs after a blood meal. Increased Tryp_SPc gene expression was especially evident for genes belonging to the recently expanded subfamilies. Our findings provide insights into the mechanism of the hematophagous trait of mosquito and the potential control of mosquito-borne diseases.

Materials and Methods

Sequence Retrieval, Annotation, and Phylogenetic Construction of Tryp_SPc Domain Family

The annotated protein databases of *A. gambiae* and *D. melanogaster* were searched using PSI-Blast (Altschul et al. 1997) to obtain Tryp_SPc proteins. A series of Tryp_SPc domain sequences were chosen that represent the diversity of the Tryp_SPc sequences for TBlastN (Altschul et al. 1997) searches against the *Anopheles gambiae* and *D. melanogaster* genome sequences with a permissive *E*-value cutoff of 1 to locate DNA hits. Each genomic hit was extended by approximately 2,000 bp upstream and downstream to ensure coverage of the full-length gene and downloaded. Genes within the downloaded sequences were predicted by GenScan (Burge and Karlin 1997) and Genewise (http://www.ebi.ac.uk/Tools/Wise2/index.html). Predicted protein sequence was queried against the Conserved Domain Database (CDD) (Marchler-Bauer et al. 2005) to ensure that they contained a Tryp_SPc domain and blasted against the nonredundant protein database to identify the highest hit sequence, which was then used as a query in the Genewise analysis to extend the nucleotide sequences. Sequences were extended to their start and stop codons. Genes containing premature stop codons or frameshifts within the translation predicted by the Genewise analysis were considered to be pseudogenes. Some annotated genes contain two or more Tryp_SPc domains, thus it is difficult to authenticate whether the two Tryp_SPc domains arrayed in tandem on a chromosome belong to a single gene or to two genes. We noted that the two or three Tryp_SPc domains within a single predicted gene diverge significantly from each other and were not due to a recent tandem duplication event. Most of the genes containing two or three Tryp_SPc domains were clustered in tandem in the chromosomes, with the different domains in a gene evolved in a parallel manner, yielding similar phylogenies for each gene (i.e., the N-terminal and C-terminal domains generated similar phylogenies), which suggested that the multiple domains belonged to single genes. For further analyses, the first Tryp_SPc protein domain sequences in each of these genes predicted to have multi-Tryp_SPc domains were used to place these genes in the phylogenetic tree of Tryp_SPc genes because many of the second and third domains are shorter than the first domain we chose to exclude them from our analysis to maximize the sequence length available for phylogenetic analysis. Genes with a predicted differing domain architecture compared with nearby phylogenetically clustered genes were reassessed and their predictions refined if necessary. Each sequence was checked manually, and adjustments and refinements were needed for some of the annotated sequences. In a similar manner, the Tryp_SPc genes in the other two mosquito genomes, that is, *Culex pipiens quinquefasciatus* and *A. aegypti* (www.vectorbase.org), were obtained.

The domain architecture of each protein sequence was determined based on the CDD (Marchler-Bauer et al. 2005). The Tryp_SPc domain sequences defined by the CDD (Marchler-Bauer et al. 2005) were aligned by ClustalX (Thompson et al. 1997) and the aligned sequences were used to construct a Neighbor-Joining phylogenetic tree using MEGA4.0 (Tamura et al. 2007). Sequences less than ~120 amino acids in length were excluded from the phylogenetic analysis. The reliabilities of the trees were assessed by 1,000 bootstrap replicates.

Expression Level Comparison

Expression data for Tryp_SPc genes in the *A. gambiae* genome were obtained from the angaGEDUCI database (Marinotti et al. 2005; Dissanayake et al. 2006, http://www.angaged.bio.uci.edu/), which provides stage- and tissue-specific microarray analyses of gene expression at different developmental stages and temporal separations following a blood meal (including 3, 24, 48, 72, 96 h, and 15 days after a blood meal) (Marinotti et al. 2005, 2006). Here, we compared the levels of gene expression at 3 h after a blood meal with the levels of expression fed with sugar instead of a blood meal using $\log_2$ transformation of their expression values. We defined "added" value by subtracting the expression level fed with sugar but without blood meal from the expression level after a blood meal. Considering the potential contribution of genes showing decreased expression after a blood meal to hematophagy, we also calculated the absolute values of the change of expression, which was defined as the "changed" value. To assign expression data from probe sets to their corresponding genes, the sequence of each probe was aligned to the Tryp_SPc genes–coding sequences by BlastN (Altschul et al. 1997). Gene expression was used only for those where a perfect one-to-one match, except the "N" in the probe set sequences, was found, which was a set of 220 Tryp_SPc genes. Expression from ~10,000 probe set sequences assigned to genes from the angaGEDUCI database (Marinotti et al. 2005; Dissanayake et al. 2006) were used to scale expression at the genome level.

Positive Selection Detected by Likelihood Ratio Test

Protein sequences of genes of subfamilies that were inferred to have recently expanded were aligned with ClustalW (Chenna et al. 2003) and back-translated into nucleotide-coding sequences. The sequences were edited manually in Jalview (Clamp et al. 2004). Sequences which showed high divergence from the other sequences were excluded from the analysis. Likelihood ratio tests implemented in PAML3.15 (Yang 1997) were employed to detect positive selection, with likelihoods under models M8 and M7 calculated and compared with identify positively selected codons (Nielsen and Yang 1998; Yang 2000).

