Predicting Insect Invasiveness with Whole-Genome Sequencing Data

Cong Huang[1,3,8#], Nianwan Yang[2#], Shuping Wang[4], Xiaodan Fan[5], Cong Pian[1], Jiapeng Luo[1], Xi Li[6], Kun Lang[1], Longsheng Xing[3], Mingxing Jiang[1], Wanxue Liu[2], Wanqiang Qian[3*], Daniel Simberloff[7*], Fanghao Wan[2,3,9*], Fei Li[1*]

[1]Ministry of Agriculture Key Laboratory of Molecular Biology of Crop Pathogens and Insect Pests, Institute of Insect Sciences, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, 310058, China

[2]State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing, 100193, China

[3]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, 518120, China.

[4]Technical Centre for Animal Plant and Food Inspection and Quarantine, Shanghai Customs, Shanghai, 200135, China

[5]Department of Statistics, The Chinese University of Hong Kong, Hong Kong SAR, China

[6]College of Computer Science, Zhejiang University, Hangzhou, 310027, China

[7]Department of Ecology & Evolutionary Biology, The University of Tennessee, Knoxville, TN, 37996, USA.

[8]Plant Protection College, Hunan Agricultural University, Changsha, 410128，China

[9]College of Plant Health and Medicine, Qingdao Agricultural University, Qingdao, 266109, China

# These authors contributed equally

*Corresponding authors, Dr. Fei Li, Dr. Fanghao Wan, Dr. Daniel Simberloff, Dr. Wanqiang Qian, **Email:** lifei18@zju.edu.cn; wanfanghao@caas.cn； dsimberloff@utk.edu; qianwanqiang@caas.cn

Abstract

Background: Invasive alien insects threaten agriculture, biodiversity, and human livelihoods globally. Unfortunately, insect invasiveness still cannot be reliably predicted. Empirical policies of insect pest quarantine and inspection are mainly designed against species that are already problematic.

Results: We conducted a comparative genomic analysis of 37 invasive insect species and six non-invasive insect species, showing that the gene families associated with defense, protein and nucleic acid metabolism, chemosensory function, and transcriptional regulation were significantly expanded in invasive insects, suggesting that enhanced abilities in self-protection, nutrition exploitation, and locating food or mates are intrinsic features conferring invasiveness in insects. By using these intrinsic genome features, we proposed an invasiveness index and estimated the invasiveness of 99 other insect species with genome data, classifying them as highly, moderately, or minimally invasive. Insects possessing all these aforementioned enhanced abilities are predicted to be highly invasive, and vice versa. Next, a logistic-regression classifier was trained to predict insect invasiveness, achieving 93.2% accuracy.

Conclusions: We present evidence that several traits may confer invasiveness in insects and these features can be used to predict insect invasiveness accurately, and we quantify insect invasiveness with an invasiveness index.

Keywords

Insect pest; Invasiveness; Genome features; Comparative genomics; Invasiveness index

Background

Invasive species threaten agriculture, biodiversity, and human livelihoods. The estimated global economic loss to invasive insects is US $70.0 billion annually [1]. Increased globalization and connectedness via trade, as well as environmental changes owing to climate change, will significantly increase invasive species threats [2, 3]. Many studies have shown some common properties in invasive species. Invasive plants have syndromes including: a tendency to be annual or biennial; increased plant height and specific leaf area; hermaphroditism with longer and earlier flowering; clonal growth and monoecy; and higher fecundity [4, 5]. Invasive birds tend to: be large; prioritize future over current reproduction; be less migratory; and be widespread in the source region [6]. In contrast, invasive freshwater fishes have smaller body size, fast reproduction, high activity, and boldness, and are omnivorous with high physiological tolerances [7]. Invasive insects are reported to tend to have some intrinsic features such as parthenogenesis, high dispersal ability, a dormant or resilient stage, and a longer adult stage [8].

These observations have yielded at least four hypotheses about invasive species: 1) enemy release hypothesis – escape from natural enemies in the original habitat [9]; 2) increased competitive ability hypothesis – efficient transfer of biological resources from enemy defense to growth and reproduction [10]; 3) novel weapons hypothesis – carrying parasites such as microsporidia that negatively affect or kill native species but not the invading species [11]; and 4) inherent superiority hypothesis – invasive species have intrinsic traits superior to those of non-invasive species, at least in new regions [12].

Accurately identifying invasiveness-related traits and predicting invasiveness of a species is important for pest risk assessments and developing national quarantine policies. However, the traits identified as associated with invasiveness are quite controversial and do not characterize all invasive species, especially in insects [13]. The controversy has hampered development of a highly accurate method of predicting invasiveness, though much effort has

been expended on this project, such as Invasive Species Predictive Schemes (ISPS) [14] and the SCOPE project [5]. A problem is that an invasion consists of several distinct stages [15], and traits that would lead a species to pass successfully through one stage may not be the same traits that would conduce to success at a different stage. With respect to risk, different stages are at issue. The first stage, transport and initial introduction, consists of a propagule arriving with human assistance in a new, distant site. Whatever traits a species has that facilitate its association with a transport vector (such as ballast water, shipping containers, or agricultural products) increase risk of transport [16]. In this paper, we focus on the next stages, establishment and spread. Once a species has arrived in a new region, do particular traits increase the risk that it will persist and spread? The propensity to establish and spread once introduced is what we define as “invasiveness” in this study, although the term “invasiveness” has also been used at times in the literature to refer to the first stage, simply arriving in a new region.

As the cost of whole-genome sequencing has decreased dramatically, hundreds of insect genomes have been sequenced [17], providing an opportunity to conduct comparative genomic analysis in insects. Comparative genomics is a powerful method to identify common genomic features in maize [18] and in certain carnivores [19]. Moreover, a recent report showed that human physical traits can be predicted from whole-genome sequencing data [20]. We thus asked whether invasive insects have common genomic features distinguishing them from other insect species. To this end, we focused on isolating invasive features at the gene family level. We collected 142 insect genomes, from which we analyzed 43 species including 37 invasive insects and six non-invasive insects with high quality assemblies. By conducting comparative genomic analysis, we found that some gene families associated with defense, energy, chemosensory function, and transcriptional regulation functions were significantly expanded in invasive insects but not in the non-invasive ones. An invasiveness index based on these families was proposed to quantify insect invasiveness. Moreover, these gene families were treated as candidate features in a machine-learning algorithm

(Determining Invasiveness based on Genome Sequences, DIGS) that we introduced to train a highly accurate classifier to predict insect invasiveness. Both the invasiveness index and DIGS were applied to predict invasiveness of the other 99 insects that were not used for training. In summary, we provide a new genetic approach to analyze quantitatively the "invasiveness" of insect species, and this approach could be extended to other biological taxa and will be improved as more genome data become available for training samples.

Results

Selection of 37 invasive insects and six non-invasive insects

From the 142 insect species for which genome data are available (Supplementary Table 1), we excluded ones lacking high-quality genome assemblies as well as those that are not confirmed as having been introduced anywhere (89 species) and thus could not be classified as either invasive or non-invasive. We confirmed: 1) 37 insects known to be invasive by literature references, including nine Diptera, two Coleoptera, fifteen Hymenoptera, five Hemiptera, and six Lepidoptera; 2) six non-invasive insects (two Diptera, one Hemiptera, and three Lepidoptera) according to the criterion of having been introduced to non-native regions but not spreading or exhibiting any signs of invasion in the introduced regions (Supplementary Table 2). From these 43 insect species, we identified 183 single-copy orthologous genes by all-vs-all BLASTP [21] against all proteins in the OrthoMCL [22]. We constructed a phylogenetic tree using these single-copy orthologous genes to infer the evolutionary relationship of these species.

The general genome features are not related to invasiveness

We calculated general genome features of these 43 insects including genome size, GC content, gene number, amount of repeat sequences, number of expanded gene families, and number of contracted gene families (Fig. 1). None of these features differed significantly between invasive and non-invasive insects, indicating that high invasiveness might be ascribed only to several key gene families closely associated with invasive traits, rather than

to general genome features (Supplementary Fig. 1 and Supplementary Table 7).

