Genome-Wide Analysis of Protein Family Diversity Provides Insights into Species-Specific Protein Family Expansions in Insects

Mehmet DAYI (mehmetdayi@duzce.edu.tr)
Duzce University

Research Article

Keywords: Genome, insects, protein diversity, phylogeny

Posted Date: September 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-864773/v1

Abstract

Insects are one of the earliest land animals with more than 400 million years old history on Earth, and they compose more than 80% of species. Insects invade a wide range of ecosystems and are considered one of the most evolutionary successful organism groups. Today, many insect species' genomes have been sequenced to encode molecular mechanisms behind this magnificent evolutionary plasticity. However, only limited genome-wide studies have been carried out to compare protein family diversity in insects. A total of 20 insect species belonging to seven insect orders and two morphogenesis groups were investigated for evolutionary relationships and to uncover protein family diversity in the present study. The phylogenetic analysis inferred from a total of 530 one-to-one single-copy ortholog genes were separated insects into two evolutionary clades based on morphogenesis. Protein family analyses showed that insects share core protein families that perform essential tasks in development and metabolic processes, such as Pkinase and Zinc Finger, cellular signaling and odorant perception (7tm), digestion of food molecules (Trypsin), and detoxification (p450) with copy number expansion compared to other protein families. Additionally, species-specific protein family expansion was observed in various protein families. This study provided insights into protein family diversity and variation among insects and highlights high copy number variation in protein families species-wide.

Introduction

Insects are recognized as the largest animal group that constitutes approximately 80% of species and one of the most diverse organism groups on Earth (Zhang et al. 2007). They are among the earliest land animals, and their existence on earth dates back to at least 400 million years (Zhang et al. 2007). Although approximately one million insect species have been identified, classified, and named to date, their actual number is believed to be in the range between 2.5 and 10 million (Zhang et al. 2007). It is accepted that they have diverged as members of one of the largest subphyla of arthropods nearly 390 million years ago. They have invaded a wide range of habitats during their long evolutionary history and have experienced rapid and successful radiation, and their evolution is recognized as faster than any other group (Gaunt and Miles 2002). The evolutionary success of insects behind this is results from enormous phenotypic and genetic diversity that allows them to overcome ecological, environmental, and biological challenges. Genetic diversity in insects has been reported higher even within an order compared to distant groups. For instance, genetic diversity in an insect order, Diptera (Severson et al. 2004), is known much higher compared to the genetic diversity of two distant species, human and zebrafish (Barbazuk et al. 2000).

Recent advances in whole-genome sequencing allowed researchers to sequence the entire genome of organisms, including insects, which provides a better understanding of genome organization, the function of genes for biological research questions in genetics, medicine, physiology, and evolution, as well as molecular mechanisms—underlying excellent adaptation skills of organisms to a wide range of environmental niches. *Drosophila melanogaster* was the first insect to have its genome sequenced (Adams et al. 2000). This species is widely used for biological research disciplines such as genetics,

biomedicine, physiology, and evolution (Jennings 2011). *Apis mellifera*, the honey bee, as a model insect species for social behavior studies (Consortium 2006), and several termite species genomes were sequenced to understand sociality in insects (Harrison et al. 2018).

In addition, researchers started to sequence entire genomes of harmful insect species globally to develop effective and environmentally friendly pest management strategies. For example, *Anoplophora glapripennis*, the Asian long-horned beetle (McKenna et al. 2016), and *Tribolium castaneum*, the red flour beetle (Tribolium Genome Sequencing et al. 2008), are concerned as global pests for trees and stored agricultural products, respectively, *Leptinotarsa decemlineata*, the Colorado potato beetle, was another insect genome have sequenced (Schoville et al. 2018). These studies unrevealed genome organization, molecular mechanisms behind adaptation to various food sources, and environments and provided insights into the evolution of insects and genomic sources for genome-wide comparison studies.

The genome size of insects sequenced so far greatly varied among insect orders (Li et al. 2019). It has been estimated that the genome size of insects ranges from 91 to 7,752 Mb, contributing to variation in the number of protein-coding genes. Genome sequence studies accumulated enormous data to perform comparative genomics studies using a wide range of insect species from various taxonomic groups and provided insights into gene family expansion/contraction related to some physiological properties such as detoxification of toxins (Thomas et al. 2020), parasitism (Wang et al. 2019), odorant receptors (Brand et al. 2018), feeding choices (McKenna et al. 2016) and genome organization (Petersen et al. 2019) and evolutionary relationships (Misof et al. 2014) by considering their evolutionary history (Thomas et al. 2020). Moreover, studies highlighted species-specific genomic characteristics, i.e., expansion of gene families responsible for insecticide resistance and immune system (Tribolium Genome Sequencing et al. 2008), digestion enzymes, and polyphagy (McKenna et al. 2016).

Protein families are the basis of molecular evolution and expansion, or contraction of gene numbers in protein domains reflect adaptation need of organisms to challenging environmental niches to survive (Lees et al. 2016). Some protein families are key players in adaptation to various environments. Those protein families are Trypsin for digestion of food molecules from different sources, 7TM (seven transmembrane) receptors to detect odorants released from environmental sources, and individuals of the same species for mating and surviving by escaping from predators (Liu et al. 2021).

Additionally, protein families involved in fundamental cellular and metabolic processes are core factors for developmental processes, shaping their developmental plasticity to adapt to various environments. It is crucial to respond and adapt to changing environments to survive and reproduce in insects. Transcriptional changes such as expression of new proteins or increasing expression levels of existing proteins and regulating their expression levels and new isoform expression of a gene are widely involved in adaptations via providing genetic plasticity or quick response to environmental stimuli in organisms. These mechanisms are mostly mediated by transcription factors, Zinc finger (Laity et al. 2001) and, signaling proteins (Bleuven and Landry 2016). Another critical factor for successful adaptation and survival is to overcome various toxic chemical components released by hosts or synthetic products such

as insecticides. A protein family Cytochrome p450 contains detoxification enzymes such as monooxygenases and performs important roles for insect defense mechanisms against these toxic components (Feyereisen 1999). In addition to regulation of expression of proteins, copy number variation of these proteins are also crucial to adaptation and, therefore, to determine ecological properties of species (Gerstein and Berman 2015).