Protein Modeling and Locating Positively Selected Sites

The protein sequences of AGAP012692 from group I and AGAP005706 from group II (see fig. 1) were submitted to 3D-PSSM (Kelley et al. 2000) (http://www.sbg.bio.ic.ac.uk/~3dpssm/index2.html) for sequence alignment and protein modeling. The best matching protein folds in the PDB database (http://www.rcsb.org/pdb/) were 2STB (E value 4×10^{-5}) and 1BRU (E value 4×10^{-4}), respectively. UCSF Chimera (Pettersen et al. 2004) was used to visualize the structure and locations of the positively selected sites that had been identified by the likelihood ratio tests.

Results

Dramatic Lineage-Specific Expansion of the Tryp_SPc Gene Family in *Anopheles*

We defined the full complement of Tryp_SPc proteins in the *A. gambiae* genome after in-depth retrieval and refinement. In total, 345 putative functional Tryp_SPc genes were identified in the *A. gambiae* genome, a statistically significant ($P < 0.01$ by χ^2 test) 32.7% [(345 − 260)/260] increase over the 260 genes identified in the *D. melanogaster* genome (Zdobnov et al. 2002). To determine whether this was a specific increase in this gene family or a part of a more general increase in gene family size in mosquito compared to *Drosophila*, we conducted a genome-wide comparison of gene family size between the two species, based on the gene family annotation in Ensembl 53 (http://www.ensembl.org/), which conducts a simultaneous analysis of sequence similarities among all genes in both taxa using the MCL algorithm and a Markov clustering algorithm to classify all proteins encoded by the two genomes (Enright et al. 2002). A total of 5,225 gene families were identified as being shared between *Drosophila* and *Anopheles*, and the ratio of *Anopheles* Tryp_SPc family repertoire size to *Drosophila* of 1.33 (345/260) was found to be significantly higher ($P < 10^{-10}$, by one-sample Student's t-test) than the mean ratio value of 1.04 for all gene family sizes between the two genomes and higher than the mean ratio value of 1.11 using from the 516 families that contain at least 2 genes in both genomes ($P < 10^{-10}$, by one-sample Student's t-test). These data suggest that the expansion of the Tryp_SPc gene family in *Anopheles* is consistent with a mosquito-specific adaptation for a requirement for more Tryp_SPc genes, potentially to digest blood meals. Besides intact genes, there are four and eight pseudogenes identified in the *Anopheles* and *Drosophila* genomes, respectively.

To evaluate the pattern of expansion of the Tryp_SPc family and to determine whether gain or loss of genes in specific subfamilies may be responsible for specific functions, we constructed a Neighbor-Joining phylogenetic tree (Saitou and Nei 1987) based on the Tryp_SPc domain protein sequences with pairwise deletion and Poisson correction distance, after excluding four short sequences from *Anopheles* and one from *Drosophila* (fig. 1). The phylogeny illustrates the magnitude of lineage-specific expansion that occurred in mosquito and fly, suggesting possible functional specialization. Intriguingly, three subfamilies (subfamilies were defined as species-specific subgroups of genes from fig. 1) with specific domain configuration, genes that consist of two linked Tryp_SPc domains (two subfamilies) and genes with three linked Tryp_SPc domains (one subfamily), were expanded in mosquito (fig. 1). In addition, two large mosquito-specific clades that contain genes with only one Tryp_SPc domain were identified (group I and II in fig. 1). Genes within these five clades, thus, are likely candidates to contribute to the unique traits of the mosquito, such as hematophagy. A large number of subfamilies were identified as specific to mosquito, but each has relatively few genes (fig. 1). Concerted evolution likely only had a very minor role in shaping the evolution of Tryp_SPc family in mosquito as sequence divergence among paralogs is high, and positive selection was detected (see following text), observations that are inconsistent with a pattern of concerted evolution (Nei and Rooney 2005).

In addition, we obtained the Tryp_SPc genes in two other mosquito genomes, *C. p. quinquefasciatus* and *A. aegypti*, which contain 403 and 380 intact Tryp_SPc genes, respectively. The *A. aegypti* and *C. p. quinquefasciatus* genomes have ~46% and ~ 55%, respectively, more Tryp_SPc genes than *Drosophila*. Although a large change in Tryp_SPc gene family size exists between mosquito and *Drosophila*, the variation in gene family size between mosquito species is much smaller, suggesting that gene family size and function have been largely conserved since the radiation of mosquitoes. Phylogenetic analysis of these Tryp_SPc genes demonstrated that the number of mosquito genes and *Drosophila* Tryp_SPc genes differ significantly, but that only a small number of gene duplications have occurred since the radiation of the mosquito lineages, suggesting that a large repertoire existed in the common ancestor of mosquitoes (supplementary fig. 1, Supplementary Material online). These observations suggest that the expansion of the Tryp_SPc genes was a common adaptive event for mosquitoes, an event that was potentially driven by the acquisition of the hematophagous trait.

