Ab Initio Construction and Evolutionary Analysis of Protein-Coding Gene Families with Partially Homologous Relationships: Closely Related *Drosophila* Genomes as a Case Study

Xia Han, Jindan Guo, Erli Pang, Hongtao Song, and Kui Lin*

State Key Laboratory of Earth Surface Processes and Resource Ecology, Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, China

*Corresponding author: E-mail: linkui@bnu.edu.cn.

Accepted: February 18, 2020

Abstract

How have genes evolved within a well-known genome phylogeny? Many protein-coding genes should have evolved as a whole at the gene level, and some should have evolved partly through fragments at the subgene level. To comprehensively explore such complex homologous relationships and better understand gene family evolution, here, with de novo-identified modules, the subgene units which could consecutively cover proteins within a set of closely related species, we applied a new phylogeny-based approach that considers evolutionary models with partial homology to classify all protein-coding genes in nine *Drosophila* genomes. Compared with two other popular methods for gene family construction, our approach improved practical gene family classifications with a more reasonable view of homology and provided a much more complete landscape of gene family evolution at the gene and subgene levels. In the case study, we found that most expanded gene families might have evolved mainly through module rearrangements rather than gene duplications and mainly generated single-module genes through partial gene duplication, suggesting that there might be pervasive subgene rearrangement in the evolution of protein-coding gene families. The use of a phylogeny-based approach with partial homology to classify and analyze protein-coding gene families may provide us with a more comprehensive landscape depicting how genes evolve within a well-known genome phylogeny.

Key words: gene family, partial homology, module architecture, subgene rearrangement, evolution.

Introduction

With the sequencing of increasing numbers of complete reference genomes of species of interest, it has become increasingly essential and important to accurately and reasonably identify protein-coding gene families genome wide, which are usually used as inputs for many aspects of phylogenomic downstream analysis, including significant family expansion and contraction (Hahn et al. 2005, 2007; De Bie et al. 2006; Demuth and Hahn 2009; Armisén et al. 2018), gene genealogical inference (Ane et al. 2006; Benton 2015; Cortesi et al. 2015; Szöllősi et al. 2015), species phylogenetic reconstruction (Maddison 1997; Delsuc et al. 2005; Liu and Pearl 2007; Degnan and Rosenberg 2009), and gene tree reconciliation (Degnan and Rosenberg 2009; Doyon et al. 2011; Ness et al. 2011; Nakhleh 2013). Although such identification has been very successful in phylogenomics, accurately and reliably reconstructing homologous relationships between genes and identifying homologous gene families are still challenging. During the course of genome evolution, many genes evolve as a whole unit and undergo various processes, such as gene duplication, loss, and horizontal transfer (Zhang 2003; Kazazian 2004; Innan and Kondrashov 2010; Doyon et al. 2011), and such genes should be homologous at the gene level. In addition, accumulating evidence from phylogenomics studies suggests that there are also many genes evolving at a subgene level, caused by evolutionary events such as gene segment duplication, fission, fusion, insertion, and deletion (Ekman et al. 2007; Ding et al. 2012; Wu et al. 2012; Meheust et al. 2016; Sibbald et al. 2019), among others. These evolutionary modes at the subgene level complicate the homologous relationships between genes and lead to the partial homology of a given set of genes (Fitch 2000;

McInerney et al. 2011; Haggerty et al. 2014), thereby inevitably affecting the accurate and reliable assignment of gene families.

In recent decades, many, if not most, evolutionary patterns at the subgene level have been examined under the concepts of protein domain and domain architecture, and many important results have been obtained. These studies uncovered the evolutionary dynamics of protein domains across large phyla, including bacteria, animals, fungi, and plants (Bjorklund et al. 2005; Ekman et al. 2007; Wang and Caetano-Anollés 2009; Kersting et al. 2012; Moore and Bornberg-Bauer 2012), and investigated the genetic mechanisms underlying domain rearrangement from domain architecture information (Bornberg-Bauer et al. 2005; Ekman et al. 2007; Fong et al. 2007; Moore et al. 2008; Buljan et al. 2010; Moore et al. 2013). These findings have furthered our understanding of how protein modularity facilitates rapid adaptation as well as species diversity (Ekman et al. 2007; Zmasek and Godzik 2011; Kersting et al. 2012; Bornberg-Bauer and Alba 2013; Moore et al. 2013). Currently, with the availability of high-quality, fully sequenced genomes, which are well resolved and closely related, we may be able to study protein modular evolution more thoroughly by using de novo-identified subgene units instead of domains. When aiming to investigate the evolutionary details of protein modular evolution within closely related species, domains might not be suitable for three reasons. First, protein domain detection relies on the predefined domain models in domain databases. These domain databases, such as Pfam (Sonnhammer et al. 1997; El-Gebali et al. 2019) and SCOP (Murzin et al. 1995; Andreeva et al. 2014), first require numbers of domain instances across species to build hidden Markov models or sensitive position-specific scoring matrices (Eddy 1998; Wilson et al. 2007). Accordingly, if a domain has just one or fewer instances, especially in the case of domains specific to poorly studied organisms, it will be overlooked and not presented in the domain databases (Bjorklund et al. 2005; Zmasek and Godzik 2011). Second, domain detection is dependent on the chosen *E* value cutoff. Domains may have evolved too fast and diverged beyond detection, especially in some clades. Lowering the *E* value cutoff may allow previously absent domains to become visible. As Moore et al. demonstrated in their study, domain loss is particularly sensitive to variation in the *E* value cutoff (Moore and Bornberg-Bauer 2012). Third, discrete annotated domains inevitably result in the loss of some information for protein sequences. Despite some efforts to increase domain coverage, such as using relaxed constraints or enabling annotation of less characterized domains such as those in ProDom (Servant et al. 2002), a proportion of proteomes still remain unassigned in domain-centric studies (Moore et al. 2008; Zmasek and Godzik 2011; Moore and Bornberg-Bauer 2012). Most importantly, proteins without domain annotation or with long unassigned regions are discarded from analysis (Bjorklund et al. 2005; Moore et al. 2013). As Zmasek et al. carefully stated, their estimates represent a lower bound for domain repertoires (Zmasek and Godzik 2011). This reduction can obviously cause the loss of domains as well as domain architectures. Thus, with such potential biases and incompleteness, any genome-wide analysis of domain evolutionary patterns may more or less miss some events at the subgene level and should also incur the loss of some information, influencing the accuracy of gene family identification. Note that as the conserved units of protein structure, evolution, and function, domains might not greatly influence the results when they are used to investigate large-scale protein evolution across distantly related species. However, for a given set of genes in closely related species, interdomain regions may still contain some useful information. It is necessary to include interdomain regions as complete as possible to infer the evolutionary process at the subgene level.

