SCIENTIFIC REPORTS

Molecular classification based on apomorphic amino acids (Arthropoda, Hexapoda): Integrative taxonomy in the era of phylogenomics

Hao-Yang Wu[1],[*], Yan-Hui Wang[1],[2],[*], Qiang Xie[1],[*], Yun-Ling Ke[3] & Wen-Jun Bu[1]

With the great development of sequencing technologies and systematic methods, our understanding of evolutionary relationships at deeper levels within the tree of life has greatly improved over the last decade. However, the current taxonomic methodology is insufficient to describe the growing levels of diversity in both a standardised and general way due to the limitations of using only morphological traits to describe clades. Herein, we propose the idea of a molecular classification based on hierarchical and discrete amino acid characters. Clades are classified based on the results of phylogenetic analyses and described using amino acids with group specificity in phylograms. Practices based on the recently published phylogenomic datasets of insects together with 15 *de novo* sequenced transcriptomes in this study demonstrate that such a methodology can accommodate various higher ranks of taxonomy. Such an approach has the advantage of describing organisms in a standard and discrete way within a phylogenetic framework, thereby facilitating the recognition of clades from the view of the whole lineage, as indicated by PhyloCode. By combining identification keys and phylogenies, the molecular classification based on hierarchical and discrete characters may greatly boost the progress of integrative taxonomy.

Taxonomy is the science of classifying, describing, identifying, and naming. With interdisciplinary endeavours becoming increasingly common in biological research, an accurate and stable classification system is prerequisite due to its role as one of the cornerstones for the integration of multiple fields. The precise description of a clade can provide both information from the past and inspiration for the future. Over the past two centuries, developments in the life sciences have involved the concomitant adjustment of the hierarchical and binomial classification systems developed by Carolus Linnaeus. Among these, perhaps the most important ontological change of the original system is that biological classifications should now be phylogenetic, *i.e.*, each group that is recognised in a classification should be monophyletic[1–4]. With the purpose of defining and naming clades in the tree of life with more explicit reference to a phylogeny, the idea of phylogenetic nomenclature has been developed that uses the phylogenetic definitions of taxa and a rank-free classification system[1,5–8]. In the context of phylogenetic nomenclature, the information provided by rank-signifying ending is limited. In addition, supraspecific names are not always explicitly associated with clades under the rank-based codes, resulting in ambiguous definitions and an impediment in the clear communication and efficient storage and retrieval of biological information. Therefore, some rank-free phylogenetic classifications[1,5–8] have been suggested to replace rank-based ones. Instead of a total negation, a series of modifications of the classic system with hierarchy have been proposed by the opponents of phylogenetic nomenclature. For example, Platnick advocated an extension of the standard rank-based classification through the implementation of rank-based definitions to the names of clades at all hierarchical levels[9],

[1]Institute of Entomology, College of Life Sciences, Nankai University, Tianjin 300071, China. [2]College of Computer and Control Engineering, Nankai University, 38 Tongyan Road, Haihe Education Park, Jinnan District, Tianjin 300350, China. [3]Guangdong Entomological Institute, Guangzhou 510260, China. [*]These authors contributed equally to this work. Correspondence and requests for materials should be addressed to Q.X. (email: qiangxie@nankai.edu.cn)

and Ward suggested maintaining a ranked phylogenetic taxonomy, at least for groups in relatively recent and species-rich branches of the tree of life[10].

Setting the controversy on the ontological issues of biological entities aside, both the proponents and opponents of the phylogenetic nomenclature have agreed that the current rank-based classification system cannot meet the need for describing the growing levels of evolutionary divergences revealed by the great advancement of phylogenetics. Indeed, from the perspective on standardisation and efficiency, such a challenge also necessitates the development of taxonomic epistemology and methodology. For example, the traditional classification systems for both animals and plants are mainly based on morphological characteristics. Therefore, for practical considerations, clades that can be named should be both strongly supported as monophyletic groups in phylogenetic analyses and have distinctive phenotypic features that allow them to be distinguished from related taxa[10], *i.e.*, correctly classified and precisely described. However, not all of the clades that we would like to study have such definitive diagnostic morphological traits. Furthermore, due to the continuous variation of many morphological traits, for many clades, it seems inevitable to employ morphological definitions that involve unique and conditional combinations of traits rather than clear and unequivocal synapomorphies[10]. In addition, because there are only a few homologous morphological traits at higher category levels due to the difficulty in establishing general ground-plans, it is hard to apply a classification system that allows for meaningful comparisons in different groups of organisms. In all, the way thus far adopted for descriptions and diagnoses of clades restrict the further development of taxonomy.

With the great advance of high-throughput sequencing technology in the last decade, it has become possible to rapidly and economically acquire large amounts of genome or transcriptome sequences. The steadily declining sequencing costs make it no longer inhibitory to analyse transcriptomes or even whole genomes, which can boost the development of a molecular-based taxonomic classification system. Currently, the major approaches that use huge molecular data sets to describe clades are based on similarity. For example, Marakeby *et al.*[11] proposed an exclusively genome-based classification and naming system, in which the proposed organism codes are assigned based on measured similarity. However, a similarity-based classification system may not accurately reflect evolutionary relationships, *i.e.*, a code may be assigned to a paraphyletic or polyphyletic clade rather than a monophyletic one, which violates the broad consensus of current biological classification. Meanwhile, with the discoveries of many lineage-specific nucleotide/amino acid residues in comparative studies, a series of character-based approaches have also been proposed, *e.g.*, ribosomal multilocus sequence typing, which assigns bacteria to genetic lineages that have identical alleles at certain genomic loci[12]. Nevertheless, the application of these approaches is restricted to certain groups, and the relationships between different sequence types are uninformative due to the lack of information on hierarchy.

Herein, we propose a hierarchical character-based molecular classification that uses group-specific traits of multilocus genes as the description and diagnoses of clades. In this classification frame, only characters with definite diagnosability can be candidates for the description of clades, *i.e.*, the amino acids or nucleotides that differ among organisms will not be treated as equal. Each clade in a lineage is given a unique code that is derived from a mining of group-specific apomorphies, *i.e.*, a trait that is found in some or all terminal groups of a clade and is inherited from a common ancestor[13]. A strategy based on the criterion of parsimony is introduced to detect apomorphies in a dataset. The codes are then arranged in a queue based on the taxonomic hierarchy and finally generate the complete diagnostic description of a terminal taxon (species) in the tree of life. The data from a recent published phylogenomic study about the evolution of hexapods[14] and from a comparative genomic study of Anophelinae mosquitoes (Diptera: Culicidae)[15] were chosen to generate two datasets for this study based on two reasons. On the one hand, insects are among the most diverse organisms on the planet, representing approximately half of all known living organisms[16]. On the other hand, several insect clades had been seen as "problematic" in the phylogenetic studies based on single or a few gene markers (*e.g.*, Polyneoptera), which makes it ideal for testing the feasibility of our system. Thousands of apomorphies were found under a strict filter criterion, covering nearly all of the nodes in the original studies. The results of the phylogenetic reconstructions based on such sites showed abundant phylogenetic signals. Two classification systems were constructed according to the category levels between the superclass and the order and between the subfamily and the species complex. Our study provides a cladistic approach to reanalyse the molecular sequence data of a phylogenomic study and shows the potential of a synthesis in systematic biology, whereby phylogenomics, molecular classification, and PhyloCode may be integrated.

Results

Apomorphy mining and the subset optimisation. After apomorphy mining with filter criteria based on various consistency index (CI) values, we generated several sub-datasets. For the dataset of Hexapoda, a total of 7,939, 8,008, and 11,241 apomorphies were identified to generate sub-datasets 1A, 1B, and 1C, respectively. For the Anophelinae dataset, a total of 422, 422, and 464 apomorphies were identified to generate sub-datasets 2A, 2B, and 2C, respectively. The phylogenetic trees inferred from the six sub-datasets are shown in Figs S1 and S2. For the phylogenetic inference on Hexapoda, only the tree based on sub-dataset 1B retrieved the same topology as the original. Meanwhile, for the phylogenetic reconstruction of Anophelinae, both the topology and support values obtained from sub-datasets 2A, 2B and 2C were acceptable. As a result, considering both phylogenetic reappearance and the informativeness for apomorphy mining, sub-datasets 1B and 2C were used for subsequent analyses.