Expression of Tryp_SPc Genes in Mosquito Is Elevated After Blood Meal, Particularly Phylogenetically Expanded Genes

Blood sucking should have a substantial influence on the physiological and functional features of the mosquito, which may be revealed through changes in the pattern of gene expression. Here, we found that the expression levels of Tryp_SPc genes are elevated substantially after a blood meal (fig. 2*A*; $P < 0.001$ by pair-sample Student's t-test; $P = 4.40 \times 10^{-6}$ by Wilcoxon signed ranks test). Furthermore, in comparison with genome-wide differences in expression, we still found a significant specific elevation in expression of Tryp_SPc gene family members after a blood meal (fig. 2*B*; $P = 2.19 \times 10^{-18}$ by independent sample Student's t-test; $P = 7.23 \times 10^{-13}$ by Mann–Whitney U test). Because decreased gene expression genes after a blood meal may also contribute to the hematophagous trait, we also computed the absolute values of difference in expression between having a blood meal and being fed sugar (without a blood meal) and found significantly more changes in expression of Tryp_SPc genes than observed

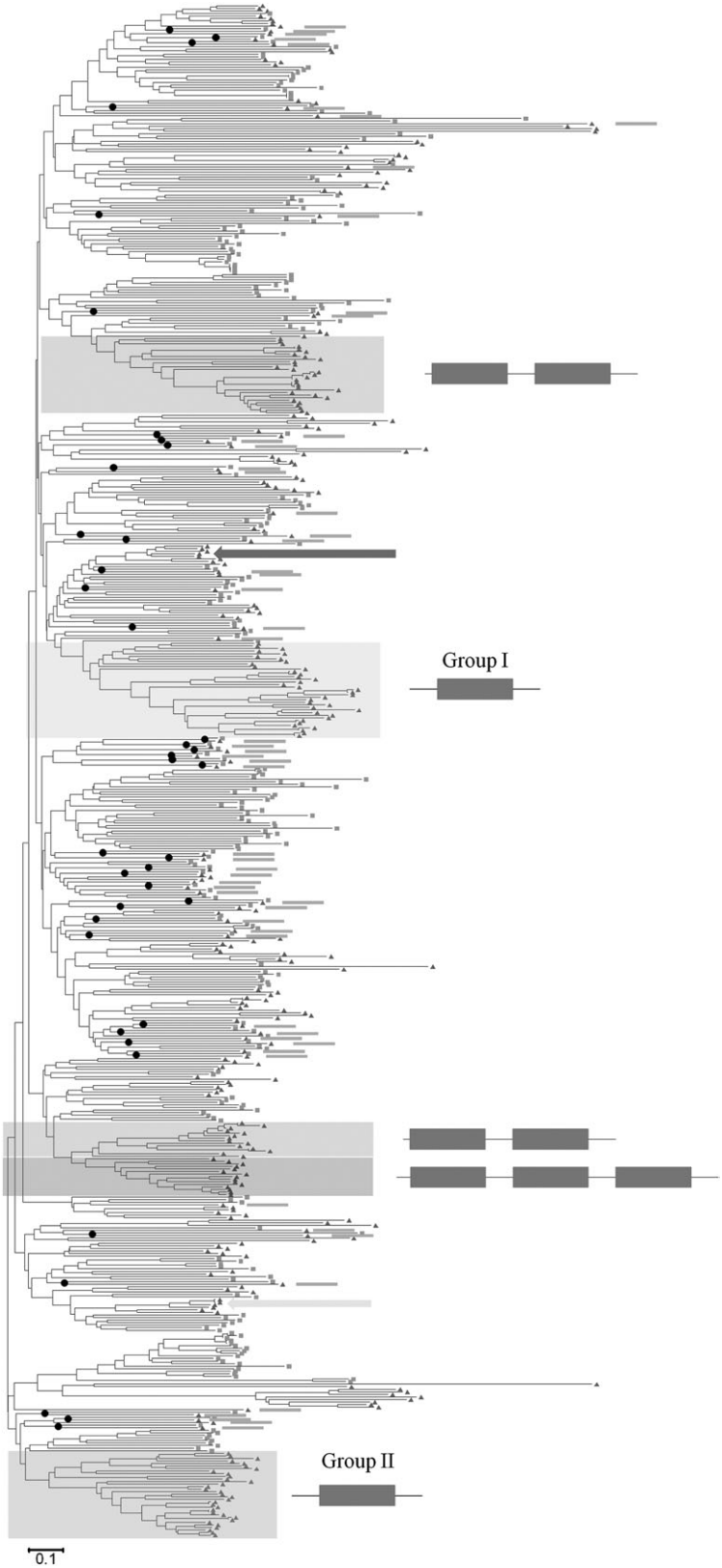

FIG. 1.—Phylogenetic relationship of Tryp_SPc domain family genes in *Anopheles gambiae* and *Drosophila melanogaster*. The genes of *A. gambiae* and *D. melanogaster* are represented by triangles and squares, respectively. The bars to the right of the genes indicate the phylogenetically stable genes in the A. gambiae genome. The box (rectangle) represents the Tryp_SPc domain. Black nodes represent genes with one-to-one orthologs between Anopheles and Drosophila. Several large-scale Anopheles-specific clusters are marked with shadows and their domain architecture is shown to the right. Only the first Tryp_SPc domain sequences of proteins containing two or three Tryp_SPc domains were used for phylogeny construction. The locations of the seven trypsin and four chymotrypsin gene clusters are represented by the arrows, respectively.