There are only a few studies to compare protein family diversity and variation of protein families in insect species. Nonetheless, these studies focused on only several protein families and were limited to a small number of insect species and orders. Therefore, it still remains largely unknown how protein families are divergent among insects and copy number variation of protein families in insects. Therefore, the present study aimed to address protein family diversity among insects using proteome data of 20 insect species from 7 insect orders generated from whole-genome sequencing present in the National Center for Biotechnology Information (NCBI) database.

Materials And Methods

Data Set

In this study, proteome data of 20 insect species belonging to 7 orders (Hemiptera, Homoptera, Isoptera, Hymenoptera, Diptera, Lepidoptera, and Coleoptera) were retrieved from NCBI (Table 1). Genomic features of the insect species are given in Table 1.

Phylogenomic

For constructing the phylogenomic tree one-to-one ortholog genes were identified using OrthoFinder v2.2.6 (Emms and Kelly 2019), and resulting in a total of 530 one-to-one ortholog were aligned using MAFFT v7.221 (Katoh and Standley 2014) with –auto option. Alignments were trimmed to remove poorly aligned positions by trimAL v1.4.rev13 (Capella-Gutierrez et al. 2009) with –automated1 option. Each trimmed one-to-one ortholog gene was concatenated to generate a super matrix. The super matrix was further used as input for RAxML v8.0.26 (Stamatakis 2014) with 500 bootstrap and partition option to set the most suitable amino acid substitution model for each ortholog gene and to construct the tree. The resulting tree was visualized using ETE toolkit (Huerta-Cepas et al. 2010).

Identification of Protein Families

The proteome file of each insect species was searched against PFAM protein database v32 with hmmer tool v3.1 (Finn et al. 2011) with a cutoff e-value 0.01 to identify protein families. The top hits were considered for further analysis.

Clustering and Correlation of Protein Families

Protein families of insects were clustered using complexHeatmap R package (Gu et al. 2016) in R v3.6.1 (Team 2013) with the options; clustering_distance_rows = "euclidean", clustering_distance_cols =

"euclidean", clustering_method = "pearson". Pairwise correlation of protein families of insect species was calculated using the built-in R function cor in R v3.6.1 (Team 2013) by applying the Pearson method, and results were visualized using factoextra v1.0.7 R package in R v3.6.1 (Team 2013).

Principal Component Analyses of Protein Families

To understand the variance of protein families among insect species, Principal Component Analysis (PCA) was performed using the built-in R function prcomp in R v3.6.1 (Team 2013), and results were visualized using factoextra and PCAtools R packages (Team 2013).

Results

Phylogenomic Relationships in Insects

Phylogenetic relationships of insects are shown in Figure 1. The phylogenetic analysis produced consistently a strong statistical support (all nodes have 100 bootstrap value) for the evolutionary relationships among insect orders.

The tree was separated into two clades; the first consists of hemimetabolous insects belonging to Hemiptera, Homoptera and Isoptera, and the second clade includes holometabolous insects belonging to the Diptera, Lepidoptera, and Coleoptera. The hemimetabolous insects formed a monophyletic group. The Hemiptera and Homoptera orders were derived from a common ancestor. The Hymenoptera insects were grouped as a monophyletic group within the holometabolous insect clade, and the Lepidoptera and Diptera orders were observed to have derived from a common ancestor and have a sister clade relationship with Coleopteran insects.

Protein Family Diversity in Insects

The number of genes in protein families greatly varied among insects (Figure 2 and Table S1). Protein families involved in various metabolic/cellular processes were found overrepresented in almost all insect species. Those include cellular process (Pkinase), digestion (Trypsin), binding activity (zf-C2H2, zf-H2C2_2, RRM_1, WD40, LRR_8, Ank_2, BTB) developmental process (Chitin_bind_4 and Homeobox), cellular signaling (7tm_1, 7tm_6 and 7tm_7), transport membrane transporter activity (MFS_1 and Sugar_tr), oxidoreductase activity (p450), GTPase activity (Ras) and immunoglobulin (Ig_3) (Figure 2 and Table S1). Gene number comparisons in each protein family revealed that order specific protein family expansion was observed in various protein families. Trypsin was found to expand in Diptera, Lig_chan was found to be expanded in Isoptera and zf-H2C2_2, was expanded in Isoptera. MADF_DNA_bdg, THAP, Kelch_1, and PIF1 protein families were expanded in *Acyrthosiphon pisum* (Homoptera). Additionally, some protein families were found to have species-specific expansion. For example, the BTB family was expanded in *Cryptotermes secundus*, *Drosophila melanogaster,* and *Formica exsecta* (Hymenoptera) species. 7tm_6, a sensory protein family, was found to expand in species belonging to two different orders, *F. exsecta* and *Tribolium castaneum* (Coleoptera). Another example of species-specific protein

family expansion was observed in Ig_3, Immunoglobulin domain protein, in *D. melanogaster*. 7tm_7 was expanded in two coleopteran species, *Anoplophora glapripennis* and *T. castaneum* and Ion_trans in *D. melano*gaster (Figure 2, Table S1). Overall, these results suggest that insect species share common protein families with various copy number expansion/contraction in specific protein families. Although copy number variation was observed in a few protein families as order-specific, most diversity in copy number was observed as species-specific.

Protein Family Variation and Correlation in Insects

To understand variation and correlation in copy numbers in protein families PCA was performed. Highest variation was observed in PC1 (47.26%) and followed by PC2 (16.81%), PC3 (9.52%), PC4 (7.82%) and PC5 (4.79%) (Figure 3).