Identifying gene families associated with insect invasiveness

It has been reported that invasive insects share some traits, such as nutrition acquisition advantage, advanced defense systems, and high reproductive ability [8]. For the inherent superiority hypothesis to be valid, we reasoned that gene families conferring functions related to invasiveness should be positively selected and most likely expanded. To this end, we analyzed the expansions and contractions of gene families in a phylogenetic context in the 43 insects using the program CAFÉ (v3.0) [23]. We found 36 gene families to have expanded in at least 13 of the 37 invasive species with the additional criterion that the ratio of number of invasive species to number of non-invasive species in which the gene family expanded exceeded 12 (the criterion was determined by testing the accuracy of invasiveness classification for a range of ratios from 4 to13; among these, a ratio of 12 achieved the highest accuracy) (Supplementary Table 8). The gene families expanded more frequently in invasive species by this criterion were treated as candidate gene families and were grouped into four categories based on their functional associations: 1) associated with defense; 2) associated with energy; 3) associated with chemosensory function; 4) associated with transcriptional regulation (Supplementary Table 8).

Next, to evaluate the contribution of these candidate gene families to invasiveness, we used a two-step logistic regression procedure (see Materials and Methods) to select the gene families whose expansion is contributing to invasiveness and to determine the relative weights of their contributions (based on the expanded gene numbers in Supplementary Table 3). The results show that in total 14 gene families are associated with invasiveness (Supplementary Table 4). The expansion pattern of these gene families varied in different invasive insects, suggesting that a variety of traits have conferred invasiveness in insects (Fig. 2). For example, invasive hymenopterans have enhanced defense ability and advanced chemosensory function, while invasive lepidopterans have enhanced abilities of defense and

energy metabolism, and particularly transcriptional regulation.

Invasiveness index for insect invasiveness

We next seek to estimate the invasiveness of insects by using the expansion indexes of four function groups. We calculated the expansion indexes for each function group in each insect species, which involves the weighting coefficients of the 14 gene families that resulted from the first-step logistic regression (see methods) and the corresponding expansion gene numbers (Supplementary Table 9). We built the invasiveness index of a species by using the second-step logistic regression model (see methods) that estimates the weighting coefficients of the four function groups with the following steps:

1) $m_j = 116.98y_{1j}+12.98y_{2j}+6.29y_{3j}+6.12y_{4j}+63.89$,

2) $$n_j = \begin{cases} \log_{10}(|m_j|), & if \ \ m_j \geq 1 \\ 0, & if \ \ -1 < m_j < 1 \\ -\log_{10}(|m_j|), & if \ \ m_j \leq -1 \end{cases},$$

3) $$z_j = 1 - \frac{1}{1 + e^{n_j}},$$

where z_j is the invasiveness index of the *j*th species and y_{1j} to y_{4j} are the expansion indexes of the four function groups of the *j*th species.

Then we calculated the invasiveness indexes of all 142 insects (Fig. 3, Fig. 4, Supplementary Table 9 and Supplementary Table 5). For the 37 invasive insects, all have high invasiveness indexes. By contrast, all six non-invasive insects have minimal invasiveness indexes (Fig. 3). For the 99 other insects, we classified their predicted invasiveness into three levels based on invasiveness indexes: high invasiveness (0.9 to 1), moderate invasiveness (0.2 to 0.9), and minimal invasiveness (0 to 0.2) (Fig. 4). Among these 99 insects, ten species have been reported to be highly invasive; eight of these were assigned a high invasiveness index and one was assigned a moderate invasiveness index (Fig. 4).

The results showed that these four aspects of capacities were generally essential for high invasiveness: defense, energy, chemosensory function, and transcriptional regulation. Highly

189 invasive insects tend to have high expansion indexes in all four function categories; insects
190 that are minimally invasive tend to have low expansion indexes in all four function groups.
191 Among the 99 insect species for which we lack adequate data on invasiveness, we predicted
192 species to be moderately invasive if they have high expansion indexes in some function
193 categories but low expansion indexes in others, while we predicted them to be highly or
194 minimally invasive according the rules just stated (Fig. 4).
195
196 **Classifying insect invasiveness by machine learning**
197 Having identified putative inherent genome features associated with insect invasiveness, we
198 adopted these features to develop a machine learning algorithm, named Determining
199 Invasiveness based on Genome Sequences (DIGS), in order to classify insects in terms of
200 invasiveness. DIGS used a random forest algorithm for feature selection and then used a
201 logistic regression model to construct a classifier; six-fold cross-validation was used to train
202 the DIGS classifier.
203
204 In each cross-validation, we used the R package “Boruta” [24] to evaluate the contributions of
205 the 36 candidate gene families to invasiveness. Boruta is designed as a wrapper with a
206 random forest classification. The key gene families selected by Boruta were next used to
207 construct a logistic regression model to estimate classification efficiency. As a result, two gene
208 families, pao retrotransposon peptidase, and putative nuclease HARBI, were stably selected
209 to predict insect invasiveness (Supplementary Table 6).
210
211 Given the high average accuracy and the balance between positive and negative samples of
212 the 43 species (Supplementary Table 6), these two gene families were used to construct a
213 logistic classifier to classify invasiveness. The results indicated that DIGS performs well in
214 classifying insect invasiveness, with an average accuracy of 93.2%. Sensitivity, specificity,
215 and precision were 88.1%, 100%, and 100%, respectively (Supplementary Table 10). Based
216 on the analysis of a ROC (Receiver Operating Characteristic) curve, the AUC (Area under the

Curve of ROC) [25] was 0.953, suggesting good performance by the DIGS classifier (Fig. 5). Next, we used this classifier to predict invasiveness of the other 99 insects, those not used for training. With a cutoff of 0.5, 56 species (56.6%) were classified as invasive.

Because we have developed two systems separately to evaluate insect invasiveness, we compared the consistency of the invasiveness index and the DIGS. Fig. 6a showed that 94.1% of highly invasive insects as determined by the invasiveness index were predicted to be invasive by DIGS, whereas 85.3% of insects predicted not to be invasive by the invasiveness index were predicted not to be invasive by DIGS (Fig. 6a). On the other hand, of those 56 insects predicted to be invasive by DIGS, 57.1% were classified as highly invasive and 33.9% as moderately invasive by the invasiveness index. Of the 43 insects predicted not to be invasive by DIGS, the invasiveness index analysis predicted 67.4% would not be invasive (Fig. 6b). These results showed substantial consistency between the invasiveness index and DIGS.

Discussion

One of the most significant challenges regarding biological invasions is to predict the risk of successful invasion. Meta-analysis of large data sets is increasingly used to predict the risk of invasion by non-native species globally, by considering, for risk of introduction, variables such as the quantity of international trade [2, 3] and the capacity of the transport vectors to assist arrival [2]; for risk of establishment, variables such as disturbance factors [2], biodiversity indices [26], and the similarity of biotic and abiotic conditions between the native locations and the location of a newly arrived IAS [3] have been used. However, the species characteristics that increase the risk of population establishment and spread once the species is introduced, which we define as “invasiveness” in this study, are still not well determined. This lacuna contributes to ineffective management and slow responses to newly arrived IAS [27]. Here, we have attempted to fill this gap by presenting a new approach based on machine learning and genome data to predict high-risk invasive insect species.

Based on the hypothesis that invasive insects tend to share particular invasiveness-related traits, we conducted a comparative genomic analysis of invasive and non-invasive insect species to identify gene families commonly expanded in invasive species. This strategy yielded 14 gene families associated with insect invasiveness. These 14 gene families can be grouped into four function categories: defense, energy, chemosensory function, and transcriptional regulation. These results support the previous hypothesis that invasive species should have certain intrinsic traits that are superior in new locations to those of non-invasive species. Invasive insects tend to have high abilities to exploit nutrition and to defend themselves, as well as advanced chemosensory abilities. We proposed an invasiveness index to quantify invasiveness by weighting the abilities of the aforementioned four groups. This invasiveness index is the first metric for evaluating insect invasiveness at the genome level, and it should aid risk assessment and provide strong theoretical support for quarantine policy decisions [28].