A previous study presented a method for de novo discovery of homologous modules solely through sequence similarity, rather than relying on previously known structural or functional domains (Wu et al. 2012). A module is a gene subsequence inherited as a basic unit without internal breaks or rearrangements across the species under comparison. After finding modules, the authors compared the modules with domains (annotated from Pfam-A) as well as exons in terms of size and boundary distance and found that the modules tended to be close to domains or extend farther than the closest domains. The presence of multiple consecutive domains in some modules may be because these modules did not have sufficient time to rearrange. Thus, they stayed together, and the whole segment could be considered a unit, at least in the *Drosophila* clade, to infer the genetic changes underlying species adaption and phenotype diversity. Indeed, the authors found a large percentage of modules lying precisely at an exon boundary, and symmetrical intron phases were enriched due to the presence of 0–0 modules, whose flanking introns were both in phase zero. These findings suggested that modules are frequently produced through exon shuffling. As the study revealed, the modules are biologically meaningful and consecutively cover the protein sequences, and they can be used to trace the evolutionary history of clade-specific modules or modules that are not found in current databases. Obviously, given the set of well-annotated genomes, these modules seem to be less biased, and furthermore, each protein sequence can be nearly completely and consecutively composed of one or more different identified modules. Using modules, Wu et al. (2012) developed an architecture-aware pipeline to reconstruct protein modular evolution. Their study reflected the distributions of module events and identified the related functions and possible mechanisms of gene fusion and fission. Because their major goal was to trace gene evolution at the module level, they did not construct gene families or analyze gene family evolution. Therefore, we may

extend the use of de novo modules and improve the accuracy of gene family identification.

By accounting for gene evolution at the module level as well as at the gene level, we can reconstruct the comprehensive gene evolutionary process, which can reflect how partial homologs occur and help us construct homologous gene families. The majority of current software tools to identify gene families rely on sequence comparisons and clustering. These tools first obtain pairwise similarity scores from an all-versus-all search by using BLAST (Camacho et al. 2009) or DIAMOND (Buchfink et al. 2015) or MMseqs2 (Steinegger and Söding 2017). After sequence similarity network construction, the clustering of highly similar genes is generally based on a clustering algorithm such as the Markov Cluster Algorithm (Enright et al. 2002), Louvian (Blondel et al. 2008), and single linkage method. The widely used methods (Enright et al. 2003; Li et al. 2003; Östlund et al. 2009; Emms and Kelly 2019) take different ways to analyze sequence similarity scores and produce different outputs. Among them is OrthoFinder, which normalizes similarity scores for gene length and phylogenetic distance and results in significant improvements in accuracy (Emms and Kelly 2015). Because most genes evolve in their entirety in a tree-like way, these methods can be run quickly and efficiently to construct gene families. In subsequent phylogenetic analysis of these gene families, we can depict gene duplication, gene loss, and gene transfer events (Doyon et al. 2011). In addition to these methods, OrthoDB is a widely used resource that delineates orthologs at varying resolution by referring to the hierarchy of species radiations (Waterhouse et al. 2013). However, some of the genes that evolved at the subgene level (McInerney et al. 2011) are constructed from different gene parts and have separate origins, and if we still partition them into groups based on similarity along almost their entire length, we will be unable to acquire a complete gene evolutionary history, whether using a tree (Omland et al. 2008; Nakhleh 2013) or a network model (Huson and Scornavacca 2011; Corel et al. 2016) to infer the gene phylogeny, resulting in information loss. Considering that genetic sequences may show similarity for partial sharing of component fragments, some methods can describe the genetic parts shared between gene families and detect composite and component gene families. CompositeSearch (Pathmanathan et al. 2018) generalized the use of sequence similarity network and outperformed the recent methods FusedTriplets and MosaicFinder (Jachiet et al. 2013). Although such families theoretically can be used as inputs to construct an *N*-rooted network (Haggerty et al. 2014) to reconstruct gene family evolution, especially for multimodule genes, relative algorithms and tools are lacking and need to be developed. In addition, the components identified in a composite gene might not be consecutive (Pathmanathan et al. 2018). To properly assign homologous genes to gene families, a more reliable and realistic method that considers the gene phylogeny to detect real homology rather than homology based purely on sequence similarity is needed. Instead of these two types of similarity-based approaches, we assume that a phylogeny-based approach that considers the origins of genetic fragments would better describe sequence relationships and classify gene families.

In this work, we extract all protein sequences of protein-coding genes from a set of closely related genomes with a known and reliable species phylogeny and use all possible constitutive subgene units for the set of protein sequences identified by Wu et al. (2012). Combining the modules and the set of protein sequences, all different extant module architectures (MAs), which are linear arrangements of each module along each sequence, can be constructed. Then, the evolutionary scenarios of the extant MAs can be inferred using STAR-MP (Wu et al. 2012), which is a maximum parsimony method based on a well-known species tree, extant Mas, and reconstructed module trees. We present a method called RASfam that can be used to construct homologous gene families. By focusing on the reconstructed architecture scenarios (RASs), we can inspect how the extant MAs originated and assign them to different groups accordingly. Next, the respective gene families are constructed for the set of closely related genomes. Most importantly, we allow the proteins whose parts have different origins to be assigned to more than one family. We think that such assigned gene families are more realistic and reasonable from an evolutionary perspective because they should reflect the complex, partially homologous relationships between genes during the course of genome evolution. For a case study, we demonstrate this idea using nine *Drosophila* genomes with a well-resolved phylogeny (Tamura et al. 2004). Indeed, we observe evolutionary patterns of the protein-coding genes that are more comprehensive than those in previous studies. Interestingly, we found that most of the expanded gene families might have evolved mainly through module rearrangement events rather than gene duplication events, especially those driven by partial gene duplication and forming single-module architectures. These results will provide us with a better understanding of complicated evolutionary patterns, in particular partially homologous relationships, of protein-coding genes in a set of closely related species.

Materials and Methods

Proteins, Modules, and Species Phylogeny

We selected nine species within the *Drosophila* genus, namely, *D. melanogaster* (dmel), *D. yakuba* (dyak), *D. erecta* (dere), *D. ananassae* (dana), *D. pseudoobscura* (dpse), *D. willistoni* (dwil), *D. mojavensis* (dmoj), *D. virilism* (dvir), and *D. grimshawi* (dgri), with a known species tree (Tamura et al. 2004). We analyzed the longest protein sequence for each protein-coding gene. The sequences and annotations are from FlyBase, and the version corresponds to that of the identified modules (Wu et al. 2012). The module identification

by Wu et al. was from a protein comparison by BlastP (Altschul et al. 1997), alignment extension by LALIGN (Huang and Miller 1991), module boundary detection by the ADDA algorithm (Heger and Holm 2003), and module family clustering by OrthoMCL (Enright et al. 2002).