PCA-based insect species (variables) grouped insects into two dimensions with 82.4% and 4.1% variabilities in dimension one and dimension 2. The highest contribution to variability in protein families in insects was observed in *Anopheles gambia* (Diptera), *Danaus plexippus (Lepidoptera)*, *Nicrophorus vespilloides* (Coleoptera), *Agrilus planipennis* (Coleoptera), *Dendroctanus ponderosae* (Coleoptera), *Laodelphax striatellus* (Hemiptera), and *Nilaparvata lugens* (Hemiptera) and other species contribution to variability was moderate (Figure 4). The lowest contribution to variability was observed in *A. pisum* (Figure 4). It was found that overall copy number variations in protein families were similar in insects and showed a high positive correlation among insects (Figure 5), except that *C. secundus* and *A. pisum* showed a lower positive correlation compared to other species. Pairwise protein family comparisons among insects showed a strong positive correlation among insects with exception of *C. secundus* and *A. pisum* that have a lower correlation with other insect species (Figure 5). *C. secundus* and *A. pisum* have a positive correlation with each other (Figure 4).

PCA for protein families (individuals) showed that most protein families were similar in insects and grouped together and positioned at the middle point of the axis. Protein families PBP_GOBP, CBM_14, Homeobox, 7tm_6, COesterase, Ras, Ig_3, LRR_8, p450, and Chitin_bind_4 was found to be positively correlated as a group, while Kelc_1, THAP, MADF_DNA_blog, BTB, Pkinase_Tyr, I-set, adh_shor, Ank_2, Sugar_tr, 7tm_1, RRM_1, WD40, and zf_C2H2 were found to have positive correlation together in a group (Figure 6). Regarding variability, in each principal component (PC), protein families zf-H2C2_2, Trypsin, Pkinase, 7tm_6, THAP, H_psq, and 7tm_7 was found to have the highest contribution to variability in almost all PCs (Figure 7).

Discussion And Conclusion

In this study, a total of 20 insect genomes from various taxonomic groups were investigated to figure out protein family expansion/contraction and diversity of protein families in insects.

The phylogenetic tree inferred from single-copy genes separated holometabolous (Diptera, Lepidoptera and Coleoptera), and hemimetabolous insects (Hemiptera, Homoptera, and Isoptera) and placed them

into two distinct clades. The holometabolous insects undergo complete metamorphosis and include four life stages; egg, larvae, pupa, and adult, while the hemimetabolous insects undergo an incomplete metamorphosis that includes egg, nymph, and adult stages (Gullan and Cranston 2014). Similar evolutionary relationships of insects were reported from previously published studies (McKenna et al. 2016; McKenna et al. 2019; Misof et al. 2014). This suggests that morphogenesis is the major contributor to insect evolution, and proteins involved in morphogenesis are essential in insect evolution.

Insects have core protein families to maintain their fundamental metabolic and cellular processes. Those protein families are Trypsin, a serine protease involved in digestion of food molecules via proteolysis activity by breaking down proteins into peptides and found in the digestive systems of many animals (Rawlings and Barrett 1994). In the present study. Trypsin was found to be expanded in all insect species. This finding suggests that feeding is essential for all insect species for survival and reproduction and provides energy to maintain cellular processes during their life cycles. Protein kinase domain (Pkinase) is involved in a wide range of cellular activity via protein phosphorylation that plays roles in several cellular process and metabolism, cell movement, transcription cell cycle regulation (Hanks and Quinn 1991), embryonic development, and response to environmental stimuli (Scheeff and Bourne 2005).

Further, a zinc finger (zf-C2H2_2) protein family also shows expansion. This protein family is known to involve in binding activity through binding DNA, RNA, and proteins (Englbrecht et al. 2004), and can control transcription of different functional genes as a response to physiological processes and external environmental stimuli in insects (Guo et al. 2018). Therefore, the zinc finger (zf-C2H2_2) family could be one of the major contributions to adaptation to various environmental conditions of insects by regulating genes based on their needs.

Chitin is a primary and significant component of the exoskeleton of insects. This polysaccharide forms complex structures in the combination of various assortments of cuticle and matrix proteins (Zhu et al. 2016). The growth, morphogenesis, and development in insects are strictly mediated by biosynthesis and modification of chitin (Zhu et al. 2016). Chitin_bind_4 is a protein family involved in the structural constituent of the cuticle by binding chitin in arthropods and provides structural materials for insect cuticles and is essential for protection, support, and locomotion during life cycles of insects (Rebers and Willis 2001). Molting and metamorphosis are core and crucial processes in the development of the insects and continue through their life cycles (Merzendorfer and Zimoch 2003). Therefore, expanding this protein family in all insects showed that molting, morphogenesis, and growth are core evolutionary conserved processes in insects.

Other protein families expanded in all insects were those involved in a binding activity such as RNA binding domain (RRM_1), repetitive proteins (WD40), and leucine-rich repeats (LRR_8), monooxygenases p450 (cytochromes), and cellular signaling protein domains (7tm_1, 7tm_6 and 7tm_7). Among these, signaling protein families, 7tm_1, 7tm_6, and 7tm_7, are essential for insects to survive and reproduction such as determine food source and selection of a mate. These protein families can be important for habitat selection and population dynamics in insects and, therefore, crucial to adapting to various

environments. P450 protein family consists of monooxygenase enzymes involved in a wide range of processes such as growth and development, insect defense via detoxification, and protection against xenobiotics such as pesticide resistance and tolerance to toxins released by plants (Feyereisen 1999).

High variation in gene number was found in zf-H2C2_2, Trypsin, Pkinase, 7tm_6, THAP, H_psq, and 7tm_7 protein families. These protein families were found to be the most dynamically changing and varying in copy numbers in all insects. Variation in these protein families may reflect species-specific adaptation needs for their environments and may have resulted from gene duplication events during insect life history.