We found that insects with high expansion indexes in all or most of the four groups tend to be highly invasive, whereas insects without high expansion indexes or with only one category with a high expansion index tend to be minimally invasive. This result appears consistent with commonsensical notions of invasive ability. However, it should be noted that this fact does not support the trade-off hypothesis, which assumes that invasive species must allocate limited energy and resources to either growth or defense [29]. That hypothesis suggests that having a trait with advantages for one function may simultaneously reduce the strength of other functions because of physical and chemical constraints, resource allocation limitations, antagonistic pleiotropy, and linkage disequilibrium [30]. However, our analysis shows that highly invasive insects may have most or even all four types of abilities including enhanced defensive, chemosensory and transcriptional regulation abilities, as well as high energy metabolism.

We applied an invasiveness index to evaluate 99 insect species that have annotated genome data. About half of these species were given a high invasiveness index value, and these species were also generally predicted to be highly invasive by DIGS, supporting the reliability of this metric. We noticed that some congeners were classified very differently in invasiveness. It is true that congeneric species even some cryptic species differ in some gene families, such as cytochrome P450 genes and UDP glycosyltransferases in two cryptic species of invasive whitefly, *Bemisia tabaci* Middle East-Asia Minor 1 (MEAM1, or 'B') and Mediterranean (MED, or 'Q') [31]. We emphasize that the present classifier was trained with a very small set of non-invasive species, these were all in just three orders, and that all the invasive species are in only five orders. It remains possible that the small sample size induced bias in feature selection and that limited phylogenetic diversity of species with adequate bases for classification as to invasiveness limits the domain of application of the striking result depicted in Fig. 3. Although the sample size for the negative training set was particularly small, the DIGS classifier still performed well, suggesting that invasive insects do have some inherent superiorities in new settings compared with non-invasive insects. These differences in inherent superiorities are reflected in the genome data. As the cost of genome sequencing decreases, more insect genomes will be available with high assembly quality. With additional genomes, larger positive and negative training datasets can be constructed to achieve better classification efficiency and to determine the extent to which our invasiveness index applies beyond the five orders for which we currently have data (the 99 species that we did not use for training are also all in the same five orders). A more robust classifier can then be trained with better, more robust performance. In addition, our approach to profiling invasive species used only the features of gene family expansions. As genomic information accumulates, other genome features such as single nucleotide polymorphisms (SNPs), copy number variants (CNV), and gene expression level at the subspecies level can be obtained and may be useful in predicting invasiveness, which should improve prediction accuracy and understanding of the molecular basis of invasiveness. Of course, the fact that a species truly has a tendency to become invasive once introduced does not mean that every introduction of

that species will lead to invasion. Aside from the fact that some probability exists for stochastic reasons alone that any population will be lost when very small (e.g., in the earliest stages of establishment), physical and biotic environmental factors differ among introductions and play some role in whether an invasion actually occurs. This fact is reflected in many examples of species that usually become invasive when introduced but fail to do so occasionally [32].

Materials and Methods

Genome resources and species selection

We downloaded 142 insect genome assemblies and the corresponding annotation data, including Coleoptera, Diptera, Hemiptera, Hymenoptera, and Lepidoptera, from the National Center for Biotechnology Information (NCBI) [33], InsectBase [34], VectorBase [35], Fireflybase (http://www.fireflybase.org/), Ensembl Genomes [36], GigaDB [37], Fourmidable [38], MonarchBase [39], and AphidBase [40] (Supplemental Table 1).

We used the genome characteristic Scaffold N50, which is positively associated with genome quality [41]. Species with a genome assembly with Scaffold N50 < 400 Kb were excluded. When a protein-coding gene has different alternative splicing forms, the longest transcript was chosen.

We analyzed 43 insect species, including 37 confirmed as invasive species by literature references, which were used as positive samples of invasive insects, and six confirmed as non-invasive by literature references, which were used as negative samples of non-invasive insects (Supplemental Table 2). The invasiveness index and the classifier were applied to the remaining 99 insects.

Gene family analysis

We used TreeFam [42], which considers phylogenetic relationship, to define gene families that

329 descended from a single gene of the most recent common ancestor. The annotated protein-
330 coding genes of 37 invasive insect species, six non-invasive insect species, and the additional
331 99 insect species were used for the application as noted above.
332
333 **Reconstruction of phylogenetic tree**
334 We performed phylogenetic analyses using proteins from all 43 invasive and non-invasive
335 species, and *Tetranychus urticae* was used as an outgroup. OrthoMCL [22] was used with
336 default parameters to identify gene groups based on sequence similarities resulting from an
337 all-against-all BLASTP search [21]. We found 183 single-copy orthologous genes shared by
338 all species. Multiple sequence alignments of orthologous genes from all species were
339 produced by MAFFT v7 [43] with default parameters, and the aligned results were trimmed by
340 trimAl [44] to remove low-quality regions with the parameter "-automated1". Finally, we
341 merged all 183 trimmed single-copy orthologous genes for each species to create a super
342 gene [45]. RAxML [46] was then used with the LG+I+F model, which is calculated with
343 ProtTest in IQ-TREE [47], to estimate a maximum likelihood tree starting with 1000 bootstraps
344 followed by likelihood optimization.
345
346 **Estimation of divergence time**
347 A nonparametric rate-smoothing method [48] and a semiparametric penalized likelihood (PL)
348 method [49] were used to estimate the divergence time with the software r8s (V1.7.1) [50]. An
349 optimal tree obtained by RAxML [46] was used as an input tree for the divergence time
350 estimation. The cross-validation approach (with parameters "cvstart=0, cvinc=1, cvnum=18")
351 was used to determine the optimal level of rate- smoothing of the PL analyses with smoothing
352 parameters varying from 1 to 1e17. We used a smoothing parameter of 1 for these data. To
353 estimate divergence time, we calculated ages of nodes within the phylogeny based on
354 calibration points. Our calibration points were: 1) the most recent common ancestor of the
355 clade including *Papilio polytes* and *Plutella xylostella*, constrained to be 140 Mya (million
356 years ago); 2) the most recent common ancestor of the clade including *Bombyx mori* and

Manduca sexta, constrained to be 39.8 Mya; and 3) the most recent common ancestor of the clade including *Aedes albopictus* and *Drosophila biarmipes*, constrained to be 157.8 Mya [45, 51].

Gene gain and loss

To identify gene family evolution as a stochastic birth and death process, we applied the likelihood model originally implemented in the software CAFÉ (v3.0) [23]. Phylogenetic tree topology and branch lengths were taken into account to infer the significance of change in gene family size in each branch. The gene number of gene families defined by TreeFam [42] in each insect and the phylogenetic tree corrected by r8s were used as input files for CAFÉ 3.0 [23].

Comparative analysis of genomic features

We calculated the genome features of all insects, including genome size, GC%, number of protein-coding genes, length of repetitive sequences, and gene number of expansion or contraction. We used t-tests (we also did permutation tests with the same result) to compare differences in genome features between all invasive and non-invasive insects, as well as fly and lepidopteran species separately. We identified repetitive sequences using the RepeatMasker [52] pipeline with “ncbi” set as the search engine and “insects” for the parameter (-species). In addition, RepBase [53] was provided as a custom library to locate associated repetitive elements in genomes of each species.

Commonly expanded gene families of invasive insects

The expanded and contracted gene families in each species, as well as the expanded and contracted gene numbers of each gene family, were extracted by using an in-house Perl script and calculated by comparing a species with its parent node in the phylogenetic tree (Supplementary Table 3). The commonly expanded gene families of invasive insects were screened with two criteria: 1) expanded in at least one-third of invasive species (>=13), or

less than or equal to half of the total number of non-invasive species (<= 3). 2) an expanding ratio, defined as the ratio of the number of invasive species to the number of non-invasive species in which the gene family is expanded, greater than 12 (the criterion was determined by testing the accuracy of invasiveness classification for a range of ratios from 4 to 13, among them, ratio of 12 achieved the highest accuracy). This protocol generated a number of candidate gene families that might be related to invasiveness. Next, we annotated these candidate gene families using the corresponding protein sequences as queries to perform a BLASTP [21] search (e-value cutoff of 1e-5) against the UniProt [54] database. We then grouped these gene families according to their annotated functions.