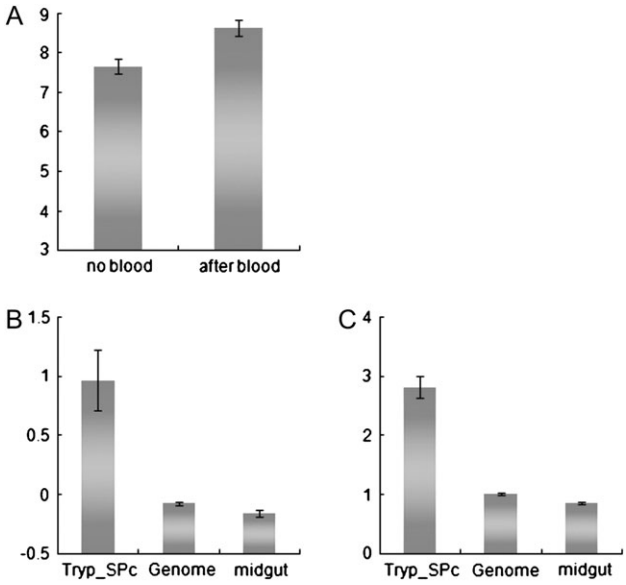

FIG. 2.—Comparison of the expression patterns of Tryp_SPc genes at 3 h after a blood meal and with no blood meal in *Anopheles gambiae*. (*A*) The mean $\log_2$-transformed values (±1 standard error) of the expression level of Tryp_SPc genes 3 h after a blood meal (after blood) and fed with sugar but without a blood meal (no blood) are plotted. (*B*) The mean added values (±1 standard error) of the $\log_2$-transformed expression level after blood meal compared with that fed with sugar but no blood meal of the Tryp_SPc genes (Tryp_SPc) and the genome-wide genes (genome) are shown, respectively, comparing levels of expression at 3 h after a blood meal to no blood meal. The added value is defined by subtracting the expression level without blood meal from the expression level after a blood meal. (*C*) The mean change values (±1 standard error) of the $\log_2$-transformed expression level after blood meal compared with that fed with sugar but no blood meal of the Tryp_SPc genes and the genome-wide genes, respectively, comparing levels of expression at 3 h after a blood meal to no blood meal. The changed value, namely the change of expression level, is defined by the absolute values of added value.

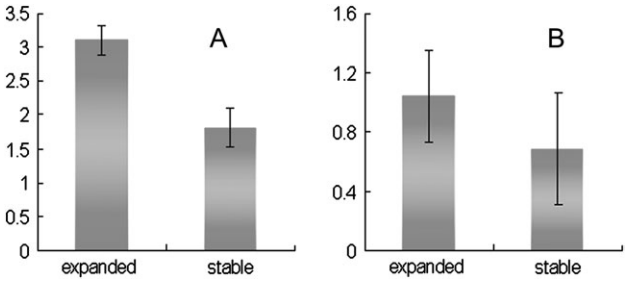

FIG. 3.—Comparison of expression change of Tryp_SPc genes between phylogenetically expanded genes and stable genes in *Anopheles gambiae*. (*A*) The mean change values of the $\log_2$-transformed expression level (±1 standard error) at 3 h after a blood meal compared with that fed with sugar but no blood meal of the phylogenetically expanded genes (expanded) and stable genes (stable), respectively. (*B*) The mean added values of the $\log_2$-transformed expression level (±1 standard error) at 3 h after a blood meal compared with that fed with sugar but no blood meal of the phylogenetically expanded genes and stable genes, respectively.

at the genome-wide level (fig. 2*C*; $P = 6.81 \times 10^{-33}$ by Mann–Whitney *U* test). In addition, we compared the patterns with that of genes expressed primarily in the midgut, a digestive organ, and obtained similar significant results (fig. 2*B* and *C*).

Genes with certain functions tend to be duplicated or retained at higher rates than others. Because expansion of gene families has been associated with new biological functions (Lespinet et al. 2002) and an evolutionary association between phylogenetic stability and function for dynamic selective pressures (Thomas 2006, 2007) exists, we grouped the genes that showed no evidence for new gene duplication on the mosquito lineage after divergence from *Drosophila*, a stable gene group (total of 58 genes) and a group of expanded (i.e., those that show gene duplication on the mosquito lineage) genes (283) that showed expansion based on the phylogenetic tree (fig. 1) and compared the expression patterns between these groups. We found that the expanded genes in mosquito show significantly more change in expression compared with the stable genes (fig. 3*A*; $P < 0.01$ by independent Student's *t*-test; $P = 0.004$ by Kolmogorov–Smirnov test) and that the expression of the expanded genes increased marginally more after blood meal than do stable ones, albeit with no statistical significance (fig. 3*B*; $P = 0.571$ independent sample Student's *t*-test; $P = 0.112$ by Kolmogorov–Smirnov test). Taken together, these findings suggested that the Tryp_SPc genes, particularly the expanded ones, are likely to be involved in the hematophagous trait of mosquito.