Overall, it was found that although all insect genomes were expanded in the same protein families, the copy number of protein families varied among insect species. This suggests that all insects need protein families involved in core biological and physiological processes to maintain a proper life cycle that includes reproduction, growth and development, and communications with other members of the same order and ecology via response to environmental stimuli. Some protein families were observed as species-specific expansion/contraction, and variation in gene number of protein families was observed in even the same order. It proposes that high genetic diversity is found in insects and species-specific protein family evolution has an essential role in the life cycles of insects. Moreover, variation in copy numbers in protein families reflects the evolution and diversification of insects as well as morphological and physiological characteristics. Current findings of this study provide insights into evolutionary conserved protein families and variation of gene numbers in protein families in insects.

Declarations

Funding

No funding was available.

Conflict of interest

The author declares no conflicts of interest.

Ethical approval

This article does not contain any experimental studies that include human participants or animals.

Availability of data and material

The data produced from the manuscript is given in the Tables.

Code availability

No custom code/scripts were used.

Acknowledgements

The author kindly thanks to Prof. Dr. Serap MUTUN and Dr. İsmail KOÇ for their suggestions.

References

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF et al. (2000) The genome sequence of Drosophila melanogaster. Science 287:2185-2195
2. Barbazuk WB, Korf I, Kadavi C, Heyen J, Tate S, Wun E, Bedell JA, McPherson JD, Johnson SL (2000) The syntenic relationship of the zebrafish and human genomes. Genome Res 10:1351-1358
3. Bleuven C, Landry CR (2016) Molecular and cellular bases of adaptation to a changing environment in microorganisms. Proc Biol Sci 283
4. Brand P, Robertson HM, Lin W, Pothula R, Klingeman WE, Jurat-Fuentes JL, Johnson BR (2018) The origin of the odorant receptor gene family in insects. Elife 7
5. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973
6. Consortium HGS (2006) Insights into social insects from the genome of the honeybee Apis mellifera. Nature 443:931
7. Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:238
8. Englbrecht CC, Schoof H, Böhm S (2004) Conservation, diversification and expansion of C2H2 zinc finger proteins in the Arabidopsis thaliana genome. BMC genomics 5:1-17
9. Feyereisen R (1999) Insect P450 enzymes. Annu Rev Entomol 44:507-533
10. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic acids research 39:W29-W37
11. Gaunt MW, Miles MA (2002) An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. Mol Biol Evol 19:748-761
12. Gerstein AC, Berman J (2015) Shift and adapt: the costs and benefits of karyotype variations. Curr Opin Microbiol 26:130-136
13. Gu Z, Eils R, Schlesner M (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32:2847-2849
14. Gullan PJ, Cranston PS (2014) The insects: an outline of entomology. John Wiley & Sons,
15. Guo Z, Qin J, Zhou X, Zhang Y (2018) Insect Transcription Factors: A Landscape of Their Structures and Biological Functions in Drosophila and beyond. Int J Mol Sci 19
16. Hanks SK, Quinn AM (1991) [2] Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. Methods in enzymology 200:38-62

17. Harrison MC, Jongepier E, Robertson HM, Arning N, Bitard-Feildel T, Chao H, Childers CP, Dinh H, Doddapaneni H, Dugan S et al. (2018) Hemimetabolous genomes reveal molecular basis of termite eusociality. Nat Ecol Evol 2:557-566
18. Huerta-Cepas J, Dopazo J, Gabaldon T (2010) ETE: a python Environment for Tree Exploration. BMC Bioinformatics 11:24
19. Jennings BH (2011) Drosophila–a versatile model in biology & medicine. Materials today 14:190-195
20. Katoh K, Standley DM (2014) MAFFT: iterative refinement and additional methods. Methods Mol Biol 1079:131-146
21. Laity JH, Lee BM, Wright PE (2001) Zinc finger proteins: new insights into structural and functional diversity. Current opinion in structural biology 11:39-46
22. Lees JG, Dawson NL, Sillitoe I, Orengo CA (2016) Functional innovation from changes in protein domains and their combinations. Current opinion in structural biology 38:44-52
23. Li F, Zhao X, Li M, He K, Huang C, Zhou Y, Li Z, Walters JR (2019) Insect genomes: progress and challenges. Insect Mol Biol 28:739-758
24. Liu N, Li T, Wang Y, Liu S (2021) G-Protein Coupled Receptors (GPCRs) in Insects-A Potential Target for New Insecticide Development. Molecules 26
25. McKenna DD, Scully ED, Pauchet Y, Hoover K, Kirsch R, Geib SM, Mitchell RF, Waterhouse RM, Ahn SJ, Arsala D et al. (2016) Genome of the Asian longhorned beetle (Anoplophora glabripennis), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. Genome Biology 17
26. McKenna DD, Shin S, Ahrens D, Balke M, Beza-Beza C, Clarke DJ, Donath A, Escalona HE, Friedrich F, Letsch H et al. (2019) The evolution and genomic basis of beetle diversity. Proc Natl Acad Sci U S A 116:24729-24737
27. Merzendorfer H, Zimoch L (2003) Chitin metabolism in insects: structure, function and regulation of chitin synthases and chitinases. J Exp Biol 206:4393-4412
28. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG et al. (2014) Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763-767
29. Petersen M, Armisen D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, Misof B (2019) Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC Evol Biol 19:11
30. Rawlings ND, Barrett AJ (1994) Families of serine peptidases. Methods Enzymol 244:19-61
31. Rebers JE, Willis JH (2001) A conserved domain in arthropod cuticular proteins binds chitin. Insect Biochem Mol Biol 31:1083-1093
32. Scheeff ED, Bourne PE (2005) Structural evolution of the protein kinase–like superfamily. PLoS Comput Biol 1:e49
33. Schoville SD, Chen YH, Andersson MN, Benoit JB, Bhandari A, Bowsher JH, Brevik K, Cappelle K, Chen M-JM, Childers AK (2018) A model species for agricultural pest genomics: the genome of the