Definitions

A module, a gene subsequence, is a single unit without internal rearrangements or breaks and can be inherited among species. The identified modules almost cover all proteins, which can be seen as composed of one or several modules without gaps. This method, without relying on domain annotations from databases, guarantees the completeness of the modules. We defined the "module architecture" (architecture) of each protein sequence as the ordered list of modules that it contained. The whole workflow can be seen in figure 1. After defining architectures, we constructed an architecture similarity network in which each node represented an MA and two nodes were connected by an edge if they shared at least one common module. From the network, we obtained 4,173 architecture connected components (ACCs), each of which had at least one path connecting any pair of two nodes, and 10,145 architecture singletons (ASTs), which shared no common modules with other architectures. Notably, we discarded 433 proteins that consisted of discontinuous modules, which resulted in our ACCs not being exactly the same as Wu's architecture family (total of 4,107). After excluding 3,708 ACCs with inferred scenarios that were consistent with those of Wu, we reconstructed the evolutionary histories of the 465 remaining ACCs. Detailed information can be found in supplementary table S1, Supplementary Material online.

We present a phylogenetic workflow for reconstructing gene evolution at the gene and subgene levels (fig. 1*A* and *B*). For each ACC, architectures with partially homologous relationships but derived from different module combinations likely evolve at the module level. The evolutionary units of ACCs are modules instead of whole genes. Therefore, we used an architecture-aware phylogenetic pipeline (Wu et al. 2012) to reconstruct ACC evolutionary scenarios and detect five types of module-level evolutionary events, namely, duplication, loss, merging, splitting, and emergence (fig. 1*C*), which represent the processes in which an ancestral module was duplicated, lost, or merged with another ancestral module, an ancestral MA split into the extant modules, and a novel module emerged, respectively. For each AST, one architecture exists alone, and its corresponding proteins can be seen to evolve at the gene level. We deployed a common phylogenomic workflow to reconstruct AST scenarios for inferring gene duplication and gene loss events.

Phylogenetic Analysis of ASTs

ASTs that evolve at the gene level fit a tree-like evolutionary model. For each AST, amino acid sequences were aligned using MUSCLE v3.8.31 (Edgar 2004) and then translated to nucleotide alignments by TranslatorX (Abascal et al. 2010) with the parameters -a -i -o. We performed model selection by jModelTest v2.1.10 (Posada 2008) with the parameters -f -g4 -BIC and without prediction of a proportion of invariable sites. Gene trees were constructed by PhyML v3.1 (Guindon et al. 2010) with 100 bootstraps and rooted by Notung v2.9 (Chen et al. 2000). Next, we used TreeFix v1.1.10 (Wu et al. 2013) with a 1:1 duplication:loss cost ratio to reconcile gene trees with the species tree. The resulting reconciled gene trees were the optimal trees with maximal likelihood and the minimum reconciliation cost. Finally, we annotated gene duplication and gene loss events on the species tree by TreeFix v1.1.10 (fig. 1*B*).

Phylogenetic Analysis of ACCs

We reconstructed the evolutionary scenario of each ACC by three steps (fig. 1*B*). First, as the evolutionary unit of an ACC is a module, we performed phylogenetic analysis of each module independently and constructed a module tree. Then, we reconciled module trees with the well-known species phylogeny repeatedly to determine module emergence, duplication, and loss. Finally, with the species phylogeny, the extant architectures in each species and the final reconciled module trees, we reconstructed ancestral MAs and inferred module evolutionary events on each branch of the phylogeny.

In the first step, we constructed module trees for each module family. For each module family, we extracted corresponding peptide sequences and aligned them with MUSCLE v3.8.31. The nucleotide alignments were then retrieved by TranslatorX with the parameters -a -i -o. To identify model parameters for SPIMAP v1.1 (Rasmussen and Kellis 2011) that could be used to reconstruct module trees, we used CAFE v4.1 (De Bie et al. 2006) to predict the birth and death rates of 9,894 module families that covered all species, and the substitution rate was determined by SPIMAP v1.1 from 7,150 single-copy module family trees that were reconstructed by PhyML v3.1. Module families with small sequences that could not be used to construct module trees were discarded.

In the second step, we used maximum parsimonious reconciliation to reconcile each module tree with the species tree. We sampled 100 module trees with replacement for the 100 reconstructed trees of each module family. We reconciled each sample tree with the species tree using FORESTER v1.050 (Zmasek and Eddy 2001) and removed any duplication nodes that predated the species tree root (preroot duplications). The resulting subtrees were rerooted and reconciled repeatedly using FORESTER until no more preroot duplications were observed.

In the third step, we combined all reconciled module trees of each ACC to reconstruct its evolutionary scenario. With the reconciled module trees, the known species phylogeny and all extant architectures of each species, we used STAR-MP v1.0