We next examined changes in the patterns of expression of each mosquito-specific subfamily of Tryp_SPc that had been identified in the phylogenetic tree (fig. 1). The number of genes within each mosquito-specific subfamily was denoted as the size of each subfamily. Intriguingly, we found a positive correlation between the increase in expression level after a blood meal with subfamily size (fig. 4*A*; $R^2 = 0.328$, $P = 0.032$ by linear regression), which means that the expression of genes within large subfamilies tends to increase more after a blood meal than genes within smaller subfamilies, although no similar correlation was observed between the change in expression value of the subfamily and the subfamily size (fig. 4*B*; $R^2 = 0.130$, $P = 0.205$ by linear regression). This intriguing observation indicated a possible need for more Tryp_SPc genes promoted by the blood sucking for mosquito. Remarkably, expressions of the group I and II genes which showed large-scale expansion in figure 1 increased by approximately 13-fold and 40-fold, respectively, after a blood meal. In contrast, the expressions of the expanded subfamilies of genes that contained two or three Tryp_SPc domains showed only minor changes in expression with a blood meal, suggesting that these three subfamilies (fig. 1) may play minor roles in hematophagy.

Tryp_SPc genes in the mosquito genome tend to be clustered in tandem on chromosomes. Here, we defined Tryp_SPc genes which are within 50 kb of a neighboring Tryp_SPc gene as being in tandem. A total of 250 genes were found in tandem and only 95 genes were dispersed. The recently expanded genes on the mosquito lineage contributed 218 tandemly arranged genes, suggesting that tandem gene duplication played a profound role in the expansion of the Tryp_SPc genes in *Anopheles*. Furthermore, we did a further analysis of relationships between phylogenetic positions and genomic locations of tandem duplicated genes by analyzing the correlation between protein sequence distance and physical distance in the chromosome

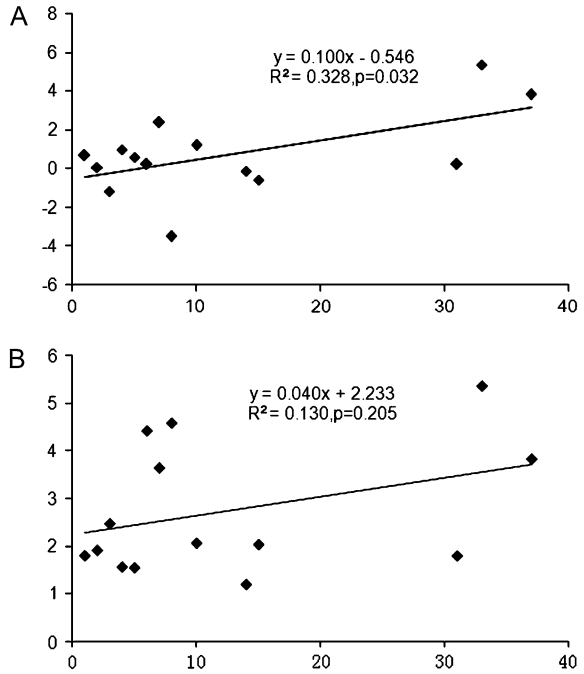

FIG. 4.—The correlation patterns of subfamily sizes with their expression added values (*A*) and change values (*B*). The number of genes within each mosquito-specific subfamily is denoted as the subfamily size of each subfamily. A positive correlation is indicated between the expression added value after a blood meal with the subfamily size (*A*; $R^2 = 0.328$, $P = 0.032$ by linear regression), but no similar correlation is observed between the changed value of the subfamily and the subfamily size (*B*; $R^2 = 0.130$, $P = 0.205$ by linear regression).

and found significant positive correlation between the two parameters ($P = 8.199 \times 10^{-9}$ by Pearson's correlation coefficient. fig. 5), supporting the significant role of tandem duplication. Expression of tandemly arranged genes changed more significantly than that of dispersed genes (fig. 6*A*; $P = 0.042$ by independent sample Student's *t*-test; $P = 0.247$ by Kolmogorov–Smirnov test), and expression of the tandem genes also increased marginally more after a blood meal than that of the dispersed genes, albeit the difference was not statistically significant (fig. 6*B*; $P = 0.351$ by independent sample Student's *t*-test; $P = 0.069$ by Kolmogorov–Smirnov test).

Expanded Genes with Increased Expression after Blood Meal in Mosquito Are Subjected to Positive Selection

Positive selection is a major driving force in the expansion of gene families for particular function, where adaptive evolution occurs frequently. Here, we evaluated the selective pressure in the two mosquito-specific clusters of Tryp_SPc genes, the group I and II identified in the figure 1 which both show expansion and significantly increased expression after a blood meal. As expected, we found evidence for positive selection among the Tryp_SPc genes, suggesting that adaptive evolution was a force driving the evolution of the hematophagous trait to digest extraneous blood (supplementary fig. 3 [Supplementary Material online] and fig. 8), consistent with the common belief that host genes involved in recognizing and interacting with

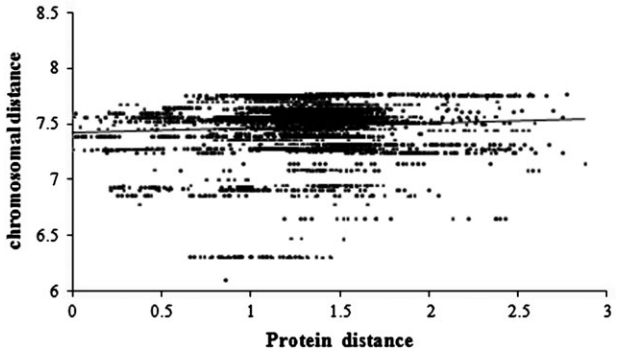

FIG. 5.—The correlation between protein sequence distance and chromosomal distance of each pair of two tandem located genes within the same chromosome. The values of physical distances in chromosome have been transformed by $\log_{10}$. The protein sequence distances were calculated by *p* distance after pairwise deletion.

environmental factors are preferential targets for adaptive diversification (Thomas 2006, 2007).