Building the code-system of Hexapoda using molecular apomorphies. From the potential apomorphies in sub-datasets 1B and 2C, a series of sites were selected based on a group of optimal criteria (Fig. 1) to construct two classification systems according to category levels between the class and order in Hexapoda and between the subfamily and species complex in Anophelinae. The results are shown in Figs 2a and 3a, which correspond to the topologies revealed by phylogenomics. The sequential arrangements of codes in two-dimensional
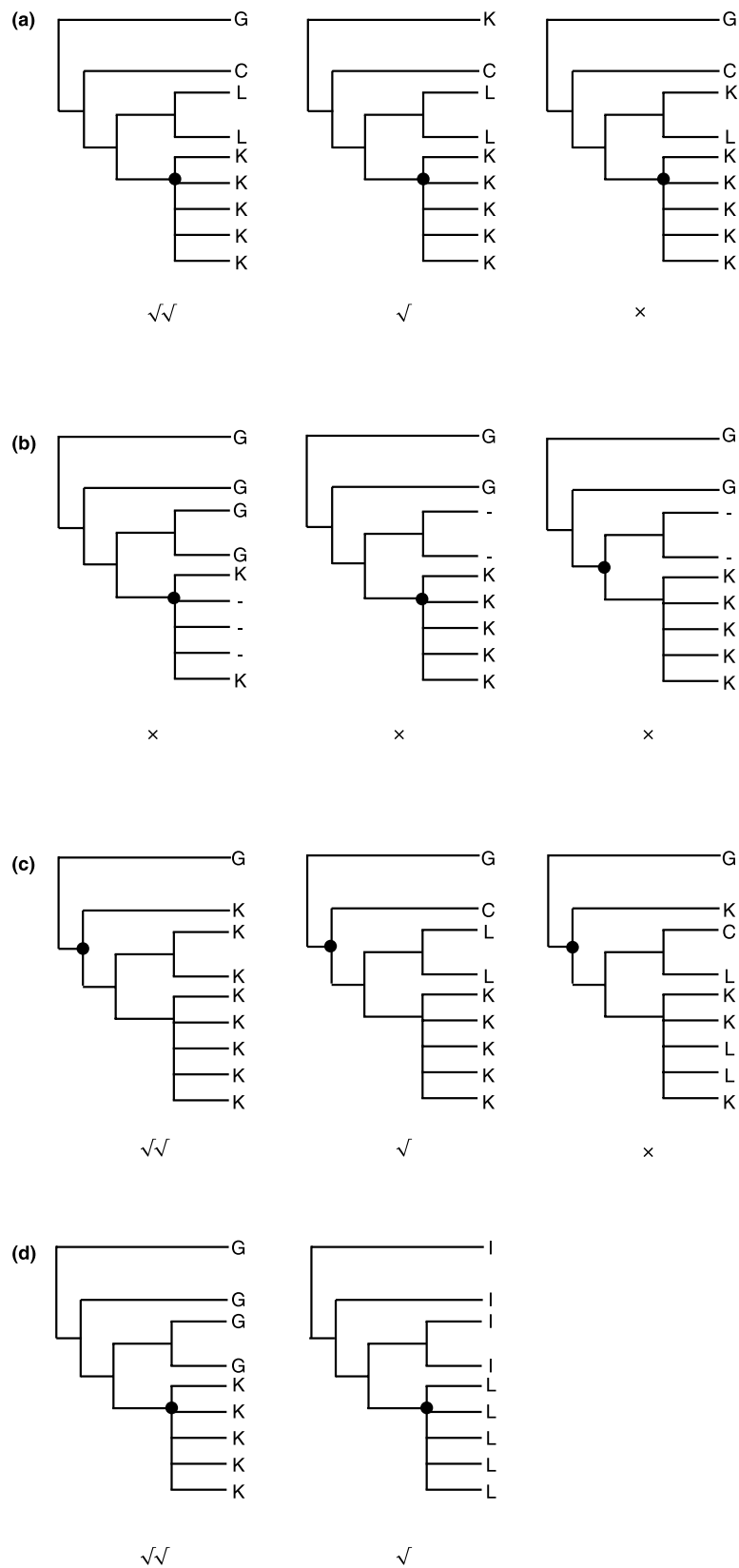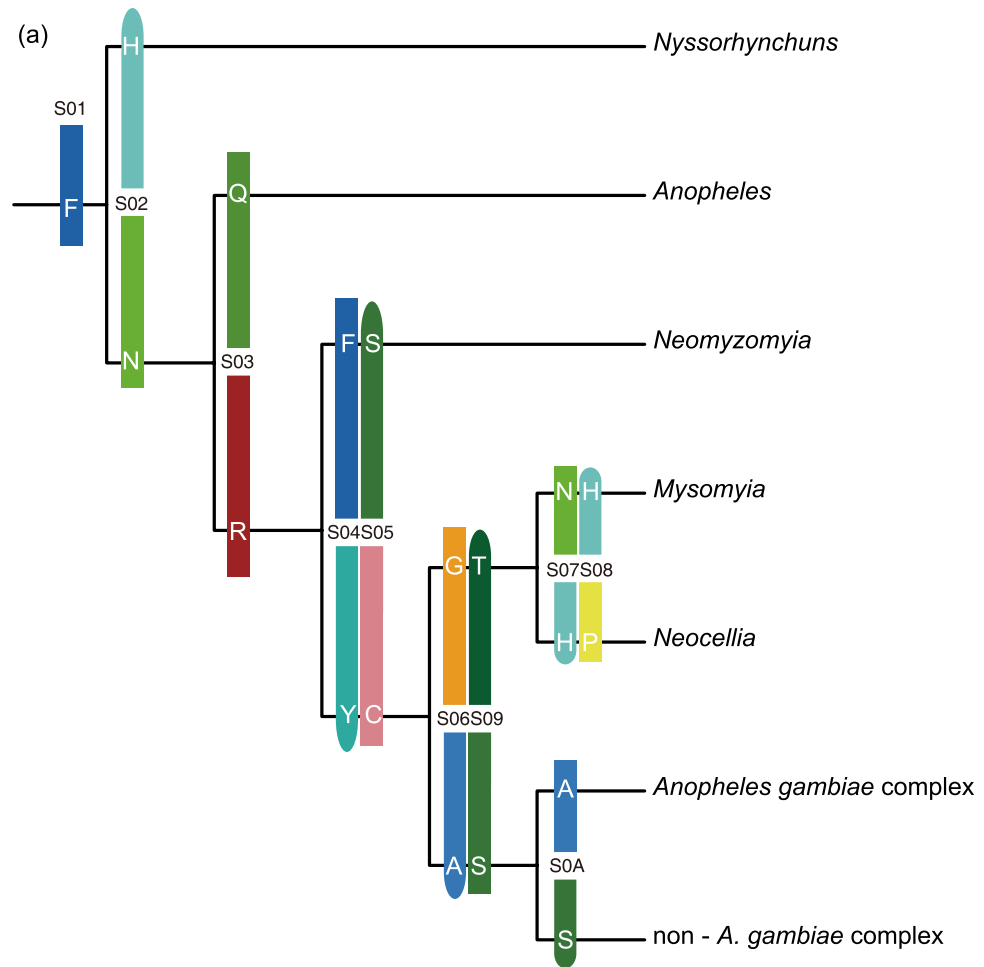
Figure 1. **Summary of the strategy in selecting apomorphies.** The black dot indicates the group of organism as the goal clade. The cross indicates that such a scenario should be refused, while the tick indicates that such a scenario can be accepted. Double ticks indicate an acceptance with high priority. (**a**) Preference of apomorphies based on the extent of overlapping. (**b**) Preference of apomorphies based on the data coverage in a site. (**c**) Preference of apomorphies based on the extent of uniqueness. (**d**) Preference of apomorphies based on the rarity of amino acids substitution.
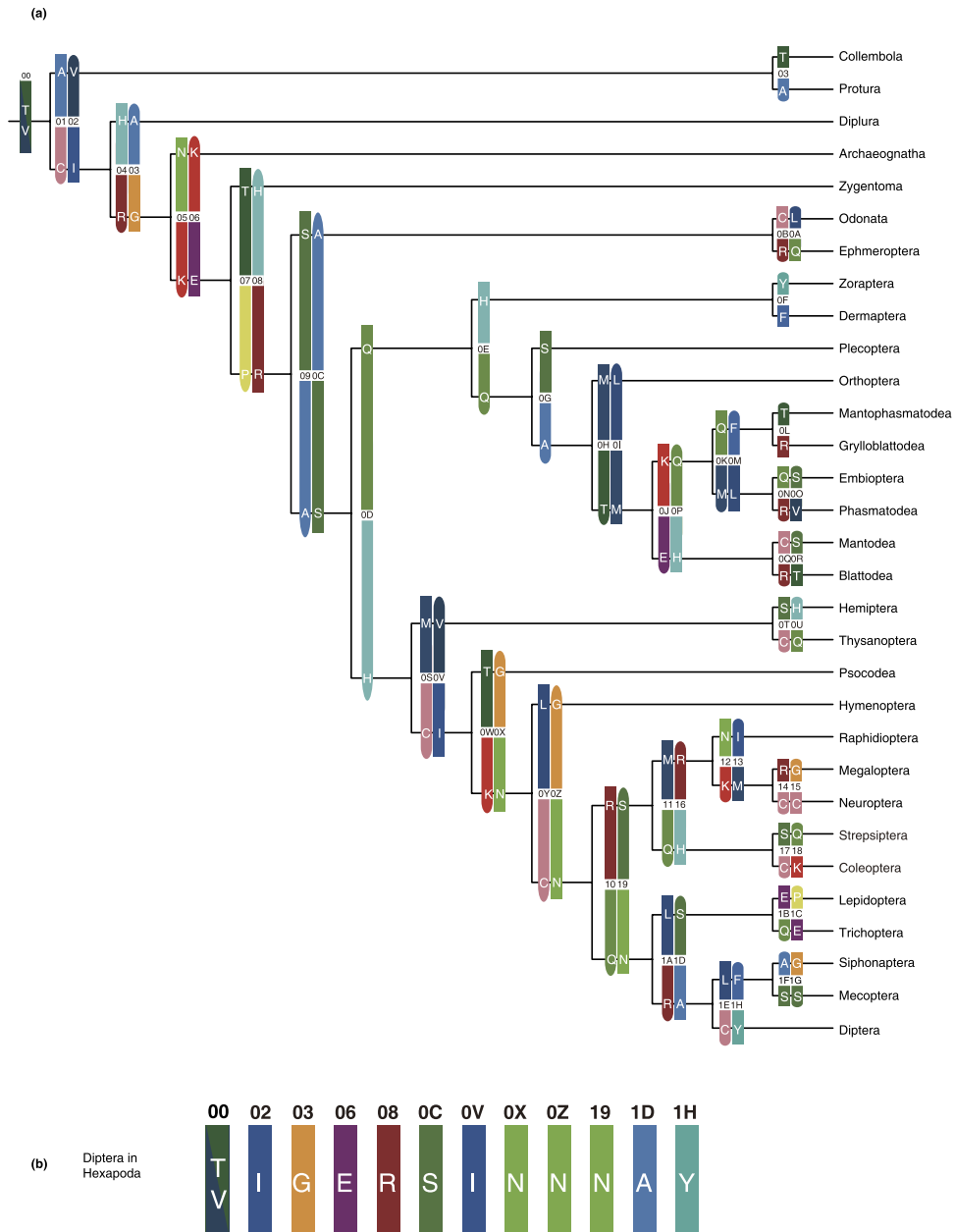
Figure 2. Molecular descriptions of clades in Hexapoda based on apomorphic amino acids. The apomorphies of amino acids are coloured based on the respective biochemical attributes. States shown in rounded rectangles indicate plesiomorphic states, while states shown in rectangles indicate apomorphic states. The diagonal indicates a binary apomorphic state. (**a**) Tree-like descriptions for clades in Hexapoda. (**b**) Combined description for Diptera.