Gene families associated with invasiveness and expansion index

To estimate the contribution of these gene families to invasiveness, we used the candidate gene families from each function group to train the first-step logistic regression model with 70% of the 43 species as the training set and the other 30% as the testing set. The partition was randomly done 30 times. The coefficient of each gene family in the logistic regression was regarded as its weight coefficient for invasiveness within the function group. If a gene family has a non-positive weight coefficient for more than one third of the partitions, we removed it and performed the same pipeline again until all remaining gene families had positive weight coefficients for more than half of the partitions. We next selected the model with the highest AUC [25] among all partitions for each function group. Finally, we used 14 gene families from four function groups and their weight coefficients to construct the expansion indexes (Supplementary table 4). The expansion index formula for a specific function group containing *n* expanded gene families associated with invasiveness is defined as follows：

$$y_j = \frac{\sum_{i=1}^{n} k_i x_{ij}}{\sum_{i=1}^{n} k_i}$$

where y_j is the expansion index of the function group for the *j*th species, *i* is the number of the *i*th invasiveness-related gene family in the function group, *n* is the total number of invasiveness-related gene families in the function group, x_{ij} is the expanded gene number of

the *i*th invasiveness-related gene family in the function group of the *j*th species, and k_i is the weight coefficient, i.e., the logistic regression coefficient, of the *i*th invasiveness-related gene family to invasiveness.

Invasiveness index formula

To estimate the weight coefficients of these expansion indexes of the five function groups to invasiveness, we used them to train the second-step logistic regression model with 70% of the 43 species as the training set and the remaining 30% as the testing set. A model with highest AUC [25] was fitted. Subsequently, the expansion indexes and their corresponding weight coefficients as well as the intercept in the logistic regression model were used to construct the invasiveness index formula in three steps:

1) $m_j = \sum_{i=1}^{g} k_i y_{ij} + b$,

2) $$n_j = \begin{cases} \log_{10}(|m_j|), & if \ \ m_j \geq 1 \\ 0, & if \ \ -1 < m_j < 1 \\ -\log_{10}(|m_j|), & if \ \ m_j \leq -1 \end{cases},$$

3) $z_j = 1 - \frac{1}{1 + e^{n_j}}$,

where z_j is the invasiveness index of the *j*th species, y_{ij} is the expansion index of the *i*th function group of the *j*th species, *g* is the total number of function groups (in this study, four), k_i is the weight coefficient of the *i*th function group, and *b* is the constant in the logistic regression. We used the first step to calculate the total weight m_j of four function groups that contribute to the invasiveness classification. The second step was used to normalize the m_j; and we calculated the invasiveness index by a logistic formula in third step.

Applying the invasiveness index formula to estimate invasiveness

To calculate the invasiveness indexes of the other 99 insects, we added one of the 99 species to the data set of the 44 species analyzed (37 invasive insects, six non-invasive insects, and the outgroup *Tetranychus urticae*) at a time, then the same methods and parameters were applied to find their single-copy orthologous genes, construct the phylogenetic tree, correct divergence time, and calculate gene gain and loss for each species. Finally, we calculated

440 invasiveness indexes of all of the 99 species using the invasiveness index formula from the
441 above section (Supplementary Table 5).
442
443 **Invasiveness classification by the machine-learning algorithm**
444 A machine-learning algorithm named Determining Invasiveness based on Genome
445 Sequences (DIGS) was built for invasiveness classification.
446
447 First, we used a random forest algorithm for feature selection and used a logistic regression to
448 estimate the classification performance of features selected by this algorithm. Six-fold cross-
449 validation was used to estimate the accuracy and stability of feature selection. To guarantee
450 that each of the six non-invasive species would be allocated to the testing set once, only one
451 non-invasive species was allocated to the testing set in each iteration of cross-validation. The
452 remaining five non-invasive species were allocated to the training set. According to the ratio
453 of 1:5 to allocate species into testing and training sets of non-invasive species, in each
454 iteration of cross-validation, the 37 invasive species were randomly distributed into six groups
455 (each group has six or seven species), one group (six or seven species) was allocated to the
456 testing set, while the remaining five groups (a total of 31 or 30 species) were allocated to the
457 training set. The R package “Boruta” [24] was used to perform feature selection with the
458 parameters of “ntree=1000” and “maxRuns=1000”, using all 36 candidate gene families of the
459 training set. This algorithm selects features with a random forest classification algorithm and a
460 statistical test. Features that do not contribute more to the classification information than
461 random features were removed. We used the expanded or contracted gene numbers of
462 species as the input data. Two features were stably confirmed to be important for
463 invasiveness by Boruta with six-fold cross-validation (Supplementary Table 6).
464
465 We then used the three gene families confirmed as important to invasiveness by previous
466 steps to construct the logistic regression model as the DIGS classifier. To treat sample size
467 imbalance, the entire dataset of negative samples was duplicated six times before being used

for training or testing in the logistic regression model [55] as follows:

$$y_j = 1 - \frac{1}{1 + e^{\sum_{i=1}^{n} k_i x_{ij} + b}},$$

where y_j is the probability that the *j*th species belongs to the invasive set, *i* is the number of *i*th invasiveness-related gene family confirmed by Boruta, *n* is the total number of invasiveness-related gene families confirmed by Boruta (here n=2), x_{ij} is the expanded gene number of the *i*th invasiveness-related gene family of the *j*th species, k_i is the weight coefficient of the *i*th invasiveness-related gene family in the classification model. By the six-fold cross-validation in DIGS, two features were stably estimated to associate with insect invasiveness, x_{1j} (pao retrotransposon peptidase) and x_{2j} (putative nuclease HARBI1) with the corresponding coefficients of 0.31 and 1.86, respectively; *b* was equal to 1.20.

Next, we used the DIGS classifier to calculate the probabilities that the species in the testing set are invasive. A species was predicted to be invasive if the probability exceeded 0.5 by DIGS; otherwise it was predicted not to be invasive.

Abbreviations

DIGS: Determining Invasiveness based on Genome Sequences; SNPs: single nucleotide polymorphisms; RAxML: Random axelerated maximum likelihood; CNV: copy number variants; ISPS: Invasive Species Predictive Schemes; SCOPE: Scientific Committee on Problems of the Environment; BLAST: Basic Local Alignment Search Tool; CAFÉ: Computational Analysis of gene Family Evolution; ROC: Receiver Operating Characteristic; AUC: Area under the Curve of ROC; GC: Guanine and Cytosine nucleotides; IAS: Invasive Alien Species; P450: Cytocrome P450; UDP: Uridine Diphosphate; MEAM1: Middle East-Asia Minor 1; MED: Mediterranean.

Acknowledgements

Not applicable.

Authors' contributions

F.L. conceived the work and designed the experiment plan; F.W., D.S., and W.Q. designed and improved the experiment plans. D.S., N.Y., and C.H. determined the invasive and non-invasive insects by reference mining; C.H., N.Y., and L.X. collected the genome data; C.H. carried out machine learning classification of invasiveness; X.F., C.P., J.L., K.L., and X.L. participated in the discussion of machine learning work. S.W., W.Q, L.X., M.J., and W.L participated in the discussion of insect invasiveness. F.L., N.Y., C.H., D.S., and F.W. wrote the manuscript. All authors have read and approved the manuscript.

Funding

This work was supported by the National Key Research and Development Project of China [2016YFC1200600, 2016YFC1201200, 2017YFC1200600]. The funders had no role in study design, data collection, and analysis, or in the decision to publish or in preparing the manuscript.

Availability of data and materials

All genomic data used in this study could be downloaded from the databases which have been listed in the supplementary table 1.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflict of interest related to the results reported in this study.