The role of interaction with an environmental factor was supported by an analysis of the three-dimensional structures and mapping of the positively selected residues to these structures. The 3D-PSSM method was employed to align the Tryp_SPc domain sequences to known protein structures in the PDB database (see Materials and Methods). For the group I genes, the two positively selected amino acid sites were found to be located on the surface of the structure, within the binding site of trypsin proteases inhibitors (fig. 7). These observations illustrate that positive selection of sites was probably crucial for adapting members of the Tryp_SPc family for digesting the blood meals in mosquito. Similarly, positively selected sites identified in the group II genes also map to the surface of the structures (supplementary fig. 4, Supplementary Material online).

Discussion

Great efforts have been made to control diseases spread by mosquitoes (Gubler 2002; Zaim and Guillet 2002). Blood sucking initiates a complex series of physiological events, including changes in gene expression in the

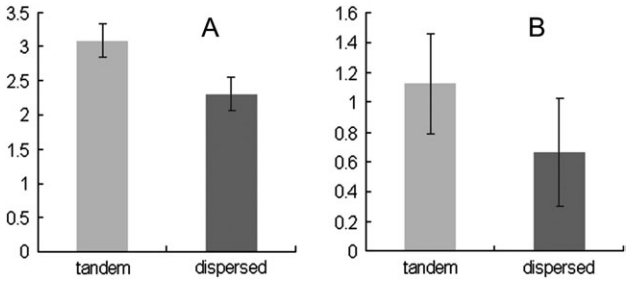

FIG. 6.—Comparison of expression change of Tryp_SPc genes between tandem-located genes and dispersed genes in *Anopheles gambiae*. (*A*) The mean change values of the $\log_2$-transformed expression level (±1 standard error) at 3 h after a blood meal compared with that fed with sugar but no blood meal of tandem genes (tandem) and dispersed genes (disperse), respectively. (*B*) The mean added values of the $\log_2$-transformed expression level (±1 standard error) at 3 h after a blood meal compared with that fed with sugar but no blood meal of tandem genes and dispersed genes, respectively.

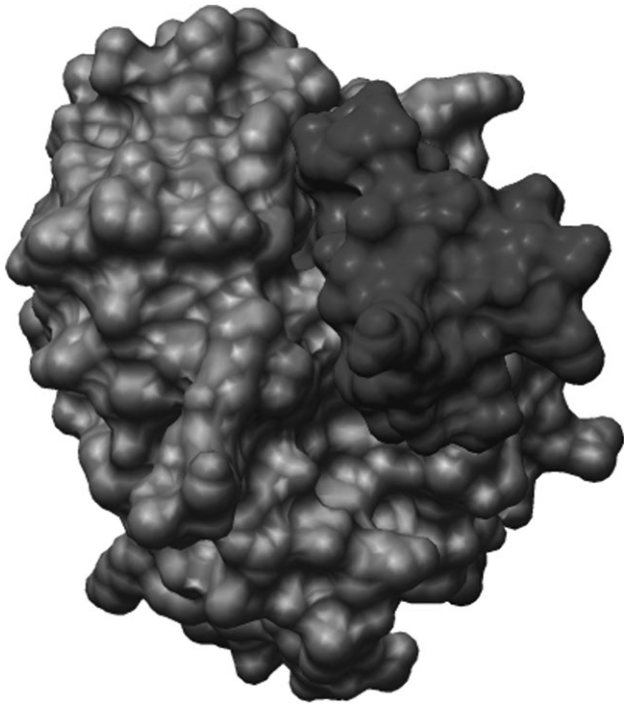

FIG. 7.—Structural modeling of Tryp_SPc domain of group I and positively selected sites. The larger region represents the Tryp_SPc protease. The grey is the Tryp_SPc protease inhibitor. The two small sites are the positively selected sites described in supplementary figure 3.

mosquito (Dana et al. 2005). Exploiting the underlying genes and elements responsible for the changes in expression and corresponding functions, particularly with the advent of the mosquito genome sequences (Holt et al. 2002; Nene et al. 2007), should facilitate research on the process of blood digestion and help to develop strategies to control mosquito-borne diseases.