Colorado potato beetle, Leptinotarsa decemlineata (Coleoptera: Chrysomelidae). Scientific reports 8:1-18

34. Severson DW, DeBruyn B, Lovin DD, Brown SE, Knudson DL, Morlais I (2004) Comparative genome analysis of the yellow fever mosquito Aedes aegypti with Drosophila melanogaster and the malaria vector mosquito Anopheles gambiae. J Hered 95:103-113
35. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313
36. Team RC (2013) R: A language and environment for statistical computing.
37. Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, Anstead CA, Ayoub NA, Batterham P, Bellair M et al. (2020) Gene content evolution in the arthropods. Genome Biol 21:15
38. Tribolium Genome Sequencing C, Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Beeman RW, Brown SJ et al. (2008) The genome of the model beetle and pest Tribolium castaneum. Nature 452:949-955
39. Wang X, Zhang Y, Zhang R, Zhang J (2019) The diversity of pattern recognition receptors (PRRs) involved with insect defense against pathogens. Curr Opin Insect Sci 33:105-110
40. Zhang G, Wang H, Shi J, Wang X, Zheng H, Wong GK, Clark T, Wang W, Wang J, Kang L (2007) Identification and characterization of insect-specific proteins by genome data analysis. BMC Genomics 8:93
41. Zhu KY, Merzendorfer H, Zhang W, Zhang J, Muthukrishnan S (2016) Biosynthesis, Turnover, and Functions of Chitin in Insects. Annu Rev Entomol 61:177-196

Tables

Table 1. The Insect Species Used in this Study.

Order	Family	Species	Number of Genes
Coleoptera	Cerambycidae	*Anoplophora glapripennis*	21,859
	Buprestidae	*Agrilus planipennis*	15,497
	Curculionidae	*Tribolium castaneum*	18,534
	Curculionidae	*Dendroctanus ponderosae*	13,457
	Chrypsomelidae	*Leptinotarsa decemlineta*	24,830
	Silphidae	*Nicrophorus vespilloides*	13,516
	Scarabaidae	*Onthaphagus taurus*	17,483
Hymenoptera	Apidae	*Apis mellifera*	15,314
	Agaonidae	*Ceratosolen solmsi marchali*	12833
	Formicidae	*Formica exsecta*	22509
Hemiptera	Delphacidae	*Laodelphax striatellus*	17512
	Delphacidae	*Nilaparvata lugens*	24319
Homoptera	Aphididae	*Acyrthosiphon pisum*	36,195
Diptera	Culicidae	*Anopheles gambia*	14102
	Drosophilidae	*Drosophila melanogaster*	30,504
	Culicidae	*Culex quinquefasciatus*	18883
Lepidoptera	Nymphalidae	*Danaus plexippus*	15,128
	Bombycidae	*Bombyx mori*	22510
Isoptera	Termopsidae	*Zootermopsis nevadensis*	14,610
	Kalotermitidae	*Cryptotermes secundus*	26728

Figures

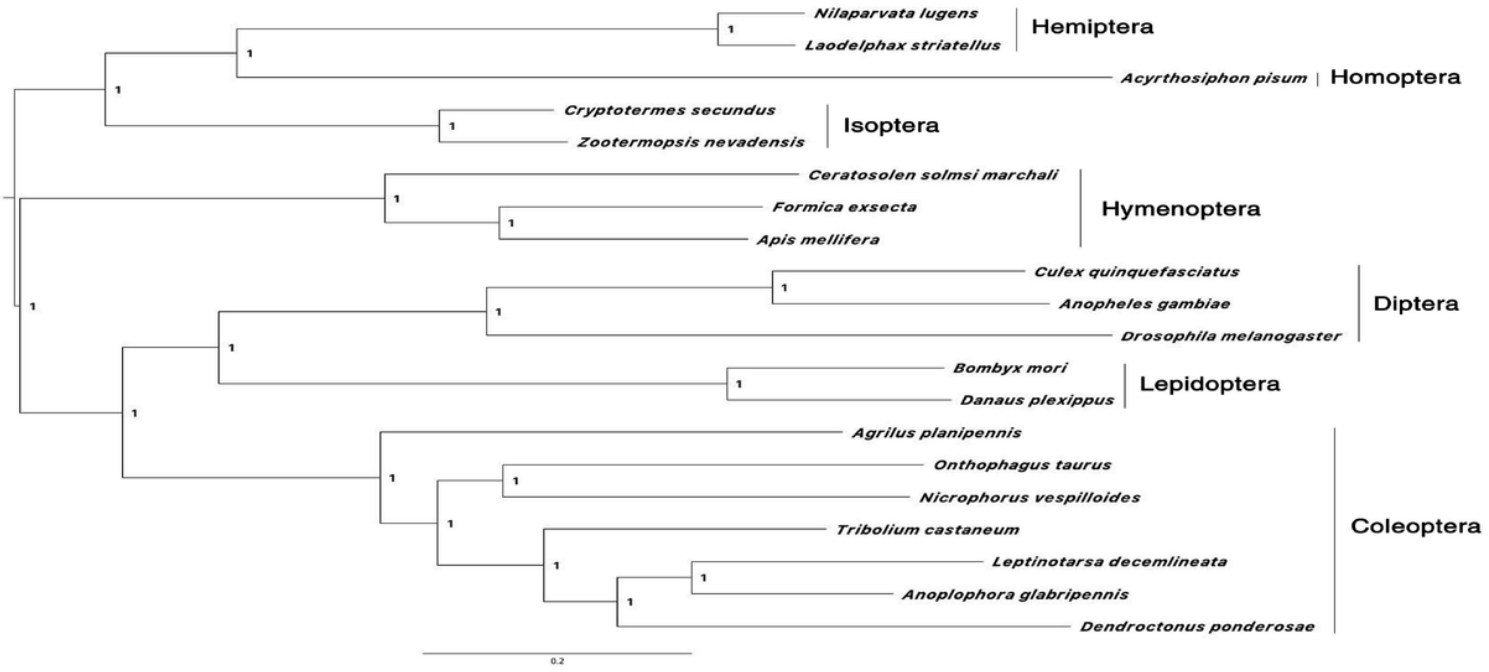

Figure 1

The Phylogenetic Tree of Insects. The Tree Was Inferred Using A Total of 530 One-to-One Single Copy Ortholog Genes with 500 Bootstrap and Partition Option Using Raxml V8.0.26.