524 **References**

525 1. Bradshaw CJ, Leroy B, Bellard C, Roiz D, Albert C, Fournier A, Barbet-Massin M, Salles
526 JM, Simard F, Courchamp F: **Massive yet grossly underestimated global costs of**
527 **invasive insects**. *Nat Commun* 2016, **7**:12986.
528 2. Early R, Bradley BA, Dukes JS, Lawler JJ, Olden JD, Blumenthal DM, Gonzalez P,
529 Grosholz ED, Ibanez I, Miller LP *et al*: **Global threats from invasive alien species in**
530 **the twenty-first century and national response capacities**. *Nature Communications*
531 2016, **7**.
532 3. Paini DR, Sheppard AW, Cook DC, De Barro PJ, Worner SP, Thomas MB: **Global threat**
533 **to agriculture from invasive species**. *P Natl Acad Sci USA* 2016, **113**(27):7575-7579.
534 4. Moravcova L, Pysek P, Jarosik V, Havlickova V, Zakravsky P: **Reproductive**
535 **characteristics of neophytes in the Czech Republic: traits of invasive and non-**
536 **invasive species**. *Preslia* 2010, **82**(4):365-390.
537 5. Pyšek P, Richardson DM: **Traits associated with invasiveness in alien plants: where**
538 **do we stand?** In: *Biological invasions.* Edited by Caldwell MM, Heldmaier G, Jackson
539 RB, Lange OL, Mooney HA, Schulze ED, Sommer U. Berlin Heidelberg: Springer; 2008:
540 97-125.
541 6. Sol D, Maspons J, Vall-Llosera M, Bartomeus I, Garcia-Pena GE, Pinol J, Freckleton RP:
542 **Unraveling the life history of successful invaders**. *Science* 2012, **337**(6094):580-583.
543 7. Cote J, Fogarty S, Weinersmith K, Brodin T, Sih A: **Personality traits and dispersal**
544 **tendency in the invasive mosquitofish (*Gambusia affinis*)**. *P Roy Soc B-Biol Sci* 2010,
545 **277**(1687):1571-1579.
546 8. Ochocki BM, Miller TEX: **Rapid evolution of dispersal ability makes biological**
547 **invasions faster and more variable**. *Nature Communications* 2017, **8**.
548 9. Colautti RI, Ricciardi A, Grigorovich IA, MacIsaac HJ: **Is invasion success explained**
549 **by the enemy release hypothesis?** *Ecol Lett* 2004, **7**(8):721-733.
550 10. Callaway RM, Ridenour WM: **Novel weapons: invasive success and the evolution of**
551 **increased competitive ability**. *Front Ecol Environ* 2004, **2**(8):436-443.

11. Vilcinskas A, Stoecker K, Schmidtberg H, Rohrich CR, Vogel H: **Invasive harlequin ladybird carries biological weapons against native competitors**. *Science* 2013, **340**(6134):862-863.
12. Sax DF, Brown JH: **The paradox of invasion**. *Global Ecol Biogeogr* 2000, **9**(5):363-371.
13. Alpert P, Bone E, Holzapfel C: **Invasiveness, invasibility and the role of environmental stress in the spread of non-native plants**. *Perspectives in plant ecology, evolution and systematics* 2000, **3**(1):52-66.
14. Whitney KD, Gabler CA: **Rapid evolution in introduced species, 'invasive traits' and recipient communities: challenges for predicting invasive potential**. *Divers Distrib* 2008, **14**(4):569-580.
15. Blackburn TM, Pysek P, Bacher S, Carlton JT, Duncan RP, Jarosik V, Wilson JRU, Richardson DM: **A proposed unified framework for biological invasions**. *Trends Ecol Evol* 2011, **26**(7):333-339.
16. Capellini I, Baker J, Allen WL, Street SE, Venditti C: **The role of life history traits in mammalian invasion success**. *Ecol Lett* 2015, **18**(10):1099-1107.
17. Rius M, Bourne S, Hornsby HG, Chapman MA: **Applications of next-generation sequencing to the study of biological invasions**. *Curr Zool* 2015, **61**(3):488-504.
18. Swarts K, Gutaker RM, Benz B, Blake M, Bukowski R, Holland J, Kruse-Peeples M, Lepak N, Prim L, Romay MC *et al*: **Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America**. *Science* 2017, **357**(6350):512-515.
19. Kim S, Cho YS, Kim HM, Chung O, Kim H, Jho S, Seomun H, Kim J, Bang WY, Kim C *et al*: **Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly**. *Genome Biol* 2016, **17**.
20. Lippert C, Sabatini R, Maher MC, Kang EY, Lee S, Arikan O, Harley A, Bernal A, Garst P, Lavrenko V *et al*: **Identification of individuals by trait prediction using whole-genome sequencing data**. *P Natl Acad Sci USA* 2017, **114**(38):10166-10171.
21. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL:

BLAST+: architecture and applications. *BMC Bioinformatics* 2009, **10**:421.

22. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: Identification of ortholog groups for eukaryotic genomes**. *Genome Res* 2003, **13**(9):2178-2189.

23. Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW: **Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3**. *Mol Biol Evol* 2013, **30**(8):1987-1997.

24. Kursa MB, Rudnicki WR: **Feature selection with the Boruta package**. *J Stat Softw* 2010, **36**(11):1-13.

25. Ling CX, Huang J, Zhang H: **AUC: a statistically consistent and more discriminating measure than accuracy**. In: *IJCAI: 2003*; 2003: 519-524.

26. De Roy K, Marzorati M, Negroni A, Thas O, Balloi A, Fava F, Verstraete W, Daffonchio D, Boon N: **Environmental conditions and community evenness determine the outcome of biological invasion**. *Nature Communications* 2013, **4**.

27. Chown SL, Hodgins KA, Griffin PC, Oakeshott JG, Byrne M, Hoffmann AA: **Biological invasions, climate change and genomics**. *Evolutionary Applications* 2015, **8**(1):23-46.

28. Worner SP, Gevrey M: **Modelling global insect pest species assemblages to determine risk of invasion**. *J Appl Ecol* 2006, **43**(5):858-867.

29. Tilman D: **Causes, consequences and ethics of biodiversity**. *Nature* 2000, **405**(6783):208-211.

30. Jessup CM, Bohannan BJM: **The shape of an ecological trade-off varies with environment**. *Ecol Lett* 2008, **11**(9):947-959.

31. Xie W, Yang X, Chen CH, Yang ZZ, Guo LT, Wang D, Huang JQ, Zhang HL, Wen YN, Zhao JY *et al*: **The invasive MED/Q *Bemisia tabaci* genome: a tale of gene loss and gene gain**. *Bmc Genomics* 2018, **19**.

32. Zenni RD, Nunez MA: **The elephant in the room: the role of failed invasions in understanding invasion biology**. *Oikos* 2013, **122**(6):801-815.

33. Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bourexis D, Brister JR, Bryant SH, Canese K *et al*: **Database resources of the National Center for**

608 **Biotechnology Information**. *Nucleic Acids Res* 2018, **46**(D1):D8-D13.
609 34. Yin C, Shen G, Guo D, Wang S, Ma X, Xiao H, Liu J, Zhang Z, Liu Y, Zhang Y *et al*:
610 **InsectBase: a resource for insect genomes and transcriptomes**. *Nucleic Acids Res*
611 2016, **44**(D1):D801-807.
612 35. Giraldo-Calderon GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, Ho N,
613 Gesing S, VectorBase C, Madey G *et al*: **VectorBase: an updated bioinformatics**
614 **resource for invertebrate vectors and other organisms related with human**
615 **diseases**. *Nucleic Acids Res* 2015, **43**(Database issue):D707-713.
616 36. Kersey PJ, Allen JE, Allot A, Barba M, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen
617 M, Davis P, Grabmueller C *et al*: **Ensembl Genomes 2018: an integrated omics**
618 **infrastructure for non-vertebrate species**. *Nucleic Acids Res* 2018, **46**(D1):D802-
619 D808.
620 37. Sneddon TP, Li P, Edmunds SC: **GigaDB: announcing the GigaScience database**.
621 *Gigascience* 2012, **1**:11.
622 38. Wurm Y, Uva P, Ricci F, Wang J, Jemielity S, Iseli C, Falquet L, Keller L: **Fourmidable:**
623 **a database for ant genomics**. *Bmc Genomics* 2009, **10**:5.
624 39. Zhan S, Reppert SM: **MonarchBase: the monarch butterfly genome database**.
625 *Nucleic Acids Res* 2013, **41**(D1):D758-D763.
626 40. Legeai F, Shigenobu S, Gauthier JP, Colbourne J, Rispe C, Collin O, Richards S, Wilson
627 ACC, Murphy T, Tagu D: **AphidBase: a centralized bioinformatic resource for**
628 **annotation of the pea aphid genome**. *Insect Mol Biol* 2010, **19**:5-12.
629 41. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz
630 P, Tyson GW: **Recovery of nearly 8,000 metagenome-assembled genomes**
631 **substantially expands the tree of life**. *Nat Microbiol* 2017, **2**(11):1533-1542.
632 42. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li RQ, Liu T, Zhang Z,
633 Bolund L *et al*: **TreeFam: a curated database of phylogenetic trees of animal gene**
634 **families**. *Nucleic Acids Res* 2006, **34**:D572-D580.
635 43. Katoh K, Standley DM: **MAFFT Multiple sequence alignment software version 7:**