tables are shown in Fig. 4. Each code assigned to an internal node in a tree contains two types of information: the state of the corresponding element and the position where the apomorphies are located. The positional information is shown in the form of sequential IDs to facilitate presentation, the annotations of which are shown in Fig. S3. A complete informative diagnostic description of the organism is composed of codes assigned to each node in its lineage from root to leaf, corresponding to the substantial part of the sequential ones, thus exhibiting a hierarchical structure. Notably, only the codes that refer to apomorphic states are actually informative for the whole description of organisms, which are meant to ensure the independence of codes. For example, the whole apomorphy-based description of Diptera and the *Anopheles gambiae* complex in Hexapoda can be shown using the sequence |T/V|IGERSINNNAY (ID: 00, 02, 03, 06, 08, 0C, 0V, 0X, 0Z, 19, 1D, and 1H, Fig. 2b) and |T/ V|IGERSINNNAY…FNRCSA (ID: 00, 02, 03, 06, 08, 0C, 0V, 0X, 0Z, 19, 1D, 1H, …, S01,S02,S03,S05,S09, and S0A, Fig. 3b), respectively. Such structure of information is similar to the rules for naming species and infraspecific taxa utilised in the latest version of PhyloCode[17]. In other words, the completely informative description of
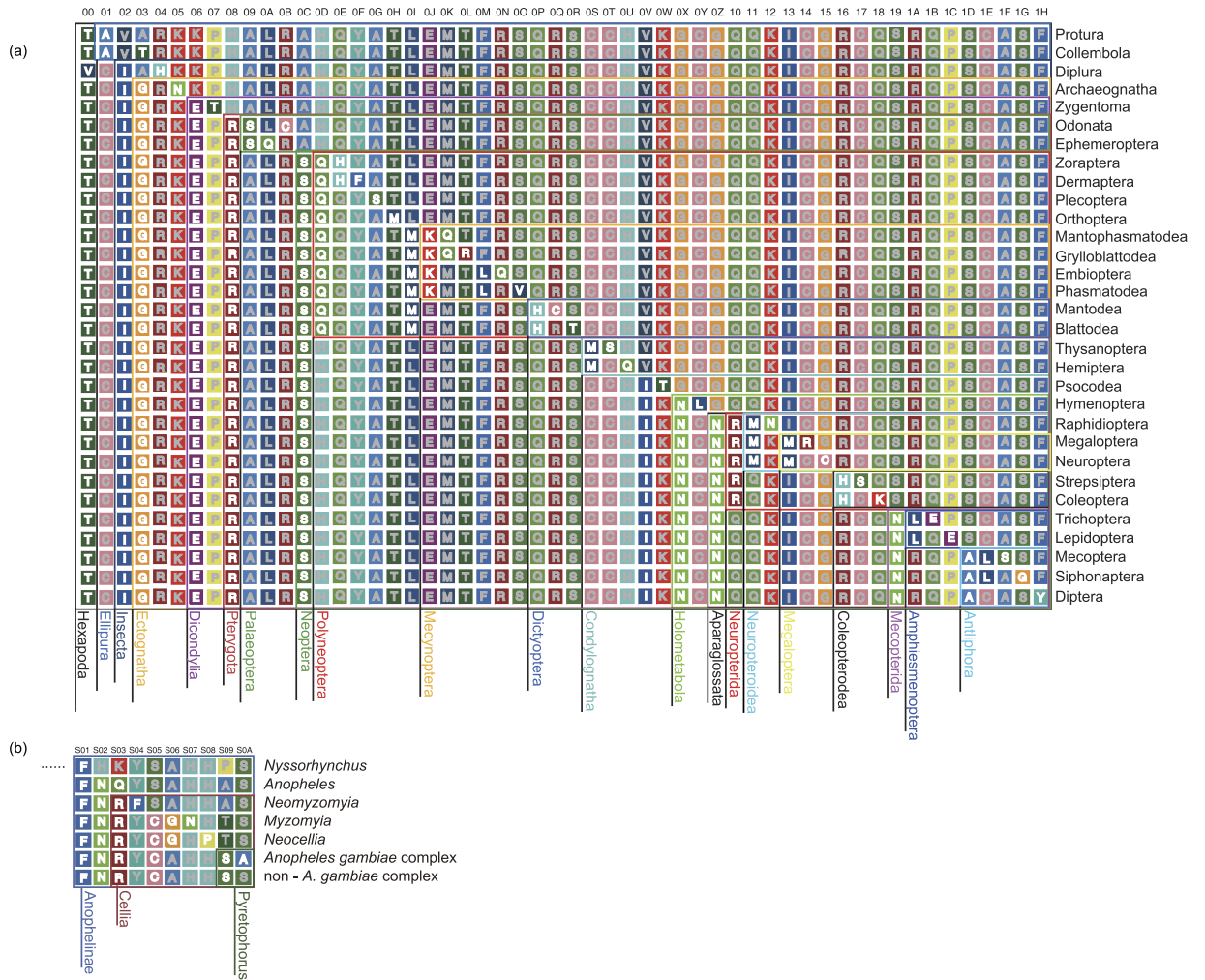
Figure 3. Molecular descriptions of clades in Anophelinae based on apomorphic amino acids. The apomorphies of amino acids are coloured based on the respective biochemical attributes. States shown as a rounded rectangle indicate plesiomorphic states, while states shown as a rectangle indicate apomorphic states. The diagonal indicates a binary apomorphic state. (**a**) Tree-like descriptions for clades in Anophelinae. (**b**) Combined description for *Anopheles gambiae* complex.

a clade can be further decomposed into two parts: a diagnostic code for a given clade and a hierarchical prefix composed of a sequential arrangement of codes assigned to the clades in which the given clade is nested.

Query test based on unknown transcriptomes. We categorised the results of the query as follows. (1) Positive, when the queried transcriptome is assigned to a right terminal node; (2) False-positive, when the queried transcriptome is assigned to a wrong terminal node; (3) False-negative, when the queried transcriptome cannot be assigned to any terminal node in the test database due to missing states; and (4) Negative, when the queried transcriptome cannot be assigned to any terminal node in the test database due to possible variations or sequencing errors. The results of the identification are shown in Table 1 and Table S1. Despite the historical controversy on monophyly, all polyneopterans and non-polyneopterans were correctly assigned to the respective group according to the codes. Among the 51 tested transcriptomes (Table S2), no false-positive or negative results were found. Due to "missing" states at the terminal nodes, three queries retrieved false-negative results. Although many of the internal nodes had "missing" states, 48 of the 51 queries obtained positive results. Approximately 60%

Figure 4. Sequential descriptions of clades based on apomorphic amino acids shown in a two-dimensional table. (**a**) Sequential descriptions of clades in Hexapoda. (**b**) Sequential descriptions of clades in Anophelinae. The number above each column is a numerical symbol. Apomorphies that are confirmed to be unique by comparing all of the organisms in the dataset are shown in white text. Non-apomorphic characters are shown in grey text. Each description for a lineage consists of two parts. The substantial parts for identification comprise apomorphic codes that are arranged following a strict hierarchical order (corresponding to the bars of discrete symbol). While the subordinate and trivial parts comprise the non-apomorphic characters, which only plays a structurally appurtenant role and contain no information for description and diagnoses (corresponding to the additional space of discrete symbologies). It should be noted that the minor variations in non-apomorphic characters are not shown for simplification, albeit the proportion of which are very small.

of these "missing" states can be attributed to missing sites of amino acids in the queried transcriptome (Table S1), which shows a semi-random distribution among transcriptomes of different sizes (Fig. S4A). The proportion of missing genes drops greatly as the size of transcriptome increases and reaches approximately zero when the size of transcriptome is over 90,000 contigs (Fig. S4B). The proportion of nodes with "missing" states first decreases steadily with an increase in the size of the transcriptome sequencing assembly but reaches a relatively stationary phase over 30,000 contigs (Fig. S4C).