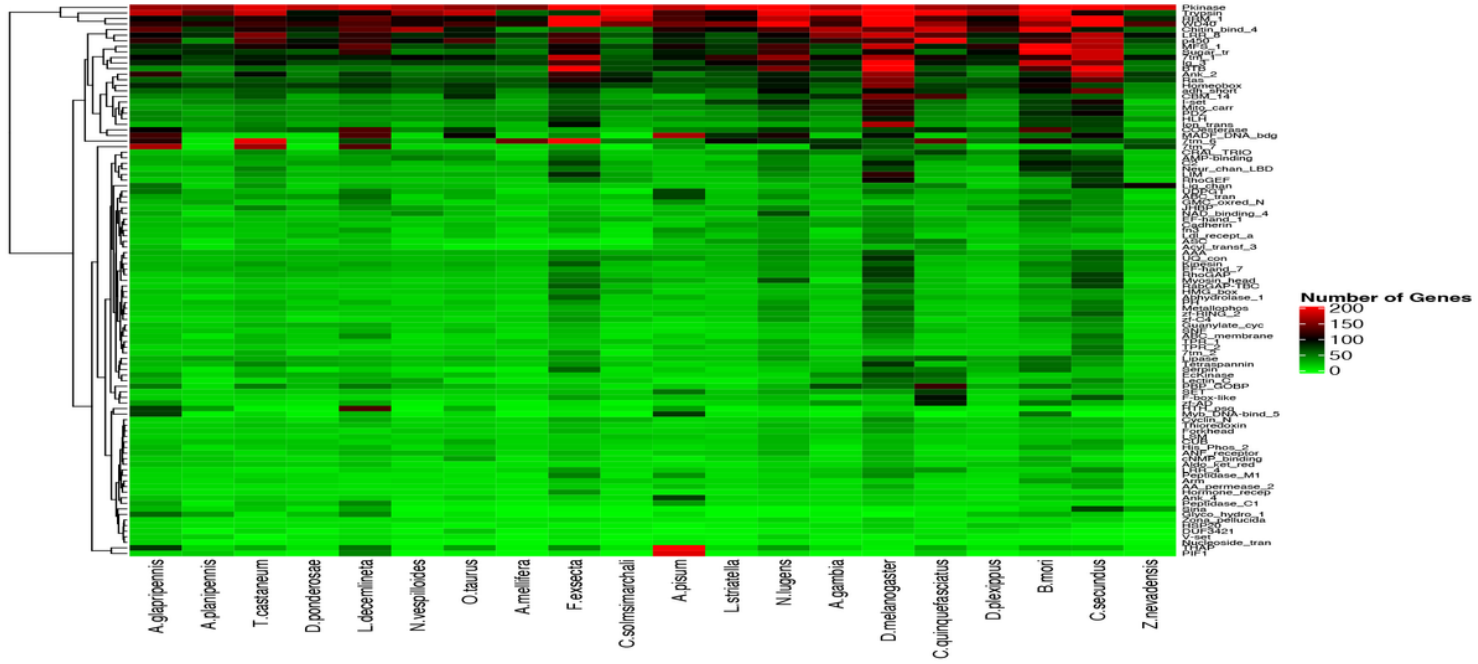

Figure 2

Cluster of the Most Abundant 100 Protein Families.

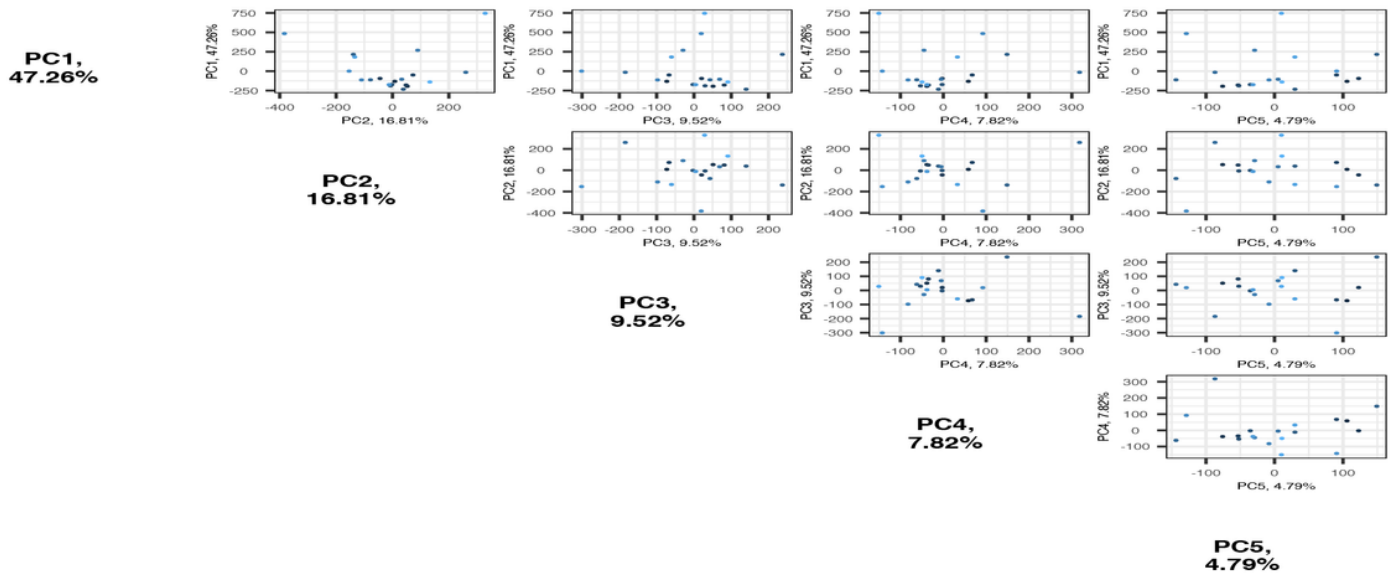

Figure 3

The Principal Component Analysis (PCA) Biplots Showing Top 5 Principal Component Variations.