improvements in performance and usability. *Mol Biol Evol* 2013, **30**(4):772-780.

44. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T: **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses**. *Bioinformatics* 2009, **25**(15):1972-1973.

45. Misof B, Liu SL, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG *et al*: **Phylogenomics resolves the timing and pattern of insect evolution**. *Science* 2014, **346**(6210):763-767.

46. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies**. *Bioinformatics* 2014, **30**(9):1312-1313.

47. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ: **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies**. *Mol Biol Evol* 2015, **32**(1):268-274.

48. Sanderson MJ: **A nonparametric approach to estimating divergence times in the absence of rate constancy**. *Mol Biol Evol* 1997, **14**(12):1218-1231.

49. Sanderson MJ: **Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach**. *Mol Biol Evol* 2002, **19**(1):101-109.

50. Sanderson MJ: **r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock**. *Bioinformatics* 2003, **19**(2):301-302.

51. McKenna DD, Scully ED, Pauchet Y, Hoover K, Kirsch R, Geib SM, Mitchell RF, Waterhouse RM, Ahn SJ, Arsala D *et al*: **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 Biol* 2016, **17**(1):227.

52. Smit AF, Hubley R, Green P: **2010 RepeatMasker Open-3.0**. *URL: http://www repeatmasker org* 1996.

53. Bao W, Kojima KK, Kohany O: **Repbase Update, a database of repetitive elements in eukaryotic genomes**. *Mobile DNA* 2015, **6**:11.

54. UniProt Consortium T: **UniProt: the universal protein knowledgebase**. *Nucleic Acids*

Res 2018, **46**(5):2699.

55. Maloof MA: **Learning when data sets are imbalanced and when costs are unequal and unknown**. In: *ICML-2003 workshop on learning from imbalanced data sets II: 2003*; 2003: 2-1.

Figures:

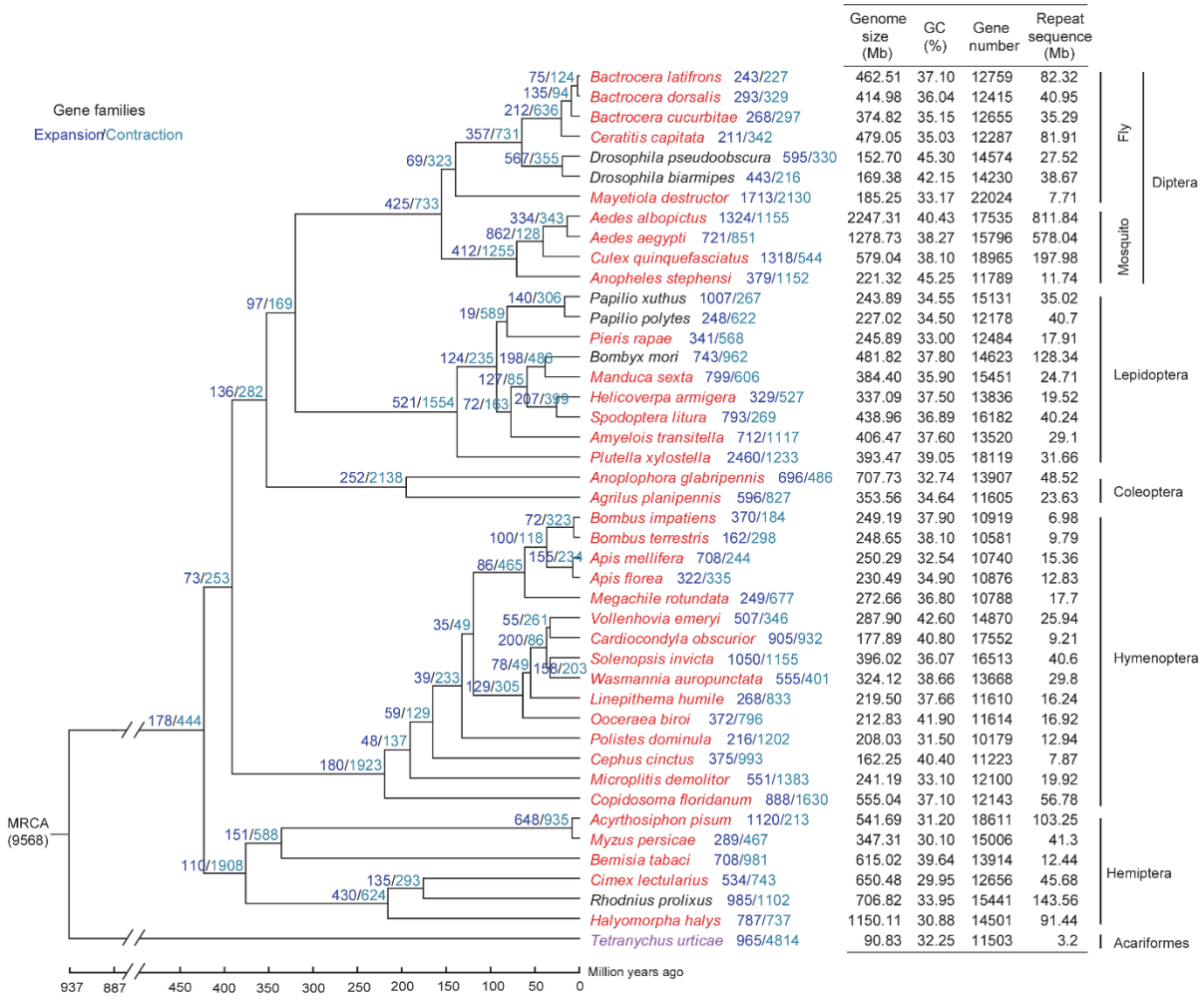


Figure 1. Phylogenetic tree and comparison of general genome features. The phylogenetic tree shows the topology and divergence time for 44 arthropods. The mite *Tetranychus urticae* was used as an outgroup. Numbers at branches and tips indicate the number of gene families that are expanded (blue color) or contracted (green color) as compared to the closest tip. MRCA = most recent common ancestor. The number in parentheses is the number of gene families in the MRCA as estimated with TreeFam software. Differences in genome size, GC content, gene number, and amount of repeat sequences between all invasive and non-invasive species, as well as between the corresponding fly species and Lepidoptera each analyzed separately, are not significant by t-test.