Phylogenetic analysis of gene and gene families has proved to be powerful tool for molecular biologists, biochemists, and virologists to understand the functions of genes (Yang 2005). Our findings demonstrate a perfect example of the correlation of evolutionary pattern with function of a gene and/or a gene family. The Tryp_SPc gene family is one of the largest gene families in insects, with ~345 putative functional members in the malaria mosquito genome. Gene family members that interact with foreign proteins typically show features such as a high turnover rate and adaptive evolution (Thomas 2006, 2007) and features observed in the mosquito Tryp_SPc gene family. Although there is little direct experimental evidence addressing the function of such large gene families, the patterns of expression and evolution strongly suggest a role of Tryp_SPc genes in hematophagy: 1) the expression of Tryp_SPc genes increases significantly after a blood meal, 2) the expression of the phylogenetically expanded genes change more than the stable genes, and a positive correlation between increasing expression level of a subfamily with the size of the subfamily suggests an adaptive requirement of more Tryp_SPc genes by mosquito for hematophagy, which is supported by the extensive duplications and positive selection–driven diversification in the *Anopheles* genome, and 3) the locations of the positively selected sites in the binding region of the Tryp_SPc protein structure suggests adaptive evolution for the process of digestion of food. Food provides an important source and motive driving the evolution of genetic complexity and change. Evidence of genomic adaptation for digestion of food has been well documented, for example, positive selection in the *AMY1* and *LCT* genes in modern human populations (Bersaglieri et al. 2004; Perry et al. 2007), adaptive origin of the digestive *RNase1B* in leaf-eating monkeys (Zhang et al. 2002; Zhang 2006). Digestion of a blood meal is associated with an environmental interaction as the sources (i.e., blood from other species interacts with the host digestive enzymes) of a mosquitoes blood meals change, and thus, the digestive process must constantly adapt to the new sources of blood meals.

Genes that contain two or three Tryp_SPc domains were identified in the *Anopheles* genome, although they appear to only play a minor role in hematophagy as only minor changes in the expression of these genes were observed after a blood meal. Additional experimental studies are needed to clarify the function of these genes. Approximately 250 of the 345 Tryp_SPc genes contain only a single Tryp_SPc domain. Genes with a single

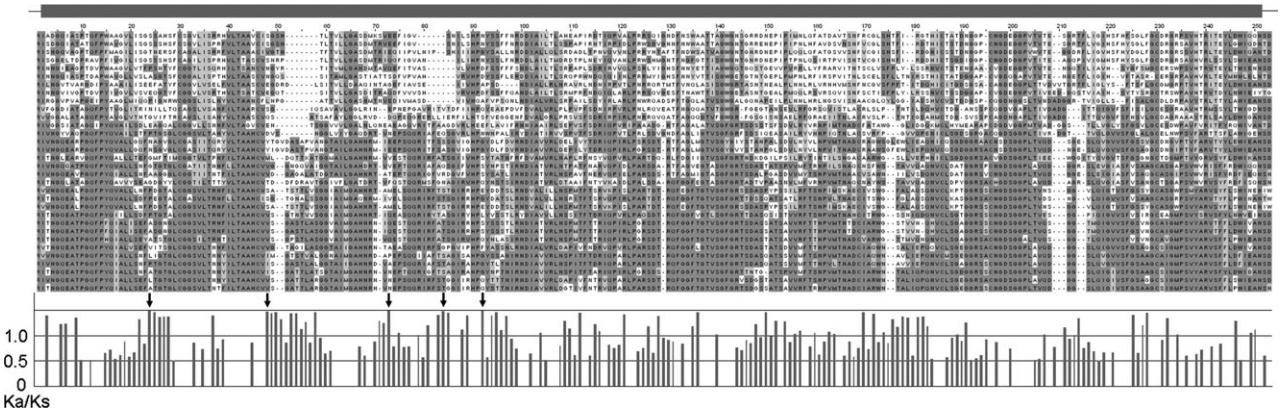

FIG. 8.—Positive selection detection result of genes in group II defined in figure 1. The top panel defines the region of Tryp_SPc domain by a strip. The histogram of low panel shows the estimated *K*a/*K*s value of each codon in the alignment by site-specific model, and the values lower than 0.5 are not shown. Five sites pointed by arrowheads are the potential positive selected sites with posterior probability >95%.

Tryp_SPc structure are responsible for the typical functions of these proteins, and most of the mosquito lineage–specific expansion involved genes of this type of gene. The two largest lineage-specific subfamilies, group I and II (fig. 1), have remarkable increases in gene expression levels after blood meal, for example, gene expression of *AGAP005671* (chr 2L:18530513–18531415) in group II increases approximately 3,223-fold after a blood meal and *AGAP007252* (chr 2L:44661593–44662510) increases approximately 917-fold (summary for all genes is in supplementary tables 1 and 2, Supplementary Material online). In the phylogenetic tree, group I genes are located close to the trypsin genes (fig. 1) suggesting that they may have a function similar to that of trypsin. The group I genes thus likely have a profound contribution to the hematophagous trait of the mosquito. Although previously it has been shown that trypsin and chymotrypsin are important in the digestion of blood after a blood mean in mosquitoes (Muller et al. 1995; Shen et al. 2000; Vizioli et al. 2001), our new genomic analysis indicates that much larger group of Tryp_SPc genes play important and significant roles in this process.

Supplementary Material

Supplementary figures 1–4 and tables 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Yu Jiang for assistance in this research. We also thank Dr Satta Yoko and several anonymous reviewers for their comments and suggestions. This work was supported by grants from the National Basic Research Program of China (973 Program, 2007CB411600), the National Natural Science Foundation of China (30621092, 30430110), and Bureau of Science and Technology of Yunnan Province.

Literature Cited

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res. 25:3389–3402.

Attardo GM, Hansen IA, Raikhel AS. 2005. Nutritional regulation of vitellogenesis in mosquitoes: implications for anautogeny. Insect Biochem Mol Biol. 35:661–675.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet. 74:1111–1120.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 268:78–94.