Discussion

We propose that the results of our study can greatly benefit molecular apomorphy-based classification. First, the diagnostic molecular codes with mutual independence allow a novel and concise molecular approach for taxonomists to define and describe clades via a series of apomorphic amino acids. For example, along the lineage from the root of Hexapoda to Diptera, the clade of Insecta including Diplura and Ectognatha can be described as the clade originating from the ancestor species possessing apomorphy 829I as inherited from 829V on the clathrin heavy chain, the clade of Ectognatha can be described as the clade originating from the ancestor species possessing apomorphy 961G as inherited from 961A on nuclear hormone receptor FTZ-F1, and so on, until the clade of Diptera, which can be described as the clade originating from the ancestor species possessing apomorphy 692H as inherited from 692F on ubiquitin carboxyl-terminal hydrolase (Figs 2b, 4a and S3). This coding can

Species	Result	Generated Barcode in Test	Designed Barcode (Database)	Category
Pantala flavescens fabricius	Pterygota	$T_{00}I_{02}X_{03}E_{06}R_{08}X_{09}X_{0B}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{09}C_{0B}$	False-negative
Ischnura elegans	Odonata	$X_{00}I_{02}X_{03}E_{06}R_{08}X_{09}C_{0B}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{09}C_{0B}$	Positive
Isonychia kiangsinensis	Ephemeroptera	$T_{00}I_{02}X_{03}X_{06}R_{08}X_{09}Q_{0A}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{09}Q_{0A}$	Positive
Ephemera sp.	Ephemeroptera	$T_{00}I_{02}X_{03}E_{06}R_{08}S_{09}Q_{0A}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{09}Q_{0A}$	Positive
Eparchus insignis	Dermaptera	$X_{00}I_{02}?_{03}E_{06}R_{08}S_{0C}Q_{0D}H_{0E}F_{0F}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}Q_{0D}H_{0E}F_{0F}$	Positive
Flavoperla sp.	Plecoptera	$X_{00}I_{02}X_{03}?_{06}R_{08}X_{0C}Q_{0D}S_{0G}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}Q_{0D}S_{0G}$	Positive
Chondracris rosea	Orthoptera	$T_{00}I_{02}X_{03}E_{06}R_{08}S_{0C}Q_{0D}M_{0H}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}Q_{0D}M_{0H}$	Positive
Gryllotalpa unispina	Orthoptera	$?_{00}I_{02}X_{03}E_{0}R_{08}S_{0C}Q_{0D}M_{0H}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}Q_{0D}M_{0H}$	Positive
Hymenopus coronatus	Mantodea	$X_{00}I_{02}X_{03}E_{06}R_{08}S_{0C}Q_{0D}M_{0I}H_{0P}C_{0R}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}Q_{0D}M_{0I}H_{0P}C_{0R}$	Positive
Phraortes sp.	Phasmida	$T_{00}I_{02}X_{03}X_{06}R_{08}X_{0C}Q_{0D}M_{0I}X_{0J}L_{0M}V_{0O}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}Q_{0D}M_{0I}K_{0J}L_{0M}V_{0O}$	Positive
Coptotermes formosanus	Blattodea	$T_{00}I_{02}X_{03}E_{06}R_{08}S_{0C}Q_{0D}M_{0I}H_{0P}T_{0S}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}Q_{0D}M_{0I}H_{0P}T_{0S}$	Positive
Eupolyphaga sinensis	Blattodea	$T_{00}I_{02}X_{03}E_{06}R_{08}S_{0C}Q_{0D}M_{0I}X_{0P}T_{0S}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}Q_{0D}M_{0I}H_{0P}T_{0S}$	Positive
Periplaneta Americana	Blattodea	$T_{00}I_{02}X_{03}E_{06}R_{08}X_{0C}Q_{0D}M_{0I}X_{0P}T_{0S}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}Q_{0D}M_{0I}H_{0P}T_{0S}$	Positive
Pedetontus sp.	Archaeognatha	$T_{00}I_{02}X_{03}N_{05}$	$T_{00}I_{02}G_{03}N_{05}$	Positive
Lepisma sp.	Zygentoma	$T_{00}I_{02}X_{03}E_{06}T_{07}$	$T_{00}I_{02}G_{03}E_{06}T_{07}$	Positive
Anoplophora glabripennis	Coleopterodea	$T_{00}I_{02}X_{03}?_{06}R_{08}S_{0C}X_{0V}N_{0X}N_{0Z}R_{10}H_{16}X_{18}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}H_{16}K_{18}$	False negative
Antheraea assama	Lepidoptera	$T_{00}I_{02}G_{03}X_{06}R_{08}S_{0C}X_{0V}N_{0X}N_{0Z}N_{19}L_{1A}E_{1C}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}L_{1A}E_{1C}$	Positive
Anthonomus grandis	Coleoptera	$?_{00}I_{02}X_{03}E_{06}R_{08}X_{0C}X_{0V}N_{0X}X_{0Z}R_{10}H_{16}K_{18}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}H_{16}K_{18}$	Positive
Bactrocera dorsalis	Diptera	$?_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	Positive
Belgica antarctica	Diptera	$T_{00}I_{02}?_{03}E_{06}R_{08}S_{0C}I_{0V}?_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	Positive
Brassicogethes aeneus	Coleoptera	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}X_{0V}N_{0X}N_{0Z}R_{10}X_{16}K_{18}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}H_{16}K_{18}$	Positive
Chrysopa pallens	Neuroptera	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}X_{0V}N_{0X}N_{0Z}R_{10}M_{11}X_{13}C_{15}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}M_{11}M_{13}C_{15}$	Positive
Colaphellus bowringi	Coleoptera	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}H_{16}K_{18}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}H_{16}K_{18}$	Positive
Corydalinae sp.	Megaloptera	$T_{00}I_{02}X_{03}?_{06}R_{08}X_{0C}X_{0V}X_{0X}N_{0Z}R_{10}X_{11}X_{13}R_{14}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}M_{11}M_{13}R_{14}$	Positive
Crioscolia alcione	Hymenoptera	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}X_{0V}N_{0X}L_{0Y}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}L_{0Y}$	Positive
Culicoides sp.	Diptera	$T_{00}I_{02}G_{03}E_{06}R_{08}?_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	Positive
Dastarcus helophoroides	Coleoptera	$T_{00}I_{02}X_{03}E_{06}R_{08}S_{0C}X_{0V}N_{0X}N_{0Z}R_{10}H_{16}K_{18}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}H_{16}K_{18}$	Positive
Delia antiqua	Diptera	$T_{00}I_{02}?_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	Positive
Fopius arisanus	Hymenoptera	$T_{00}X_{02}G_{03}E_{06}?_{08}X_{0C}I_{0V}N_{0X}L_{0Y}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}L_{0Y}$	Positive
Hypothenemus hampei	Coleoptera	$T_{00}I_{02}?_{03}E_{06}?_{08}?_{0C}?_{0V}N_{0X}?_{0Z}R_{10}H_{16}K_{18}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}H_{16}K_{18}$	Positive
Ips typographus	Coleoptera	$T_{00}X_{02}?_{03}?_{06}R_{08}?_{0C}X_{0V}?_{0X}X_{0Z}R_{10}X_{16}K_{18}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}H_{16}K_{18}$	Positive
Lymantria dispar	Lepidoptera	$T_{00}I_{02}?_{03}X_{06}R_{08}X_{0C}X_{0V}N_{0X}X_{0Z}N_{19}L_{1A}E_{1C}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}L_{1A}E_{1C}$	Positive
Musca domestica	Diptera	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}X_{0V}X_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	Positive
Nevrorthus apatelios	Neuroptera	$?_{00}?_{02}?_{03}X_{06}?_{0C}X_{0V}?_{0X}?_{0Z}R_{10}?_{11}X_{13}C_{15}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}M_{11}M_{13}C_{15}$	Positive
Nicrophorus vespilloides	Coleoptera	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}H_{16}K_{18}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}H_{16}K_{18}$	Positive *
Oropsylla silantiewi	Siphonaptera	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}G_{1E}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}G_{1E}$	Positive *
Osmia cornuta	Hymenoptera	$T_{00}I_{02}G_{03}E_{06}R_{08}X_{0C}X_{0V}N_{0X}L_{0Y}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}L_{0Y}$	Positive
Polistes metricus	Hymenoptera	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}X_{0X}L_{0Y}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}L_{0Y}$	Positive
Raphidia ariadne	Raphidioptera	$T_{00}X_{02}G_{03}?_{06}R_{08}X_{0C}X_{0V}?_{0X}X_{0Z}?_{10}X_{11}N_{12}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}R_{10}M_{11}N_{12}$	Positive
Rhodinia newara	Lepidoptera	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}X_{0V}N_{0X}N_{0Z}N_{19}L_{1A}E_{1C}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}L_{1A}E_{1C}$	Positive
Samia ricini	Lepidoptera	$T_{00}I_{02}G_{03}X_{06}R_{08}S_{0C}X_{0V}N_{0X}X_{0Z}N_{19}L_{1A}E_{1C}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}L_{1A}E_{1C}$	Positive
Sitodiplosis mosellana	Diptera	$?_{00}I_{02}G_{03}?_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}?_{1D}Y_{1H}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	Positive
Stomoxys calcitrans	Diptera	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	Positive *
Telchin licus	Lepidoptera	$?_{00}?_{02}G_{03}X_{06}R_{08}X_{0C}X_{0V}?_{0X}?_{0Z}N_{19}?_{1A}E_{1C}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}L_{1A}E_{1C}$	Positive
Telenomus podisi	Hymenoptera	$T_{00}X_{02}?_{03}E_{06}R_{08}S_{0C}X_{0V}N_{0X}L_{0Y}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}L_{0Y}$	Positive
Teleopsis whitei	Diptera	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	Positive *
Tetramorium bicarinatum	Hymenoptera	$T_{00}I_{02}?_{03}E_{06}R_{08}S_{0C}X_{0V}N_{0X}L_{0Y}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}L_{0Y}$	Positive
Thaumetopoea pityocampa	Lepidoptera	$?_{00}?_{02}?_{03}?_{06}R_{08}?_{0C}X_{0V}?_{0Z}N_{19}X_{A}E_{1C}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}L_{1A}E_{1C}$	Positive
Themira biloba	Diptera	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}A_{1D}Y_{1H}$	Positive *
Xyela alpigena	Holometabola	$?_{00}?_{02}?_{03}X_{06}X_{08}?_{0C}X_{0V}N_{0X}?_{0Y}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}L_{0Y}$	False-negative
Yponomeuta evonymellus	Lepidoptera	$X_{00}I_{02}X_{03}X_{06}R_{08}S_{0C}X_{0V}N_{0X}X_{0Z}N_{19}L_{1A}E_{1C}$	$T_{00}I_{02}G_{03}E_{06}R_{08}S_{0C}I_{0V}N_{0X}N_{0Z}N_{19}L_{1A}E_{1C}$	Positive