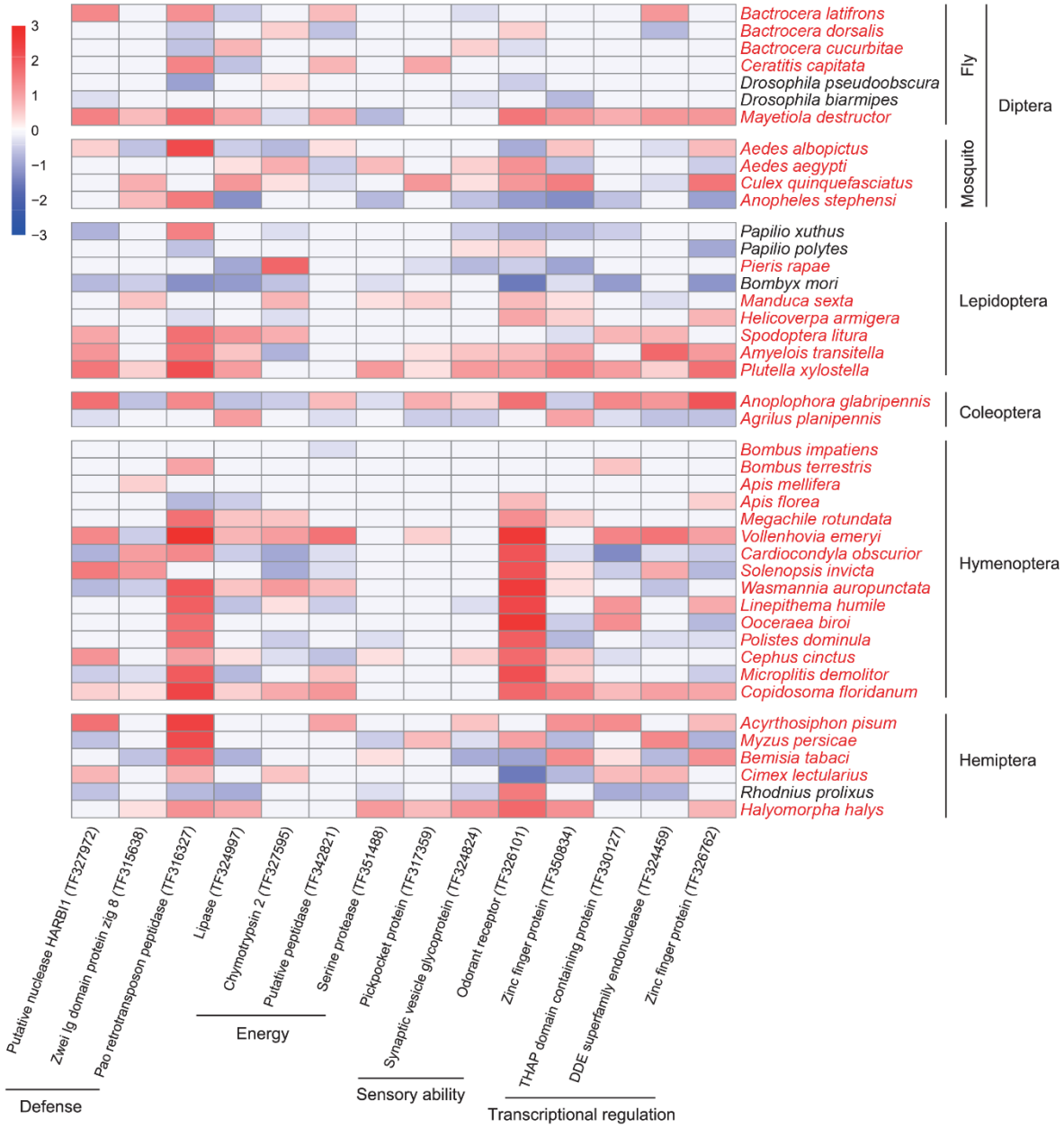
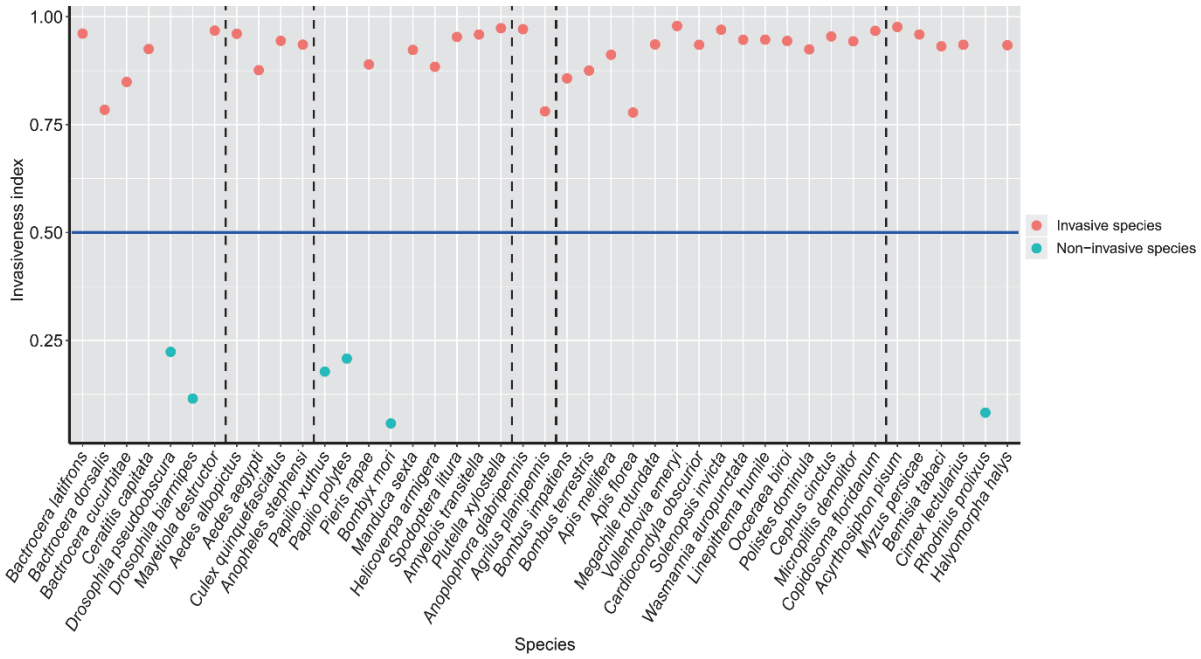


Figure 2. The comparison of expanded and contracted gene number in invasiveness-related gene families between invasive species (red lettering) and non-invasive species (black lettering). The expansion and contraction gene numbers were converted by log10. y = log10(|x|) (x≥1) or y = -log10(|x|) (x≤-1), where x represents the expanded gene number (x≥1) or the contracted gene number (x≤-1) and y was used in the heatmap.



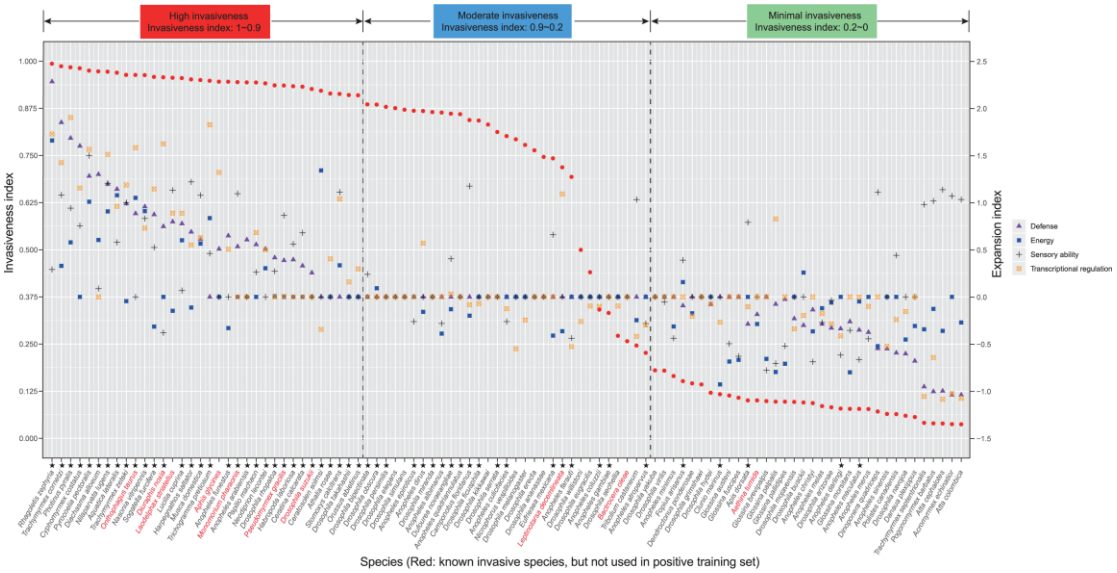
686

687 **Figure 3.** Invasiveness indexes of all forty-three species. Red dots represent invasive insects