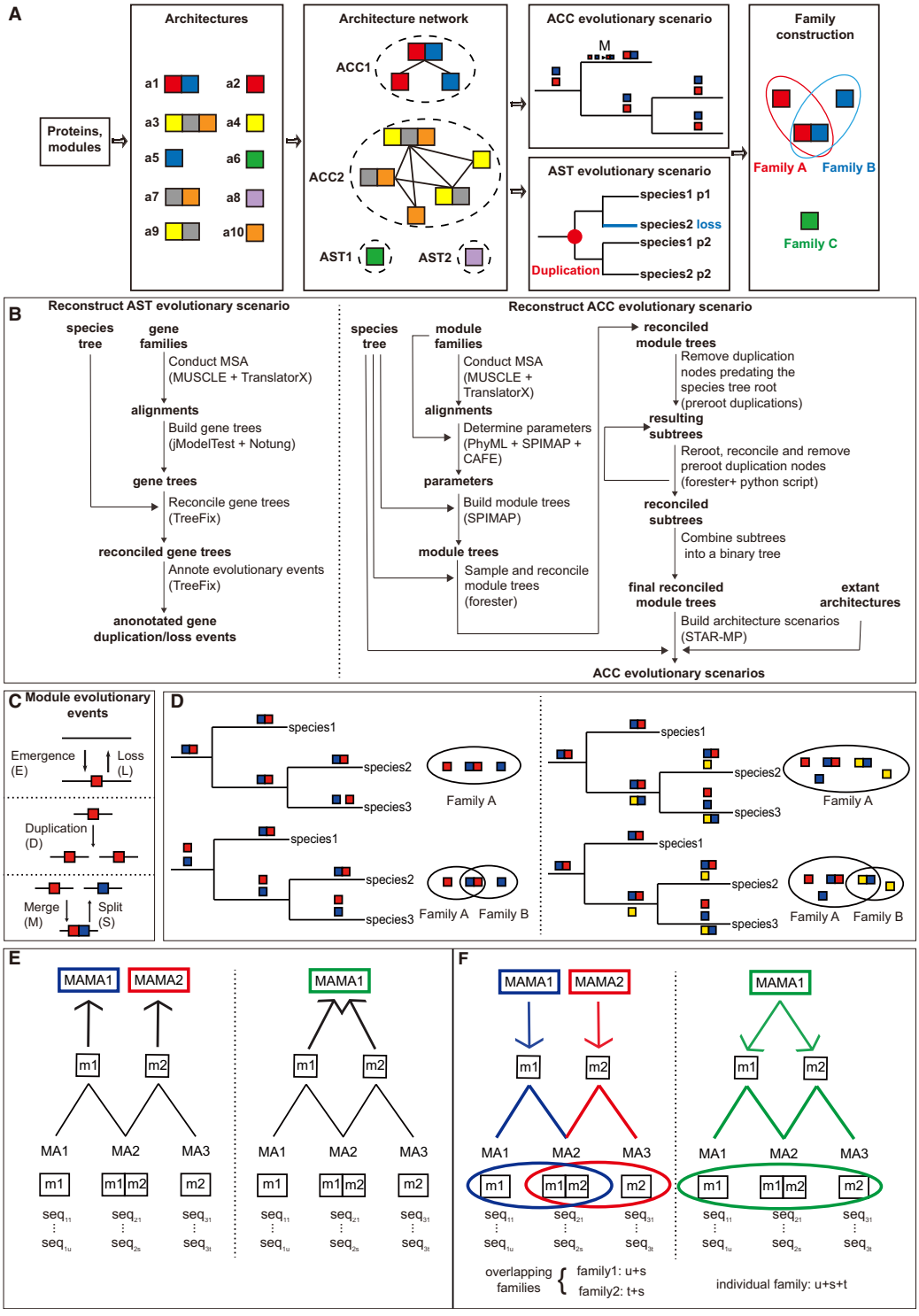

Fig. 1.—Workflow used to analyze gene evolution at the gene and subgene levels. (*A*) Overview of the three major steps. (*B*) Reconstruction of the gene evolutionary history for each AST and ACC. (*C*) Five types of module evolutionary events. (*D*) Homologous gene family construction based on the RASs. The gene family to which a MA belongs depends on how the MA originated. (*E*) The schematic diagram of the RASfam algorithm demonstrates how to determine the most ancient module architecture(s) (MAMAs) of an extant MA. We traced the MAMA of each extant MA by detecting the origin of each of its modules. Meanwhile, the number of the inferred MAMAs is derived from the result of architecture scenario reconstruction. For example, as shown on the left side of (*D*), the top ACC reconstructed one MAMA, while the bottom two, although both ACCs presented two extant single-module architectures and one extant multimodule architecture. When RASfam is applied, "m1" and "m2" in (*E*) represented the blue and red module (architecture), respectively. Two scenarios were demonstrated: On the left side of (*E*), "MAMA1" and "MAMA2" also represent the blue and red module (architecture), respectively, and the extant blue-red architecture descended from MAMA1 and from MAMA2. On the right side of (*E*), "MAMA1" represented the blue-red architecture, from which the extant blue-red architecture originated. (*F*) The schematic diagram of the RASfam algorithm demonstrates how to construct the respective gene family of a MAMA.

(Wu et al. 2012), a maximum parsimony method, to reconstruct the evolutionary history of each ACC and infer the module-level evolutionary events.

Homologous Gene Family Construction

We named our phylogenetic approach RASfam: A RAS-based approach for homologous gene family construction (supplementary algorithm, Supplementary Material online). Here, a homologous gene family is defined as a set in which all of the members share, at least partly, a common evolutionary origin. This definition implies that different parts of all the sequences of one gene family may have different histories. Each AST thus forms an individual family. Intuitively, it seems difficult to classify gene families for each ACC. Surprisingly, with the RASs, analysis becomes simple, but the results are reasonable (fig. 1*D*). For each ACC, there were at least two extant MAs consisting of at least two different modules. For each extant MA of the ACC, we first traced its most ancient MA (MAMA) by detecting the origin of each of its modules based on the RAS (fig. 1*E*). The modules may derive from different MAMAs or from the same MAMA. Then, we can reliably and naturally construct the corresponding gene families for the ACC according to the following rules: 1) the number of gene families being classified is equal to the number of MAMAs inferred and 2) each extant MA, and therefore its corresponding sequences, will belong to one or more families in accordance with its path(s) linked to one or more MAMAs, respectively (fig. 1*F*).

Gene Family Comparison

To assess the difference between our workflow and other methods for defining gene families, we selected OrthoFinder v2.3.8 (Emms and Kelly 2019) and CompositeSearch (Pathmanathan et al. 2018) as representatives, the former of which views a gene as a whole unit and the latter of which considers partial sharing of gene fragments, and used them to define gene families independently. We used OrthoFinder to construct gene families with the parameters -f -t 40 -a 10. Gene families constructed by CompositeSearch were first analyzed using BlastP (Camacho et al. 2009) with the parameters -query -out -seg yes -soft_masking true -max_target_segs 5000 -outfmt -db -num_threads 50, then using cleanblastp with the parameters -i -n 1, and finally using the compositeSearch command with parameters -i -n -m composites -e 1e-05 -p 30 -c 80 -l 20 -t 1. We then calculated some indices of the gene families and calculated the relationship between our families and families defined by other methods. For 10,145 AST families and 3,245 ACC-derived nonoverlapping families, we classified the relationships into 5 types: "identity," "included," "inclusion," "overlap," and "inclusion and overlap" (fig. 2*A*). Multimodule architecture is abbreviated as "MMA." In figure 2*A*, yellow solid circles denote homologous gene families, and gray hollow circles denote gene families identified by OrthoFinder or CompositeSearch. For these relationships, "included" means that a homologous gene family has a smaller size and is included in the corresponding gene family identified by the methods and vice versa. For homologous gene families in this category, "included-A" means that other homologous gene families overlap with the family identified by the methods, and "included-B" means that other homologous gene families are also included in this gene family. "Overlap" means that a homologous gene family shares some but not all of its members with some related gene families identified by the methods. "Inclusion and overlap" means that a homologous gene family includes gene families and overlaps with other gene families that were identified by the methods. Next, for 1,832 ACC-derived overlapping families, the relations were divided into 3 types: "MMA partially divided," "MMA solely divided," and "MMA partially and solely divided." These relations describe the fate of an MMA based on methods to construct gene families. "MMA solely divided" means that proteins with an MMA are assigned to a unique family. "MMA partially divided" describes proteins with an MMA that are assigned to different families. "MMA partially and solely divided" indicates that some MMAs in an ACC are assigned to a unique family and some MMAs are partially divided into different families.

Size Changes of the Gene Families

From a total of 15,222 gene families, we selected 7,820 families in which each species had at least 1 protein. We used CAFE v4.1 to estimate one global lambda (lambda = 0.00061) and analyzed the size changes of these gene families. Using $P < 0.01$, we extracted the significant results. Gene families with significant expansion or contraction on only one leaf branch were considered species-specific expanded or contracted families.

Others

Gene ontology (GO) enrichment analysis was performed using the ClusterProfile R package, and the enriched GO terms and their corresponding adjusted *P* values were summarized and visualized using REVIGO (http://revigo.irb.hr/, last accessed March 10, 2020)(Supek et al. 2011). Other visualization was performed by the following tools: EvolView (Zhang et al. 2012) and R v3.3.3. Data in the study were processed and analyzed with Perl, Python, and R scripts.