Table 1. Results of query test of the 51 unknown transcriptomes. "X" represents missing amino acid residues, while "?" represents missing gene. "*" represents complete-matching result.

be further extended along the lineages within Diptera, taking the position of the *Anopheles gambiae* complex in Diptera-Anophelinae as an example (Figs 3b, 4b and S3). Such an approach can be especially useful for the

category levels at which uniquely diagnostic morphological variations are occasionally rare and could hardly be used to distinguish closely related taxa.

In addition, the application of a hierarchical prefix in the description could benefit the classification from the aspect of hierarchy, in which the information regarding nesting and mutual exclusivity can be definitely provided. For example, the prefix of Diptera is |T/V|IGERSINNNA (ID: 00, 02, 03, 06, 08, 0C, 0V, 0X, 0Z, 19, and 1D); therefore, Diptera must be nested in the clades corresponding to codes in the prefix, *e.g.*, Insecta (02-I), Pterygota (08-R), and Holometabola (0X-N), but can never be nested in clades such as Polyneoptera (0D-Q). Clades are much closer if they share more similar prefixes and are judged to be the closest or sister groups if the prefixes are the same but the diagnostic codes are different. With this structure, our approach can utilise an explicit phylogenetic definition by which the clades are fully defined and described under a phylogenetic framework and show explicit references to a particular phylogenetic hypothesis, thus coinciding with the principles of PhyloCode.

It may be suggested that there are similarities between the proposed molecular apomorphy-based classification and DNA barcoding, which is a technique of specimen classification that serves an important role in assessing and describing biological diversity[18–23] by using a DNA sequence from some gene (*e.g.*, cytochrome c oxidase subunit I (*COI*)) as a species-specific barcode[24–27]. As complete clade descriptions in our approach can be interpreted as a series of diagnostic codes along the lineages, our approach may also act as a sort of identification key and a diagnostic "barcode". Although either approach can implement standardisation, practicability, and generality in rapid identification, the criterion of the similarity-based barcoding methods[28,29] that is used to distinguish one species from others is the use of some categorised thresholds to describe gaps in genetic distances. Moreover, similarity-based strategies also restrict current DNA barcoding to a leaves-only processing that only provides one level of resolution and does not focus on precise information regarding the relationships among barcodes. Indeed, current DNA barcoding methods cannot describe a clade higher than the species level through explicit reference to phylogeny, although this is not the main question that they are designed to answer.

In contrast, the criterion applied here for distinguishing one clade from the others is qualitative rather than quantitative. A series of molecular apomorphies are used as clade identification tags that are unique for certain groups of organisms and completely distinct among different groups. Apomorphies at the amino acid level are mainly used to ease the ambiguity resulting from molecular homoplasy. As opposed to the mere assemblage of mutually exclusive characters used in diagnostic barcoding[30,31], after apomorphy mining and filtering under a set of criteria, the apomorphies of various nodes should be arranged according to the rank of the node in the lineage from root to leaf, by which the hierarchical prefix and the diagnostic code of a clade can be given and combined in a particular order. In this sense, the whole description of clades in this classification system is highly hierarchical, thus facilitating phylogenetic descriptions and diagnoses of clades that we would like to study at various category levels. In this study, the concrete establishment of classification systems in Hexapoda and Anophelinae has shown the feasibility of hierarchical molecular apomorphy-based classification in multiple category levels. Additionally, at the root end, some apomorphic nucleotides/bases have been shown to exist and demonstrate the existence of group specific molecular attributes in bacteria, archaea, and eukaryotes[32]; meanwhile, at the leaf end, apomorphic amino acids/nucleotides have been successfully discovered in several species which could hardly be distinguished when using morphological characters[33,34]. Therefore, benefiting from the unambiguous attribute of apomorphy, a hierarchical molecular apomorphy-based classification system has the ability to put forward a general criterion without sensitivity resulting from various genetic distances among all living organisms on Earth.

The general procedures for database construction are also different between DNA barcoding methods and the hierarchical molecular apomorphy-based classification, as shown in Fig. 5a,b. In DNA barcoding methods, the sequences of gene markers are directly deposited in the database and linked with the identification information. While in the hierarchical apomorphy-based classification, two related sub-databases should be formed simultaneously. One database contains sets of core-orthologs of each homologous gene, in which the information of apomorphic amino acids is imbedded, and the other contains descriptions, *i.e.*, sequences of codes for organisms based on discrete and apomorphic amino acids. The sub-database of core-orthologs should be updated regularly according to orthology annotation databases, such as KEGG[35], OrthoDB[36], OMA[37], while the sub-database of descriptions can follow a similar way that joins published data into a fully versioned and dynamic framework[38,39].

It should be noticed that the molecular apomorphy-based classification does not abandon the existing monolocus data used in molecular identification. On the contrary, compiled homologous sequences should be explored as much as possible to discover the molecular apomorphic sites, no matter whether they are short sequences or genome or transcriptome data, and no matter whether they are amino acids or nucleotides. Therefore, such inclusiveness offers an optimal utilisation of the existing sequence data for the apomorphy-based barcoding system, thus leading to standard and efficient molecular identifications in a post-genomic era. Simultaneously, the approach for molecular classification proposed here does not imply a replacement of the existing biological classification system. In fact, the apomorphy-based molecular classification should be seen as an epistemological and methodological complement and extension that could be compatible with classification systems with different ontological declarations.