Casu RE, Jarmey JM, Elvin CM, Eisemann CH. 1994. Isolation of a trypsin-like serine protease gene family from the sheep blowfly Lucilia cuprina. Insect Mol Biol. 3:159–170.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acid Res. 31:3497–3500.

Clamp M, Cuff J, Searle SM, Barton GJ. 2004. The Jalview Java alignment editor. Bioinformatics. 20:426–427.

Dana AN, Hong YS, Kern MK, Hillenmeyer ME, Harker BW, Lobo NF, Hogan JR, Romans P, Collins FH. 2005. Gene expression patterns associated with blood-feeding in the malaria mosquito *Anopheles gambiae*. BMC Genomics. 6:5.

Dissanayake SN, Marinotti O, Ribeiro JMC, James AA. 2006. angaGEDUCI: *Anopheles gambiae* gene expression database with integrated comparative algorithms for identifying conserved DNA motifs in promoter sequences. BMC Genomics. 7:116.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584.

Gorman MJ, Paskewitz SM. 2001. Serine proteases as mediators of mosquito immune responses. Insect Biochem Mol Biol. 31:257–262.

Gubler DJ. 2002. Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. Trends Microbiol. 10:100–103.

Hoffmann JA, Kafatos FC, Janeway CA Jr, Ezekowitz RAB. 1997. Phylogenetic perspectives in innate immunity. Science. 112:190.

Holt RA, Subramanian GM, Halpern A, et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science. 298:129–149.

Iwanaga S, Kawabata S, Muta T. 1998. New types of clotting factors and defense molecules found in horseshoe crab hemolymph: their structures and functions. J Biochem. 123:1–15.

Kelley LA, MacCallum RM, Sternberg MJE. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol. 299:501–522.

Lespinet O, Wolf YI, Koonin EV, Aravind L. 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res. 12:1048–1059.

Levashina EA, Langley E, Green C, Gubb D, Ashburner M, Hoffmann JA, Reichhart JM. 1999. Constitutive activation of toll-mediated antifungal defense in serpin-deficient Drosophila. Science. 285:1917–1919.

Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z. 2005. CDD: a conserved domain database for protein classification. Nucleic Acids Res. 33:D192–D196.

Marinotti O, Calvo E, Nguyen QK, Dissanayake S, Ribeiro JMC, James AA. 2006. Genome-wide analysis of gene expression in adult *Anopheles gambiae*. Insect Mol Biol. 15:1–12.

Marinotti O, Nguyen QK, Calvo E, James AA, Ribeiro JMC. 2005. Microarray analysis of genes showing variable expression following a blood meal in *Anopheles gambiae*. Insect Mol Biol. 14:365–373.

Muller HM, Catteruccia F, Vizioli J, Dellatorre A, Crisanti A. 1995. Constitutive and blood meal-induced trypsin genes in *Anopheles gambiae*. Exp Parasitol. 81:371–385.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 39:121–152.

Nene V, Wortman JR, Lawson D, et al. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science. 316:1718–1723.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 148:929–936.

Patthy L. 1999. Genome evolution and the evolution of exon-shuffling—a review. Gene. 238:103–114.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R. 2007. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 39:1256–1260.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem. 25:1605–1612.

Rawlings ND, Barrett AJ. 1994. Families of serine peptidases. Methods Enzymol. 244:19–61.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.

Shahabuddin M, Kaslow DC. 1994. Plasmodium: parasite chitinase and its role in malaria transmission. Exp Parasitol. 79:85–88.

Shen Z, Edwards MJ, Jacobs-Lorena M. 2000. A gut-specific serine protease from the malaria vector *Anopheles gambiae* is downregulated after blood ingestion. Insect Mol Biol. 9:223–229.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol. 24:1596–1599.

Tang H, Kambris Z, Lemaitre B, Hashimoto C. 2006. Two proteases defining a melanization cascade in the immune system of *Drosophila*. J Biol Chem. 281:28097–28104.

Thomas JH. 2006. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. Genome Res. 16:1017–1030.

Thomas JH. 2007. Rapid birth–death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLos Genet. 3:e67.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acid Res. 25:4876–4882.

Vizioli J, Catteruccia F, Torre Ad, Reckmann I, Müller H-M. 2001. Blood digestion in the malaria mosquito *Anopheles gambiae*: Molecular cloning and biochemical characterization of two inducible chymotrypsins. Eur J Biochem. 268:4027–4035.

Wang Y, Luo W, Reiser G. 2008. Trypsin and trypsin-like proteases in the brain: Proteolysis and cellular functions. Cell Mol Life Sci. 65:237–252.

Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, Barillas-Mury C, Bian G, Blandin S, Christensen BM. 2007. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. Science. 316:1738–1743.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics. 13:555–556.

Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J Mol Evol. 51:423–432.

Yang Z. 2005. The power of phylogenetic comparison in revealing protein function. Proc Natl Acad Sci USA. 102:3179–3180.

Zaim M, Guillet P. 2002. Alternative insecticides: an urgent need. Trends Parasitol. 18:161–163.

Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM. 2002. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. Science. 298:149–159.

Zhang J. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat Genet. 38:819–823.

Zhang J, Zhang Y, Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat Genet. 30:411–415.

Yoko Satta, Associate Editor

Accepted June 29, 2009