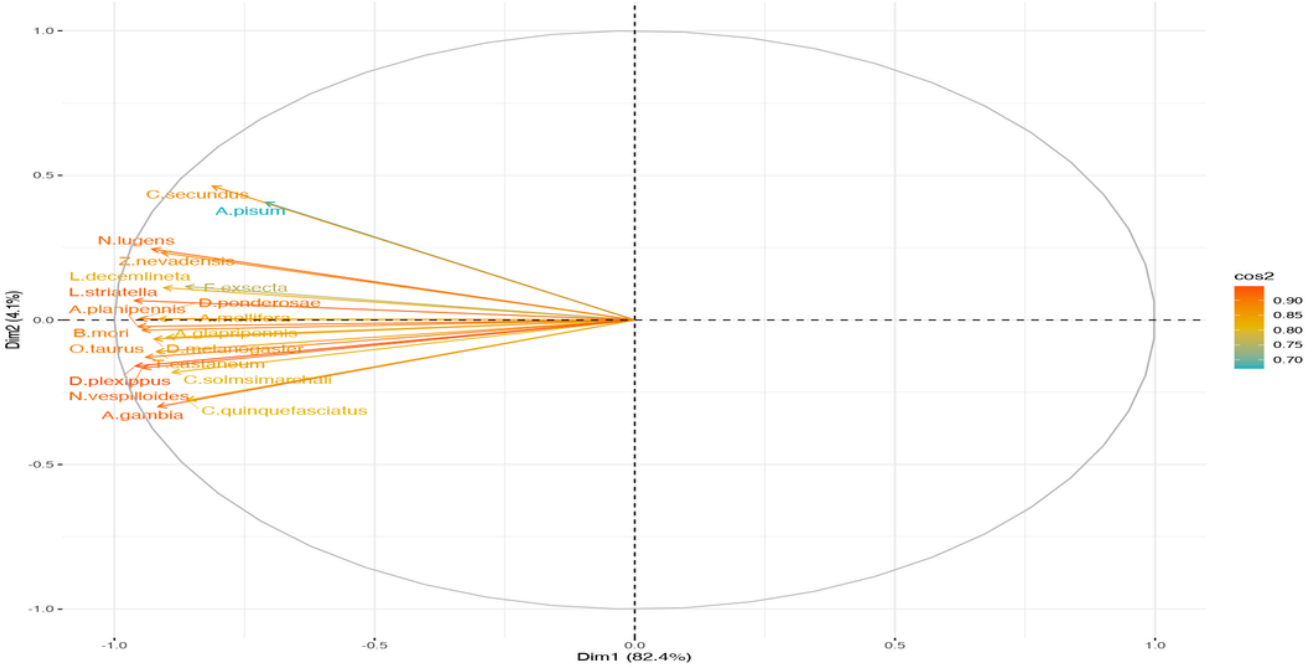

Figure 4

The Principal Component Analysis (PCA) Based on PFAM Copy Numbers in Insects.

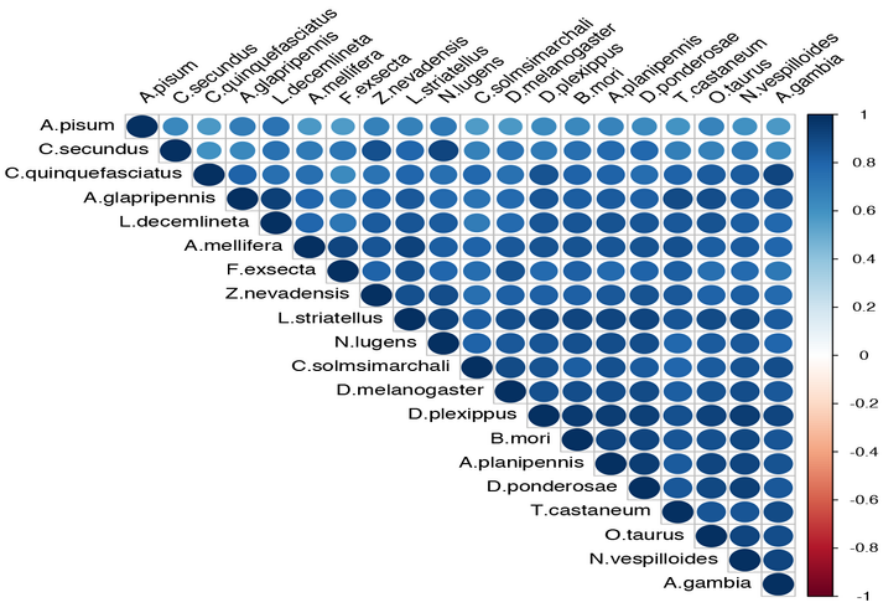

Figure 5

The Correlation Matrix Showing Pairwise Correlation in Insect Pairs Inferred from Protein Family Copy Numbers.