Several challenges may occur during the application of a hierarchical apomorphy-based classification system. First, as shown in the query test results, missing genes and missing amino acid sites in sequenced transcriptomes will affect the efficiency and accuracy of this approach to some extent, thereby causing ambiguities in descriptions and identifications. Such challenge could hopefully be overcome in two ways in the future. As shown in Fig. S4B, the proportion of missing genes resulting from the incompleteness of the transcriptome is greatly reduced as the size of the transcriptome increases. That indicates that the problem of missing genes can be effectively solved with increases in sequencing throughput. While for the proportion of missing amino acid sites resulted from fragmented sequencing, such cases can be hopefully solved by increasing the length of reads and by increasing the completeness of assembly in the progressive high-throughput sequencing techniques[40–42]. Thus, the proportion of false-negative results can be reduced correspondingly.
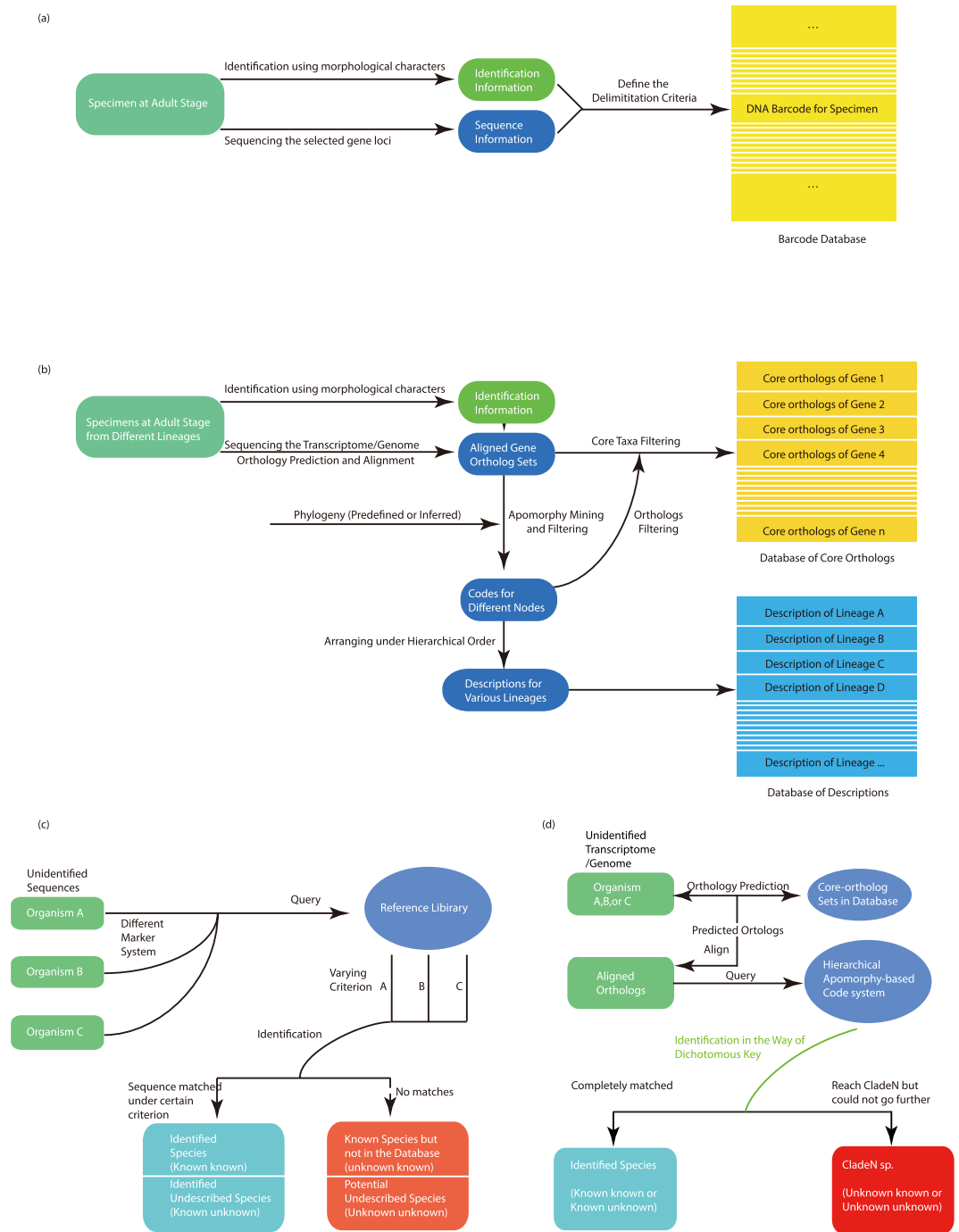
Figure 5. Comparison of DNA barcoding methods and the hierarchical molecular apomorphy-based classification system. Complete definition of known known, known unknown, unknown known, unknown unknown can be found in the study produced by Collins and Cruickshank[59]. (**a**) General workflow of database construction in the previous barcoding methods. (**b**) General workflow of database construction in the hierarchical molecular apomorphy-based system. (**c**) General workflow of identification in the previous barcoding methods. (**d**) General workflow of identification in the hierarchical molecular apomorphy-based classification system.

Because the apomorphy-based system relies highly on tree topology, such a system may encounter problems if a phylogenetic ground plan has not been well established for the group under consideration. In the phylogenomic era, the currently recognised challenges include non-random distributed missing data, great rate heterogeneity, and serious incomplete lineage sorting (ILS), among others. According to a comprehensive survey using simulated and empirical big data[43], several gene-tree-based coalescent (ASTRAL, MP-EST) and supertree (MRP) methods consistently recovered the true species tree as the number of genes increased to 1,000 even in the presence of 70% non-random missing data either in sampled taxa or in genes with high ILS and rate heterogeneity.

In other words, the reliability of the reference phylogenomic trees used for molecular apomorphy-based coding can be checked by such methods and have good opportunities to be convincing with the continuously increasing –omics-based data currently and in the future.

It may be argued whether all of the qualified nodes should be encoded in a fully resolved tree when considering the convenience of taxonomic practices. Although a complete set of codes may appear to be excessive, especially compared to traditional keys and the morphological characters used in them, fully encoded clade descriptions in fact provide other conveniences. Because the storage of codes and the procedure of decoding can both be accomplished computationally, restrictions in the length of descriptions and diagnoses are indeed relaxed. In addition, similar to other methods for molecular identification, the apomorphy-based encoding can be relaxed from the restrictions of sexual dimorphism and developmental stages and the professionalism requirement for taxonomic practitioners, thus simplifying the workflow of taxonomic practices. Furthermore, because both the description and identification of a certain clade follow a tree-climbing procedure on a strictly evaluated phylogenetic tree, the dichotomous-key-like presentation of successive codes can facilitate identification, evaluation and comparisons among closely related organisms in a fully phylogenetic way. In this sense, a complete set of codes can finally achieve an accurate and strictly phylogenetic description of biological organisms. Therefore, we propose that full encoding should be encouraged rather than simply translating traditional levels or ranks into codes that are often arbitrary and may result in subsequent controversy.

It should be noted that only one apomorphy per clade was used as code in the designed database in this study for the convenience of illustration. In fact, such a scheme is expandable in a real classification system. Because rare mutants may occur at even the most conservative sites in some individuals of any species (*i.e.*, diagnostic exceptions where case subtaxa deviate from the otherwise diagnostic identity of a given state), the redundancy of multiple codes for the same clade in identification keys will be necessary when using apomorphic amino acids. The strategy of adopting multiple codes rather than one for the description and identification of a certain clade can be viewed as a supraspecific extension of the"near-minimal" set of SNPs (single nucleotide polymorphism), which are commonly used in species-level rapid identification[44]. Such a strategy may be especially important in tackling some "problematic" taxonomic groups or organisms that have experienced recent speciation. To increase the available amount of qualified apomorphies, the coverage of sites in the sequence matrix used for apomorphy mining should be as high as possible. Nevertheless, such a redundancy of apomorphies does not mean a conditional combination of traits. In fact, each apomorphy in codes with redundancy is independent from the others. The strategy of multiple encoding is only to avoid the error resulting from minor and practical constraints. On the other hand, together with the development of sequencing technologies, the mining of apomorphic amino acids can be improved by reducing missing data. Although the extent of the reference genome coverage remains biased in that there is a dearth of non-vertebrate genomes throughout the tree of life[45], hundreds of genomes and more than hundreds of thousands of transcriptomes in eukaryotes have been sequenced. Furthermore, the development of third-generation sequencing can greatly boost the process of genome sequencing with broader taxon sampling. The eliminated need for excessive reagents and the harnessing of the processivity of DNA polymerase in third-generation sequencing allow an increase in the integrity of throughput and a decrease in the time and cost of sequencing[42]. Moreover, with the gradual accomplishment of genome annotation and orthology prediction in distantly related taxa, the number of genes that can be used for apomorphy mining will be increased accordingly. Therefore, using redundant apomorphies as identification tags for one clade is both necessary and feasible, and we propose that the permanent code for an organism should have more than one column site for each clade.

Furthermore, the challenge resulting from the high specialisation and the non-generalisation of the marker system can also be relieved or even overcome with advances in sequencing technology. In contrast to the early age of DNA barcoding using Sanger sequencing, it has become realistic in the era of high-throughput sequencing to generate a large amount of molecular data from different loci simultaneously, efficiently, and economically. The explosively increasing amount of genome and transcriptome data may even permit apomorphy mining in almost all of the extant organisms in the future. As a result, molecular classification studies have been largely freed from the restrictions of sequencing and the number of markers used. Moreover, it is realistic to compare markers from thousands of available gene sequences based on the existence of apomorphic amino acids for a certain clade, which can lead to the progressive optimisation of the marker system. Therefore, with an even more rapid accumulation of genome and transcriptome data in the future, the hierarchical apomorphy-based classification system can achieve standardisation and, thus, the gradual fixation of the marker system.

The broadness and depth of genomic and transcriptomic data enable researchers to obtain more reliable topologies in phylogenetic reconstructions and provide more opportunities to discover informative group-specific amino acids and/or nucleotides. Benefiting from these advances, we are now able to reframe the methodology of description in taxonomy. The hierarchical molecular apomorphy-based classification system proposed in this study can be very helpful in leading to precise descriptions of clades from the most microscopic but essential aspect of evolution and may even develop as an alternative, efficient approach for organism identification. Furthermore, the hierarchical apomorphy-based classification system provides a practical way of standardising the phylogenetic descriptions and nomenclature of clades, thus offering a potential methodological implementation of PhyloCode and facilitating its development. In this sense, the hierarchical apomorphy-based classification system can serve as a primer of integrative taxonomy[46–48] linking phylogenomics, molecular classification, and phylogenetic nomenclature.