Results

Constructed Gene Families with Partially Homologous Relationships

Applying our analysis pipeline RASfam (fig. 1) to the set of 111,208 protein sequences from the 9 genomes, we detected

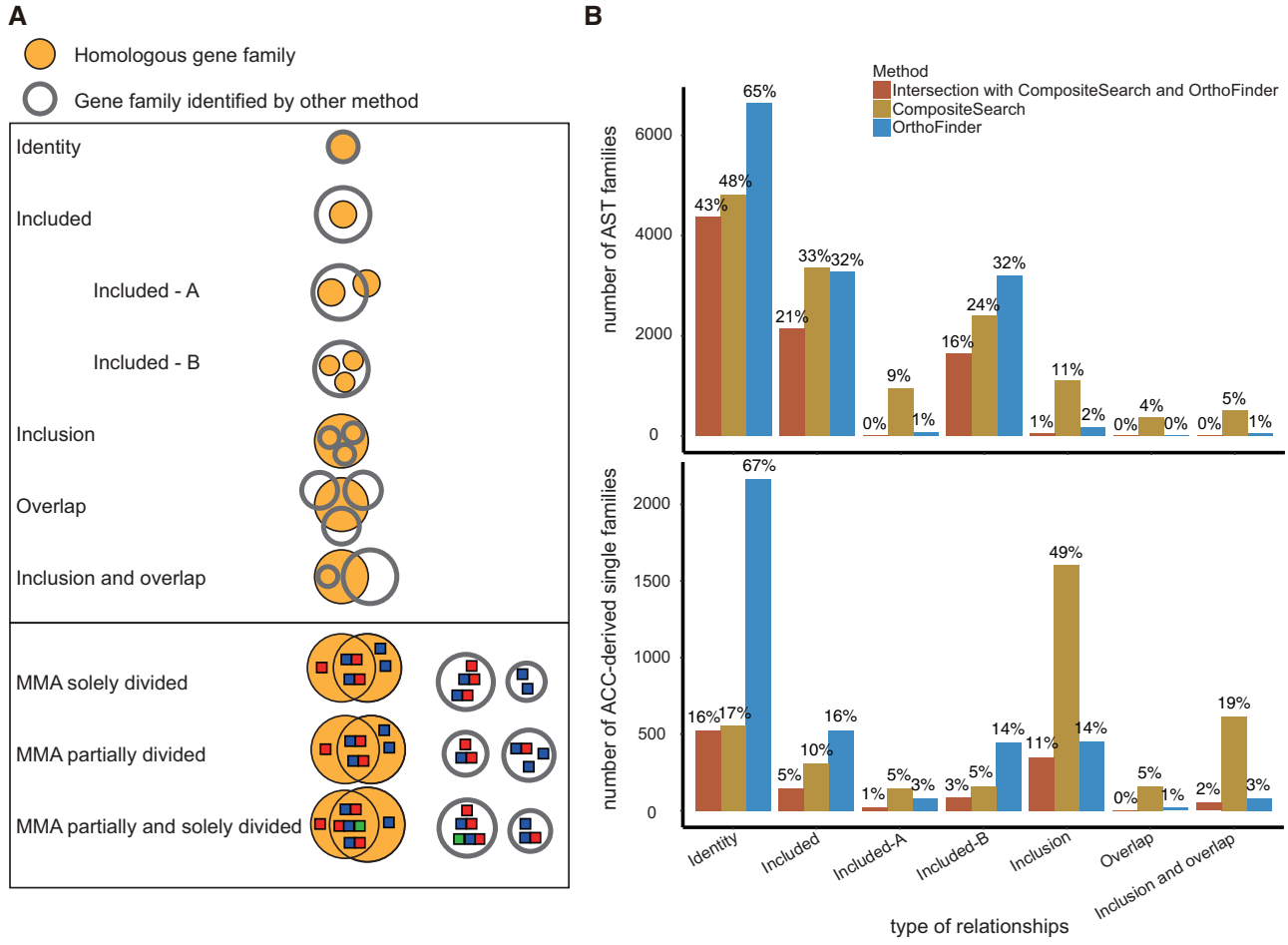

Fig. 2.—The relationships between different gene families. (*A*) We allow for five types of relationships and two detailed relations of "included" ("included-A" and "included-B") for AST families and ACC-derived single families and three types of relationships to describe ACC-derived overlapping families. For a homologous gene family that was included, "included-A" means that other homologous gene families overlapping with the family were identified by the methods, and "included-B" means that other homologous gene families were also included in this family. Multimodule architecture is abbreviated "MMA." "MMA solely divided" means that proteins with an MMA were assigned to a unique family. "MMA partially divided" describes proteins with an MMA that were assigned to different families. "MMA partially and solely divided" represents cases in which some MMAs in an ACC were assigned to a unique family and some MMAs were partially divided into different families. Yellow solid circles denote homologous gene families, and gray hollow circles denote gene families identified by OrthoFinder or CompositeSearch. (*B*) Distribution of relationships for homologous gene families and gene families obtained with other methods.

22,840 different modules and 24,312 different MAs. Assuming that an edge existed when 2 MAs shared at least 1 module, an architecture network comprising 4,173 ACCs and 10,145 ASTs was created. The RASs of 4,018 ACCs were inferred (supplementary table S1, Supplementary Material online). Among these ACCs, 310 were computed by ourselves, and 3,708 were computed previously by Wu et al. (2012). By carefully examining these RASs (fig. 1*D* and Material and Methods), we identified 5,077 ACC-derived families. Importantly, of the 5,077 ACC-derived families, 1,832 were derived from 790 ACCs; thus, they shared partial homologs with other families (supplementary table S2, Supplementary Material online). With the 10,145 AST families, we constructed 15,222 gene families with a mean family size of 6.22 genes. As shown in table 1, 82.3% (12,527) of these gene families were single-copy families with at most 1 gene for each species. Of these single-copy families, 29.7% (4,520) were one-to-one orthologous across all 9 species. In addition, only 6.5% (992) were species-specific families.

To assess the quality of the constructed gene families, we applied the recently popular approach OrthoFinder (Emms and Kelly 2019) and the newest method considering partial sharing of sequence fragments, CompositeSearch (Pathmanathan et al. 2018), to the same set of protein sequences. As shown in table 1, OrthoFinder identified the fewest families (14,220), whereas CompositeSearch identified the largest number of families (21,733). Surprisingly, more than one-third of the families identified by

Table 1
Indices of Different Gene Families

Method	Family Size		No. of Families				
	Mean	Median	Total	ssF	scF	scF-allSP	F-allSP
RASfam	6.22	6	15,222	992 (6.5%)	12,527 (82.3%)	4,520 (29.7%)	6,155 (40.4%)
OrthoFinder	7.82	9	14,220	981 (6.9%)	10,977 (77.2%)	5,994 (42.2%)	8,235 (57.9%)
CompositeSearch	5.12	1	21,733	8,311 (38.2%)	19,457 (89.5%)	4,028 (18.5%)	5,416 (24.9%)

NOTE.—ssF, species-specific families; scF, single-copy families; scF-allSP, single-copy families with all species present; F-allSP, families with all species present.