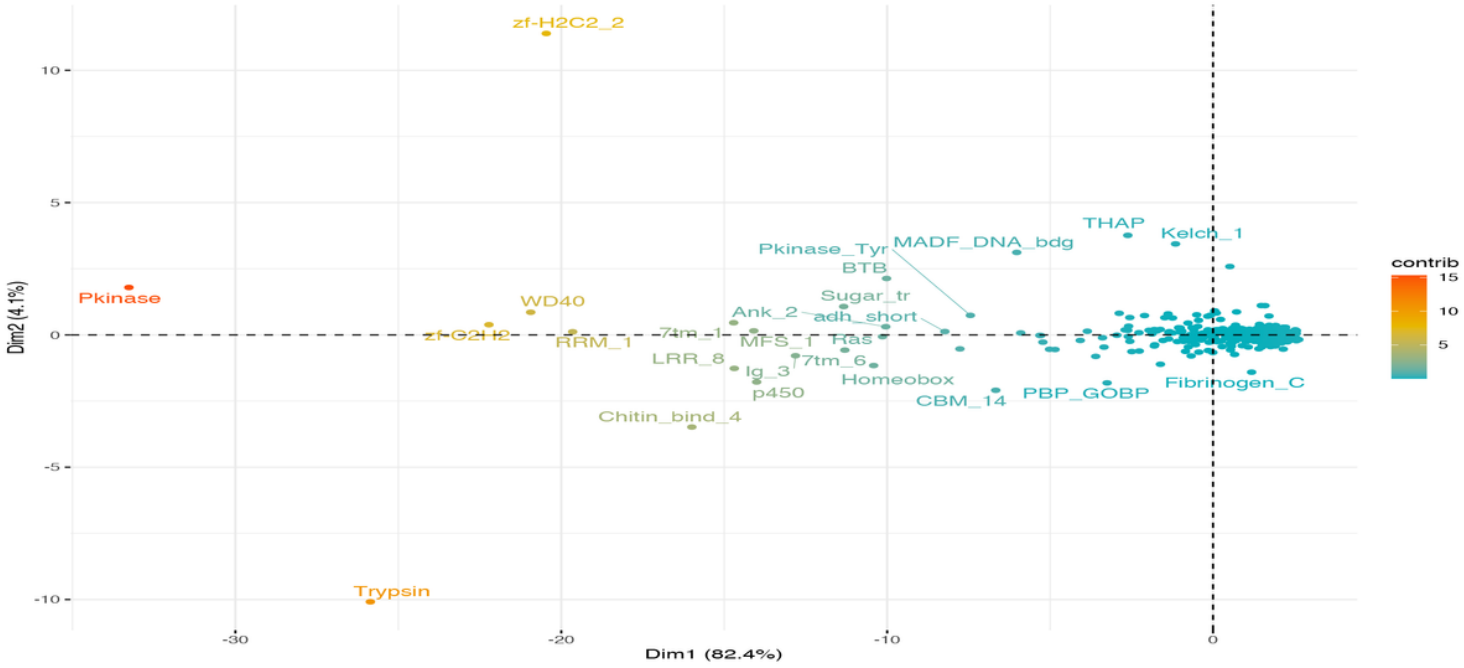

Figure 6

The Principal Component Analysis (PCA) of Protein Families in insects.

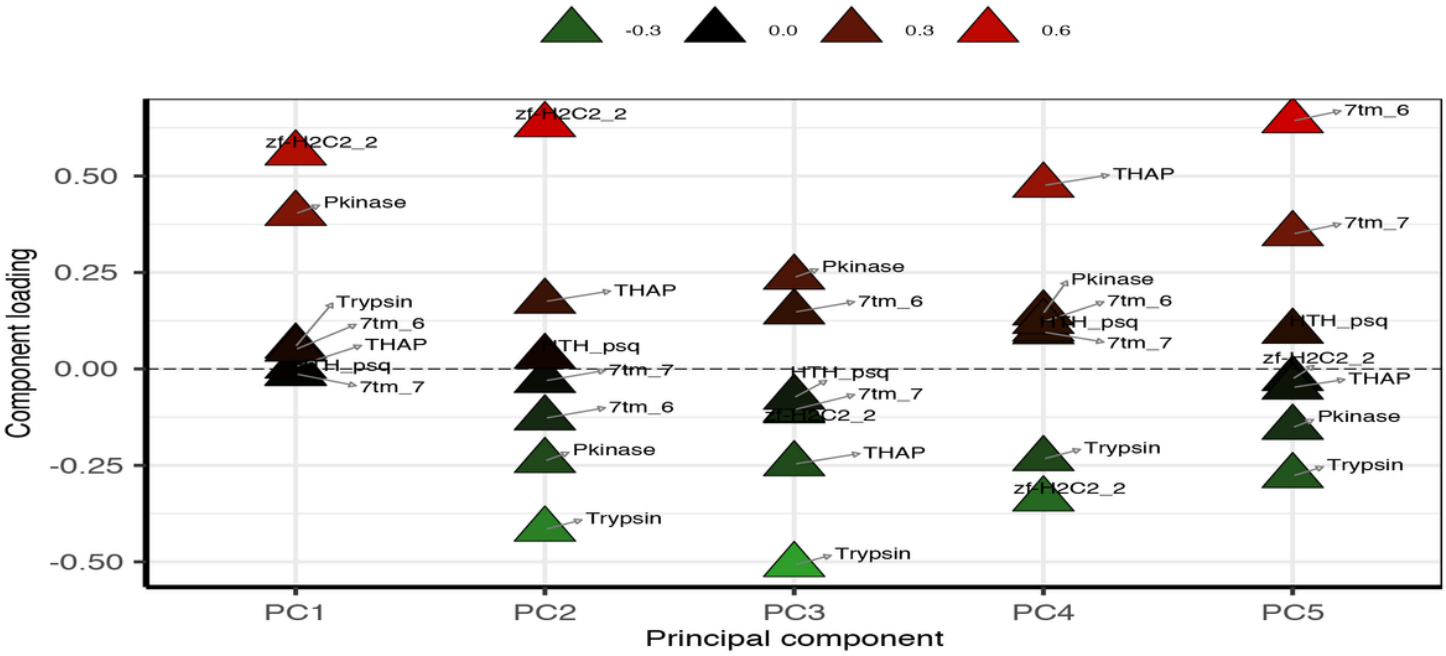

Figure 7

The Loading Plots Showing the Most Variance PFAM Domains In Insects.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.xlsx