Materials and Methods

Apomorphy mining and sub-dataset generation. The transcriptome and/or genome data from a recent published phylogenomic study of Hexapoda[14] and from a comparative genomic study of Anophelinae[15] were chosen to generate two source datasets for this study. At the nucleotide level, due to evolutionary saturation

and convergence, it is more common for two distantly related groups of organisms to share the same base status in a site, *i.e.*, molecular homoplasy. Therefore, contrary to previous strategies of molecular classification, we instead used the diversity of amino acids as clade description in this study. Considering the performance (signal versus noise) in the phylogenetic reconstruction, we chose the amino acid sequence matrix Supermatrix C_AA (provided by Misof *et al.*[14]) with a protein domain-based partitioning scheme as the source dataset for apomorphy search and named it Dataset 1. For the apomorphy mining of the genomes in the comparative genomic study by Neafsey *et al.*[15], the matrix was constructed as follows. Seventeen genomes belonging to seven lineages in Anophelinae (Table S3) were downloaded from GenBank. Gene orthology was predicted using HaMStR v 13.2[49] with the same core-orthologs as those used in the phylogenomic study of Misof *et al.*[14]. Sequences were then initially aligned using MAFFT[50] and then refined using MUSCLE[51]. After the exclusion of random similar sites using ALICUT[52] based on the score given by Aliscore[53], sequences were concatenated as Dataset 2 using an in-house script.

Thereafter, we applied a parsimony method to explore potential apomorphies in the datasets. First, the sequence matrices and the corresponding treefiles were imported into PAUP* 4.0b10[54]. For Dataset 1, the topology of the treefile was fixed to the final result in the original study[14]. For Dataset 2, the topology of the treefile was the same as the one provided by the Broad Institute (https://olive.broadinstitute.org/projects/anopheles). All of the taxa in Hexapoda and Anophelinae were defined as ingroups. State optimisation of parsimony was set to DELTRAN. After the log-file option was activated, sequence data were then used to obtain a labelled tree with a complete list of apomorphies (Describetrees/root = outgroup plot = phylogram labelnode = yes apolist = yes). The results revealed all of the possible apomorphies of the corresponding dataset. Sites with ambiguous changes were abandoned. For species in distantly related groups, it is still possible that the evolution of amino acids also show convergence to some degree. To avoid this scenario, the CI of each apomorphy was then used as a filter criterion. Generally, the changes are very likely to be homoplasious if the value of CI is lower than 0.3. Herein, considering the randomness of the missing data in the majority of existing transcriptomic data, three more stringent filter criteria (CI varied from 1.0 to 0.8) were applied to ensure the credibility of the apomorphy call. A series of in-house shell scripts were used to filter the apomorphies listed in the log-file given by PAUP and to generate various sub-dataset files for the subsequent analyses, in which apomorphies filtered under different criteria were assembled.

Phylogenetic analyses of the subsets. To compare the phylogenetic signal versus noise of apomorphies filtered by different criteria, three amino acid sequence sub-datasets for each of the two original datasets were generated, which were named sub-datasets 1A, 1B, 1C, and 2A, 2B, 2C, with the filter criterion of CI equal to or higher than 1.0, 0.9, and 0.8, respectively. For each sub-dataset, we performed phylogenetic analysis with maximum likelihood (ML) methods using the Pthreads-parallelised version of RAxML, v.8.0.12[55]. The provisional partition schemes were made with data blocks based on orthologous genes boundaries, and then PartitionFinder v. 1.1[56] was used to infer further model estimation and partition schemes. As for sub-datasets trimmed from Dataset 1 and the corresponding reduced partition schemes, we used the same substitution model as the original one for consistency. The randomised stepwise addition parsimony trees were used as the starting tree. For each dataset, 100 searches were performed for the ML tree, with the robustness of branches tested by 100 bootstrap replicates. Rapid bootstrap analysis and the search for the best scoring ML tree were performed in a one program run (-f a).

Design of a classification system of Hexapoda and Anophelinae using molecular apomorphies.
According to its performance in the phylogenetic analyses, sub-dataset 1B in the case of Hexapoda and sub-dataset 2C in the case of Anophelinae were chosen for designing the clade codes. The principle described here consists of assigning each internal node in the phylogenetic tree a unique "code", which represents the molecular apomorphy shared by that clade. Two classification systems corresponding to the category levels between superclass and order and between subfamily and species complex were constructed. Apomorphies of clades with extremely low coverage of taxon sampling were abandoned. The arrangements of codes were designed to be as hierarchical as possible, and they show a similar order within the tree obtained by phylogenetic inference. The following criteria were adopted to select a series of the most proper apomorphies from sub-datasets 1B and 2C.

(1) Apomorphies for one clade should have no overlap with its sister-group, *i.e.*, if an apomorphic state of one clade is also found in its sister-group, such apomorphy should not be used because it is difficult to ascertain whether the possible apomorphy is a bona fide one or some type of plesiomorphy (Fig. 1a).
(2) To ensure reliability, missing taxa that are supposed to share apomorphies should be limited. In addition, if the number of missing data is relatively large in the clade that is supposed to share the apomorphies, the reliability of the apomorphies becomes questionable. Apomorphies should never be used, no matter how "good", if the states of the sister group or the basal clade of the analytic group are missing (ambiguous) (Fig. 1b).
(3) Apomorphies that are unique, *i.e.*, autapomorphic, in a clade should be given the higher priority (Fig. 1a).
(4) Apomorphies that are unique in the target clade are preferred. Apomorphies with binary or multinary states can also be candidates, but they should be less desirable unless they are shared by sub-clades and show a consistent evolutionary trend within the tree (Fig. 1c). For those preferred apomorphies, preference increases with the strength of the attributes that are close to autapomorphies (Fig. 1d).
(5) For the purpose of simplification, it is more acceptable that fewer gene markers are involved. Therefore, the preference for a gene increases as the range of the taxa possessing a corresponding amino acid apomorphy broadens.

Following these criteria, we made an apomorphy-screening pipeline (Fig. S5) and selected one apomorphy for each node for the convenience of illustration. However, for groups that contain only one taxon in our sample (*e.g.*, Protura, Zoraptera and Mantophasmatodea), the apomorphies of these nodes were discarded because such characters are not shared by members within each taxon. The description of these apomorphic characters involves the state changes of each amino acid, the names of genes that possess those amino acid sites, the positions of those sites compared to their homologous genes in Drosophila, and the available structure information of the corresponding proteins, which were inferred with SWISS-Model online workplace[57,58] and filtered by similarity (higher than 40%).

Identification of unknown transcriptome as a proof-of-concept test. Considering the incompleteness of transcriptome data due to missing genes and missing sites in sequenced genes, it could be argued that transcriptomic data can be adequate for query identification. Therefore, we carried out a group of tests to see the identification ability of an unknown transcriptome based on diagnostic apomorphies. In view of the taxon coverage of transcriptome data, the sequence assemblies of 15 *de novo* sequenced transcriptomes in this study and 36 transcriptomes obtained through GenBank (and not used in the works of Misof *et al.*[14]), which together represent 18 orders of insects (Table S2), were used for querying the designed classification system of Hexapoda. The test results of Polyneoptera were additionally summarised considering both the taxon coverage of the dataset and the historical controversy.

The workflow of identification is illustrated in Fig. 5d. Contrary to the leaves-only processing of DNA through barcoding (Fig. 5c), the identification can be interpreted as a query along the lineage from the root to the leaf. The first step of identification is predicting the orthology of the queried transcriptome or genome using the core-orthologs of database, followed by alignment between the predicted orthologs and the core-orthologs. Subsequent identifications then follow a series of progressive steps corresponding to the fixed sequential steps as those in an identification key. If the matching judgment of a certain internal node based on the state of apomorphic amino acid is fulfilled, then the identification will progress and continue repeatedly until a terminal node is reached or no matches are found, at which point the final generated query description is obtained by assembling the matched codes in a hierarchical order. If the matching judgment encounters some missing state and the corresponding step is ambiguous, then the subsequent steps will progressively check the state of each child node until a clear judgment can be given (Fig. S6).