688 and green dots represent non-invasive insects. The dashed lines separate the species into

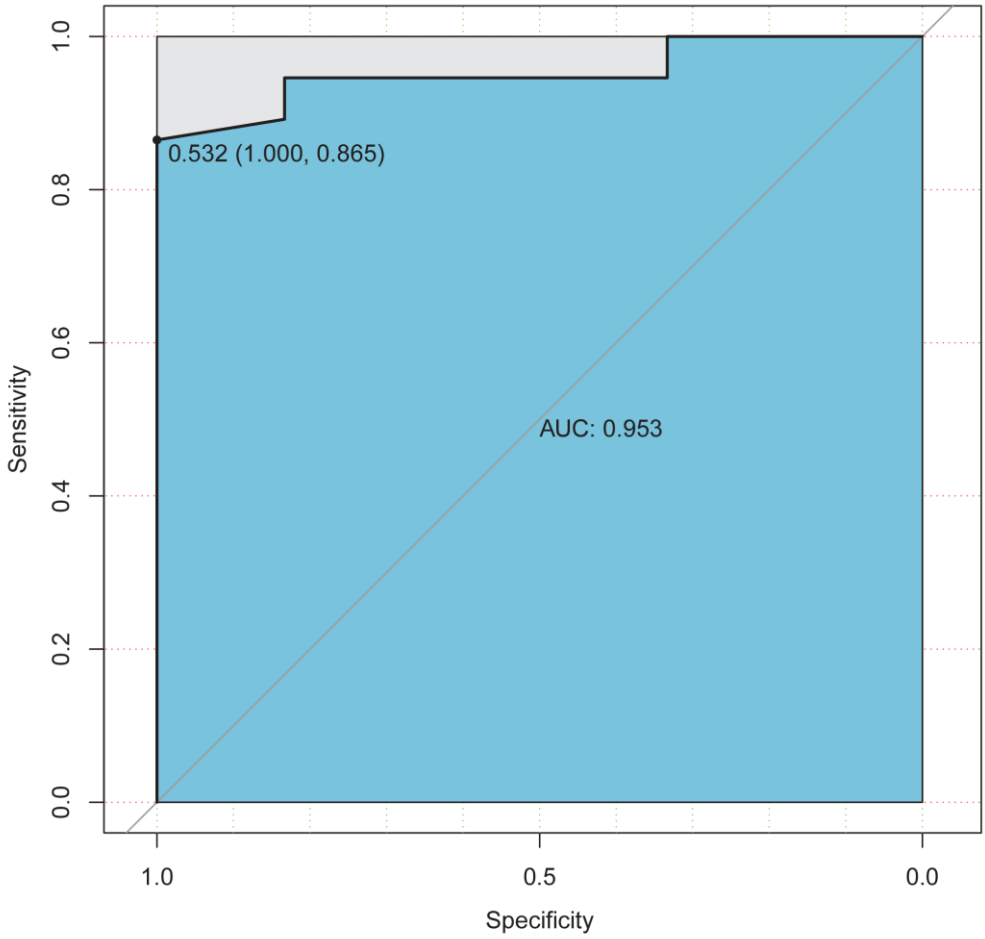
689 different taxa: fly, mosquito, Lepidoptera, Coleoptera, Hymenoptera, and Hemiptera from left

690 to right. The blue solid line represents the cutoff value of 0.5.



691

692 **Figure 4.** Invasiveness indexes and gene family expansion indexes of the other 99 insect
693 species. The symbol "●" represents the invasiveness index. Three levels of invasiveness
694 (high, moderate, and low) were classified by the invasiveness index cutoffs at 0.9 and 0.2.
695 Fourteen identified invasiveness-related gene families are categorized into four function
696 groups as defense, energy, chemosensory function, and transcriptional regulation. The
697 symbol "▲" represents the expansion index of gene families in defense function group, the
698 symbol "■" represents the expansion index of gene families in the energy function group, the
699 symbol "+" represents the expansion index of gene families in the chemosensory function
700 group, and the symbol "⊠" represents the expansion index of gene families in the
701 transcriptional regulation function group. The species in red lettering were confirmed to be
702 invasive but excluded in the 43-species sample set because of their relatively low-quality
703 genome assemblies (a scaffold N50 < 400 Kb), while the ones in black were species with no
704 evidence to confirm them as either invasive nor non-invasive (generally because they have
705 not been confirmed to have been introduced anywhere). The symbol "★" above a species'
706 scientific name means this species was predicted to be invasive by DIGS (Determine
707 Invasiveness based on Genome Sequences).

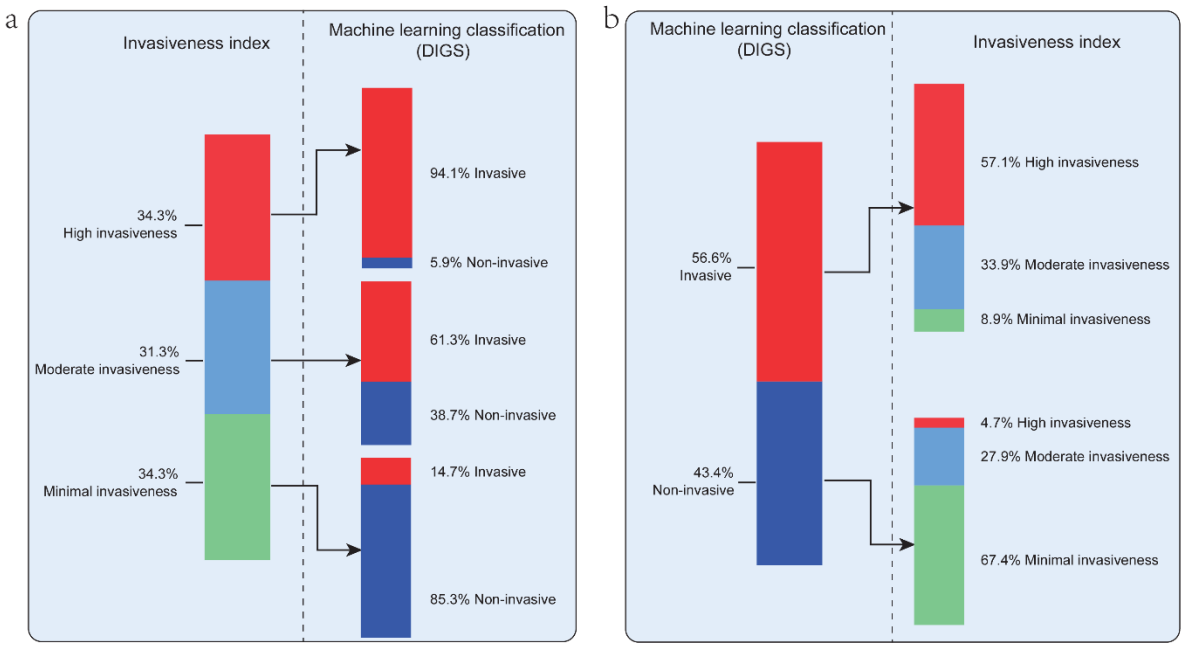


708

709 **Figure 5.** The ROC curve of the DIGS. ROC curve: the receiver operating characteristic

710 curve, AUC: the area under the curve of ROC, sensitivity: true positive rate, specificity: true

711 negative rate, DIGS: Determine Invasiveness based on Genome Sequences.



712

713 **Figure 6.** Invasiveness classification by DIGS and invasiveness level assessment by

714 invasiveness index. a) The percentage of invasive and non-invasive species classified by

715 DIGS in each invasiveness level assessed by the invasiveness index was calculated. b) The

716 proportions of species with different levels of invasiveness in invasive and non-invasive

717 categories classified by DIGS is shown.