CompositeSearch were species specific, whereas only ~6.5% of those identified by RASfam and 6.9% of those identified by OrthoFinder were species specific.

Among the 10,145 AST families that we identified, ~43.1% were identically detected by the 3 different methods (fig. 2*B*). Specifically, most of the AST families (~65.4%) were identical to those detected by OrthoFinder, indicating that, for the genes evolving in their entirety, RASfam and OrthoFinder may construct very similar gene families. Interestingly, we found that ~32.2% of the AST families were contained in the OrthoFinder families, suggesting that the identified modules were more comprehensive and thus that the associated MAs were more complete.

On the other hand, based on the STAR-MP inference results for the 4,035 ACCs, 790 of the ACCs were split into smaller groups, and 3,245 of them remained as a whole set of related MAs. For these 3,245 ACC-derived families that did not share proteins with other families (called nonoverlapping families), 66.8% were identical to the OrthoFinder families (fig. 2*C*). However, only ~17.1% were identical to the CompositeSearch families, and more, ~49.4%, of the families contained two or more CompositeSearch families, indicating that RASfam may capture much more complex partially homologous relationships than CompositeSearch. For the 790 ACCs that needed to be split, we detected 1,832 ACC-derived overlapping families that shared proteins with other families. Nonetheless, we found that OrthoFinder constructed 468 of the 790 ACCs as nonoverlapping families, most of these families were designated as MMAs solely, CompositeSearch identified 201 ACCs as nonoverlapping families, and most of these families were designated as MMAs solely (supplementary table S3, Supplementary Material online). For example, the RabX5-PB family and RpL23A-PA family are good illustrations of an MMA being solely divided (fig. 3). In this case, our method allowed the multimodule protein to belong to two families, whereas OrthoFinder and CompositeSearch assigned it to only one family (fig. 3*A*). Importantly, the corresponding RAS provided by our workflow reflected how the partial homology originated, which evolutionary events might have occurred along the genome phylogeny, and when those events might have occurred (fig. 3*B*). Interestingly, such complex partially homologous relationships between genes could be modeled in an *N*-rooted gene network ($N = 2$ in this case) representing the gene remodeling history (Haggerty et al. 2014), from which the genealogy was inferred by using the RAS as a guide (fig. 3*C*).

Gene Families with Significant Size Changes Inferred Using CAFE

Based on the classification of the gene families for all the protein-coding genes of the nine closely related genomes, in general, statistical significance of the family size changes could be inferred across the species phylogeny by using stochastic birth–death process-based models such as CAFE (De Bie et al. 2006). For each family that has at least one member in each genome, this inference may reveal the pattern of significant family size changes, including expansion or contraction, suggesting that adaptive responses to selection on some biological processes might have occurred through these changes in the gene families (Rubin et al. 2000; Francino 2005; Innan and Kondrashov 2010). Among the 7,820 families (AST: ACC derived = 1.56) that satisfied the inference condition, we found that 405 (5.18%) underwent significant size changes (P value < 0.01). For the 242 significantly expanded families with various biological functions (supplementary table S4, Supplementary Material online), we found 15 GO terms related to 22 families, such as defense response, oxidoreductase activity, and odorant binding, among others (supplementary table S5, Supplementary Material online), which have previously been identified in expanded families in the respective species (Hahn et al. 2007).

Interestingly and importantly, as shown in figure 4, the majority of the significantly expanded families were ACC derived in the respective lineages ($\chi^2 = 220.96$, df = 1, $P < 2.2e\text{-}16$). In addition, for the 191 species-specific families that were significantly expanded, a similar trend was observed in all species except *D. pseudoobscura* (ACC derived: all = 0.5) (supplementary fig. S1, Supplementary Material online). Hence, such expanded ACC-derived families should allow us to investigate potential evolutionary patterns at the subgene level, which may reveal the partially homologous relationships between the genes in those families. This is important because, due to CAFE modeling only the family size changes, nothing from the inference can reveal the detailed sequence relationships, in particular the partial