References

1. De Queiroz, K. & Gauthier, J. Toward a phylogenetic system of biological nomenclature. *Trends Ecol. Evol.* **9,** 27–31 (1994).
2. Backlund, A. & Bremer, K. To be or not to be – principles of classification and monotypic plant families. *Taxon* **47,** 391–401 (1998).
3. Nixon, K. C. & Carpenter, J. M. On the other "phylogenetic systematics". *Cladistics* **16,** 298–318 (2000).
4. Judd, W. S., Campbell, C. S., Kellogg, E. A., Stevens, P. F. & Donoghue, M. J. *Plant systematics. A phylogenetic approach.* (Third edition. Sinauer Associates, Sunderland, MA, XV) pp. 611 (2007).
5. De Queiroz, K. & Gauthier, J. Phylogeny as a central principle in taxonomy: Phylogenetic definitions of taxon names. *Syst. Zool.* **39,** 307–322 (1990).
6. De Queiroz, K. & Gauthier, J. Phylogenetic taxonomy. *Annu. Rev. Ecol. Syst.* **23,** 449–480 (1992).
7. De Queiroz, K. & Donoghue, M. J. Phylogenetic nomenclature, three-taxon statements, and unnecessary name changes. *Syst. Biol.* **60,** 887–892 (2011).
8. De Queiroz, K. & Donoghue, M. J. Phylogenetic nomenclature, hierarchical information, and testability. *Syst. Biol.* **62,** 167–174 (2012).
9. Platnick, N. I. The poverty of the PhyloCode: a reply to de Queiroz and Donoghue. *Syst. Biol.* **61,** 360–361 (2012).
10. Ward, P. S. Integrating molecular phylogenetic results into ant taxonomy (Hymenoptera: Formicidae). *Myrmecol. News* **15,** 21–29 (2011).
11. Marakeby, H. *et al.* A system to automatically classify and name any individual genome-sequenced organism independently of current biological classification and nomenclature. *PLoS One* **9,** e89142 (2014).
12. Jolley, K. A. *et al.* Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* **158,** 1005–1015 (2012).
13. Hennig, W. *Phylogenetic Systematics.* (University OF Illinois Press, Urbana Chicago London.) pp. 89 (1966).
14. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346,** 763–767 (2014).
15. Neafsey, D. E. *et al.* Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science* **347,** 1258522 (2015).
16. Chapman, A. D. *Numbers of living species in Australia and the World.* (2009) Available at http://www.environment.gov.au/node/13876/. (Date of access: 28th November 2015).
17. Cantino, P. D. & de Queiroz, K. International code of phylogenetic nomenclature (the PhyloCode), Version 4c. (2010), Available at http://www.phylocode.org/. (Data of access: 15th September 2015).
18. Zhou, X. *et al.* Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Gigascience* **2,** 4 (2013).
19. Sivaramakrishnan, K. G., Janarthanan, S., Selvakumar, C. & Arumugam, M. Aquatic insect conservation: a molecular genetic approach. *Conserv. Genet. Resour.* **6,** 849–855 (2014).
20. Chesters, D. & Zhu, C. D. A protocol for species delineation of public DNA databases, applied to the Insecta. *Syst. Biol.* **63,** 712–725 (2014).
21. Bilgin, R., Utkan, M. A., Kalkan, E., Karhan, S. U. & Bekbolet, M. DNA barcoding of twelve shrimp species (Crustacea: Decapoda) from Turkish seas reveals cryptic diversity. *Mediterr. Mar. Sci.* **16,** 36–45 (2015).
22. Chariton, A. A. *et al.* Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. *Environ. Pollut.* **203,** 165–174 (2015).
23. Lopes, C. M. *et al.* DNA metabarcoding diet analysis for species with parapatric vs sympatric distribution: a case study on subterranean rodents. *Heredity (Edinb)* **114,** 525–536 (2015).
24. Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. DNA points the way ahead in taxonomy. *Nature* **418,** 479 (2002).
25. Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. A plea for DNA taxonomy. *Trends Ecol. Evol.* **18,** 70–74 (2003).
26. Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proc. Biol. Sci.* **270,** 313–321 (2003).
27. Hebert, P. D. N., Ratnasingham, S. & deWaard, J. R. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. Biol. Sci.* **270** Suppl 1, S96–99 (2003).

28. Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S. & Francis, C. M. Identification of Birds through DNA Barcodes. *PLoS Biology* **2,** e312 (2004).
29. Ratnasingham, S. & Hebert, P. D. N. A DNA-based registry for all animal species: the barcode index number (BIN) system. *PLoS One* **8,** e66213 (2013).
30. Bertolazzi, P., Felici, G. & Weitschek, E. Learning to classify species with barcodes. *BMC Bioinformatics* **10,** S7 (2009).
31. Weitschek, E., Van Velzen, R., Felici, G. & Bertolazzi, P. BLOG 2.0: a software system for character-based species classification with DNA Barcode sequences. What it does, how to use it. *Mol. Ecol. Resour.* **13,** 1043–1046 (2013).
32. Xie, Q., Wang, Y., Lin, J., Qin, Y., Wang, Y. & Bu, W. Potential key bases of ribosomal RNA to kingdom-specific spectra of antibiotic susceptibility and the possible archaeal origin of eukaryotes. *PLoS One* **7,** e29468 (2012).
33. Pavan, M. G. *et al.* A nuclear single-nucleotide polymorphism (SNP) potentially useful for the separation of *Rhodnius prolixus* from members of the *Rhodnius robustus* cryptic species complex (Hemiptera: Reduviidae). *Infect. Genet. Evol.* **14,** 426–433 (2013).
34. Li, J. *et al.* Discriminating Dabry's sturgeon (*Acipenser dabryanus*) and Chinese sturgeon (*A. sinensis*) based on DNA barcode and six nuclear markers. *Hydrobiologia* **757,** 185–196 (2015).
35. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).
36. Waterhouse, R. M., Tegenfeldt, F., Li, J., Zdobnov, E. M. & Kriventseva, E. V. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.* **41,** 358–365 (2013).
37. Altenhoff, A. M. *et al.* The OMA orthology database in 2015: function predictions, better plant support, synteny view, and other improvements. *Nucleic Acids Res.* **43,** 240–249 (2015).
38. Riedel, A., Sagata, K., Suhardjono, Y. R., Tänzler, R. & Balke, M. Integrative taxonomy on the fast track—towards more sustainability in biodiversity research. *Front. Zool.* **10,** 15 (2013).
39. Riedel, A., Sagata, K., Surbakti, S., Tänzler, R. & Balke, M. One hundred and one new species of *Trigonopterus* weevils from New Guinea. *ZooKeys* **280,** 1–150 (2013).
40. Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum. Mol. Gen.* **19,** 227–240 (2010).
41. Liu, L. *et al.* Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* **2012,** 251364 (2012).
42. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Gen. Biol.* **14,** R51 (2013).
43. Xi, Z., Liu, L. & Davis, C. C. The impact of missing data on species tree estimation. *Mol. Biol. Evol.* **33,** 838–860 (2016).
44. Frey, J. E., Guillén, L., Frey, B., Samietz, J., Rull, J. & Aluja, M. Developing diagnostic SNP panels for the identification of true fruit flies (Diptera: Tephritidae) within the limits of COI-based species delimitation. *BMC Evol Biol.* **13,** 106 (2013).
45. Richard, S. It's more than stamp collecting: how genome sequencing can unify biological research. *Trends Genet.* **Pii**, S0168-9525(15)00088–8 (2015).
46. Schlick-Steiner, B. C., Steiner, F. M., Seifert, B., Stauffer, C., Christian, E. & Crozier, R. H. Integrative taxonomy: a multisource approach to exploring biodiversity. *Annu. Rev. Entomol.* **55,** 421–438 (2010).
47. Yeates, D. K., Seago, A., Nelson, L., Cameron, S. L., Joseph, L. & Truemann, J. W. H. Integrative taxonomy, or iterative taxonomy? *Syst. Entomol.* **36,** 209–217 (2011).
48. Pante, E., Schoelinck, C. & Puillandre, N. From integrative taxonomy to species description: one step beyond. *Syst. Biol.* **64,** 152–160 (2015).
49. Ebersberger, I., Strauss, S. & Von Haeseler, A. HaMStR: profile hidden Markov model based search for orthologs in ESTs. Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories, Vienna, Austria. URL https://sourceforge.net/projects/hamstr/ (2009).
50. Katoh, K. & Standley, D. M. MAFFT: Multiple alignment program for amino acid or nucleotide sequences, version 7. Immunology Frontier Research Center, Osaka University, Suita, Osaka, Japan. URL http://mafft.cbrc.jp/alignment/software/linux.html/ (2013).
51. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. 195 Roque Moraes Drive, Mill Valley, CA 94941, *USA* URL http://www.drive5.com/muscle/downloads.htm/ (2004).
52. Kück, P. ALICUT: a PerlScript which cuts ALISCORE identified RSS. version 2.0 edition. Department of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig (ZFMK), Bonn, Germany. URL https://www.zfmk.de/de/forschung/forschungszentren-und-gruppen/utilities/ (2009).
53. Misof, B. & Misof, K. Aliscore, A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Department of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig (ZFMK), Bonn, Germany. URL https://www.zfmk.de/en/research/research-centres-and-groups/aliscore/ (2009).
54. Swofford, D. L. PAUP* Phylogenetic Analysis using Parsimony (* and other Methods). Version 4. [Computer software and manual]. Sinauer Associates, Sunderland, Massachusetts, URL http://www.paup.csit.fsu.edu/ (2002).
55. Stamatakis, A. RAxML: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies, Version 8.0.12. Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany. URL https://github.com/stamatak/standard-RAxML/ (2014).
56. Lanfear, R., Calcott, B., Ho, S. Y. & Guindon, S. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. The Molecular Evolution Lab Macquarie University, Sydney. URL http://www.robertlanfear.com/partitionfinder/ (2012).
57. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Biozentrum, University of Basel, Basel 4056, Switzerland. URL http://swissmodel.expasy.org/ (2006).
58. Biasini, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Biozentrum, University of Basel, Basel 4056, Switzerland. URL http://swissmodel.expasy.org/ (2014).
59. Collins, R. A. & Cruickshank, R. H. Known knowns, known unknowns, unknown unknowns and unknown knowns in DNA barcoding: a comment on Dowton *et al. Syst. Biol.* **63,** 1005–1009 (2014).

Acknowledgements

Author Contributions

Q.X. and W.-J.B. conceived and designed the research; H.-Y.W., Y.-H.W., Q.X. and Y.-L.K. collected the specimens; H.-Y.W. and Y.-H.W. performed the experiments and analysed the data; H.-Y.W., Y.-H.W., Q.X. and Y.-L.K. wrote and revised the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Wu, H.-Y. *et al.* Molecular classification based on apomorphic amino acids (Arthropoda, Hexapoda): Integrative taxonomy in the era of phylogenomics. *Sci. Rep.* **6**, 28308; doi: 10.1038/srep28308 (2016).