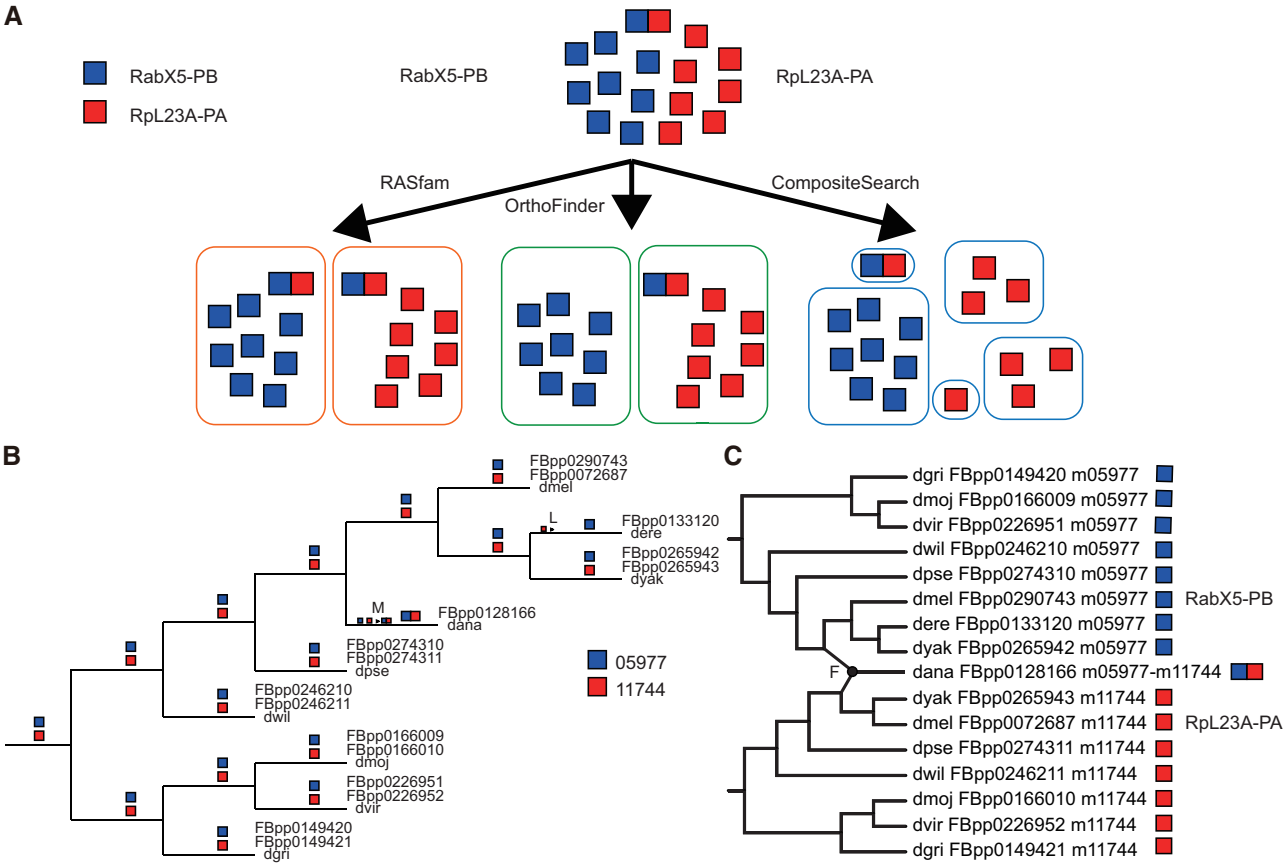

Fig. 3.—The inferred gene families and evolutionary history of *RabX5-PB* and *RpL23A-PA*. (*A*) The different families were constructed using different methods. (*B*) The architecture scenario of the ACC contains *RabX5-PB* and *RpL23A-PA*. M denotes a merge event, and L denotes a loss event. (*C*) The sequence evolutionary history inferred by combining two module trees. Each module tree was reconstructed by SPIMAP and reconciled with the species tree using maximum parsimonious reconciliation. The two module trees were merged manually with Adobe Illustrator software. F denotes a fusion event occurring on the node with an in-degree of two.

homology between genes in each of the significantly expanded families.

We used the architecture scenarios reconstructed by STAR-MP to trace how the architectures in each focal family and the number of their corresponding sequences might have changed between each of the extant species and its most recent ancestor. For example, significant expansion in the sperm-specific dynein intermediate chain (*Sdic*) gene family in *D. melanogaster* resulted in five copies, whereas only one copy was observed in its most recent common ancestor. Based on the scenario inferred by STAR-MP, these four new *Sdic* copies might have originated from partial duplication of the ancestral *Cdic* gene (supplementary fig. S2, Supplementary Material online). *Sdic* copies were previously reported to be fixed by adaptive responses to natural selection (Nurminsky et al. 1998, 2001) and to have evolved a novel function that might enhance male fitness (Yeh et al. 2012). Another example is the *Cyp6a21/Cyp6a9* family, which has expanded from one copy to five copies in *D. mojavensis*, including three duplicated genes and one novel gene. This novel gene's MA existed in the extant family but was absent in its recent ancestor's family (supplementary fig. S3, Supplementary Material online). The expansion of CYP-related genes may specialize the detoxification ability of *D. mojavensis*, allowing it to tolerate both toxic cactus necroses and high-desiccation deserts (*Drosophila* 12 Genomes Consortium et al. 2007; Markow and Ogrady 2007). More interesting examples are shown in supplementary figures S4–S6, Supplementary Material online.

Pervasive Module Rearrangements in Expanded Gene Families

Once incorporating evolution at the subgene level into inference on size changes of gene families, the evolutionary patterns will be complex, as exemplified by the aforementioned patterns. In addition to gene duplication, other events such as module duplication, loss, fusion, fission, and emergence might also have occurred, leading to architecture formation in respective gene families. Note that architecture formation is

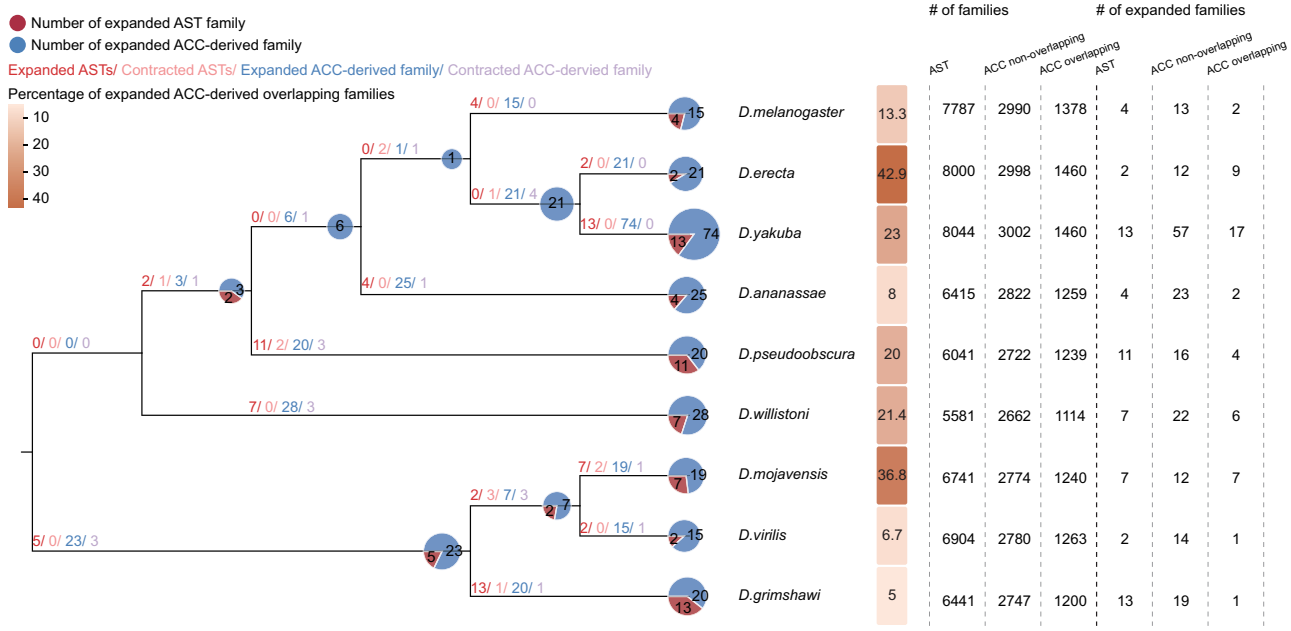

Fig. 4.—Gene family expansion and contraction. The number of family expansions and contractions is given on each branch of the species tree. The colors of numbers represent the amount of size change of the corresponding family type. The pie shows the expanded gene families: Its red part represents the AST families, and the blue part indicates the ACC-derived families. The percentages of significantly expanded ACC-derived overlapping families in all expanded ACC-derived families along each leaf branch are shown in the heatmap.

inferred when a novel architecture is present in extant species but absent in the corresponding ancestor or when an architecture has more copies in extant species than in the corresponding ancestor. Such an evolutionary pattern in each family is also important for a species, even though family size might not change or change insignificantly. To explore evolutionary patterns at the subgene level, we focused on the 5,077 ACC-derived families and investigated their architecture change patterns based on the evolutionary scenarios inferred by STAR-MP.

For the simplicity of comparison, only two levels are contrasted for each gene family, namely, the family of the extant species and that of its recent ancestor. A simple index called foldchange is defined as the ratio of extant species family size to its recent ancestor family size. The overall distribution of the foldchanges can be calculated accordingly (supplementary fig. S7, Supplementary Material online). Meanwhile, for different types of foldchanges, we examined the evolutionary patterns of gene families in the respective species (supplementary table S6, Supplementary Material online). Most branches were completely unchanged (25,998, 66.8%), and many only lost genes (5,189, 13.3%). The unchanged architectures and copy numbers may simply be a result of the short time that has passed within the closely related species or may be important for some gene families. There were 187 unchanged ACC-derived families and 4,039 unchanged AST families in the 9 species. We found that the enriched GO term of the AST families was related to "mRNA splicing, via spliceosome" and that the major functions of the ACC-derived families were involved in some basic biological processes, such as replication, protein import, metabolism, cellular respiration, and digestion. In addition, we observed that 86 ACC-derived families were unchanged only in the *Drosophila* subgenus and that 92 ACC-derived families were unchanged only in the *Sophophora* subgenus. Interestingly, the enriched GO terms for these unchanged families were different between the two subgenera; they were related to rhodopsin biosynthesis, synapse organization, and sleep in *Drosophila* subgenus, whereas in the *Sophophora* subgenus, they were mainly related to the Notch signaling pathway, the open tracheal system, cilium organization, locomotor rhythm, and other processes (supplementary fig. S8, Supplementary Material online).

Except for the branches presenting no changes or only loss events, supplementary figure S9 and table S6, Supplementary Material online, present the distributions of the lost, duplicated, and novel genes for each family in each species. Surprisingly, novel genes were generated uniquely in some of the families with a foldchange <1 (1,244, 19.2%), indicating that new architectures might have occurred, although most ancestral genes were lost. For the families of unchanged size, a similar pattern was observed: some of them (4,053, 13.4%) generated novel genes while losing all ancestral genes, implying that the types of architectures have completely changed while the number of genes has remained the same. Interestingly, for the families with a foldchange >1, most (1,667, 72.9%) did not lose genes, and many of them (1,150, 69.0%) gained novel genes, indicating that most such

families might have generated novel architectures while maintaining the ancestral architectures as well. Additionally, some of the families (443, 26.6%) without loss events generated only duplicated genes, and a few families (74, 4.4%) generated both duplicated and novel genes. Similarly, for the families with a foldchange >1 and with loss events, most (565, 91.3%) generated only novel genes, and fewer families (25, 4.0%) generated only duplicated genes. This result demonstrates that the ACC-derived families expanded mainly through module rearrangements by creating novel genes instead of undergoing simple gene duplication.

Dominant Module Rearrangement Type for Architecture Formation

Using our pipeline, we also provided a much more complete landscape of gene evolution at the gene and subgene levels. With our 4,107 RASs for ACCs and 8,902 reconstructed genealogies for ASTs, we studied the distribution of each type of evolutionary event (supplementary fig. S10 and table S7, Supplementary Material online). These evolutionary events might lead to architecture formation as well as architecture loss. Architecture formation at the gene level can only increase gene copies, whereas that at the subgene level may lead to novel genes. Along a branch, a single-module evolutionary event may occur with other single-module events, thus causing not only single occurrences but also the cooccurrence of module events. We classified each of the possible combinations of module events as rearrangement events by using abbreviations for combinations of module events such as "D," "S-D," "S-L," and so on (fig. 5). Each rearrangement event could fall into one of seven rearrangement types: module duplication, partial gene duplication, module loss, fission, fusion, terminal loss, and recombination. Note that a "D" rearrangement event along a branch indicates the occurrence of only module duplication, which could result from the duplication of an ancestral single-module architecture or each module of the ancestral multimodule architecture, resulting in whole-gene duplication. Similarly, the "L" rearrangement event results in whole-gene loss. To determine the patterns that might explain the formation of architectures, we examined the rearrangement events along terminal branches.

Module duplication ("D"), module loss ("L"), and fission ("S") were inferred when only the corresponding module event occurred along the branch. Note that fission here requires both products of the split to be present. Partial gene duplication was considered to have occurred when one of the split modules duplicated after the ancestral architecture split ("S-D"). Terminal loss was determined by ancestral architecture split and loss of one of the split modules ("S-L") and if the remaining module duplicated ("S-L-D"). Fusion was inferred when the extant architecture was formed by the simple merging of two ancestral modules ("M"), the merging of two ancestral modules after one or both of them duplicated ("D-M"), the merging of two ancestral modules, whereas the other ancestral module was lost ("L-M"), or even the merging of two ancestral modules, with some ancestral modules duplicating and other being lost ("D-L-M"). Finally, a module from the ancestral multimodule architecture participating in the formation of the extant multimodule architecture was placed in the recombination category, which included cases in which the split module merged with another ancestral module after the ancestral architectural split ("S-M"), one of the split modules was then lost ("S-L-M") or duplicated ("S-D-M") or one of the split modules was lost and the other was duplicated ("S-D-L-M").

In total, 11,045 rearrangement events were present along the terminal branches, including architecture formation and loss. The majority (6,856, 62.07%) of these rearrangements were single-step events, and the remaining (4,189, 37.93%) were multiple-step events (fig. 5), which was consistent with the findings of a previous study on *Drosophila* domains, suggesting that 63.6% of new arrangements have an exact single-step solution (Moore et al. 2013). In contrast, we detected rearrangements that may have been the result of more complex rearrangements. Importantly, a complex chain of events, which might have a high cost but still has a probability of occurring, can explain some architecture formation. In figure 5, we illustrate the percentage distribution of the seven rearrangement types.

Simple module loss and module duplication could be characterized as gene loss and duplication as they did not form novel architectures. Among the five other types (5,979 events), terminal loss (2,187, 36.58%), and fusion (1,996, 33.38%) were the major drivers of novel architectures ($\chi^2 =$ 2,522, df = 4, $P <$ 2.2e-16), and only 4.1% (245) of the novel architectures arose due to fission. It seems plausible that proteins are more likely to lose a nonfunctional fragment than to produce two individual fragments because a mutation is more likely to result in a premature stop codon than to produce new start and stop codons. In addition, compared with fusion, which directly merges ancestral single-module architectures, recombination, which requires ancestral fragment merging by multiple steps, is rare (515, 8.61%). To further understand which rearrangement events have contributed to the formation of single-module and multimodule architectures, we investigated the rearrangement event distributions of these types of architectures among the nine species (fig. 6). For multimodule architecture formation, "M" was the most frequent category in each species, and it is likely that gene fusion is the dominant multimodule architecture formation mechanism (χ^2 test, $\chi^2 =$ 9,479, df = 13, $P <$ 2.2e-16). In contrast, most single-module architecture formation was mediated by terminal loss, except in *D. grimshawi*, in which the most common rearrangement type was partial gene duplication (χ^2 test, $\chi^2 =$ 12,869, df = 13, $P <$ 2.2e-16). Interestingly, partial gene duplication also occurs at a high frequency in *D. yakuba*. The single-module architectures formed by partial gene

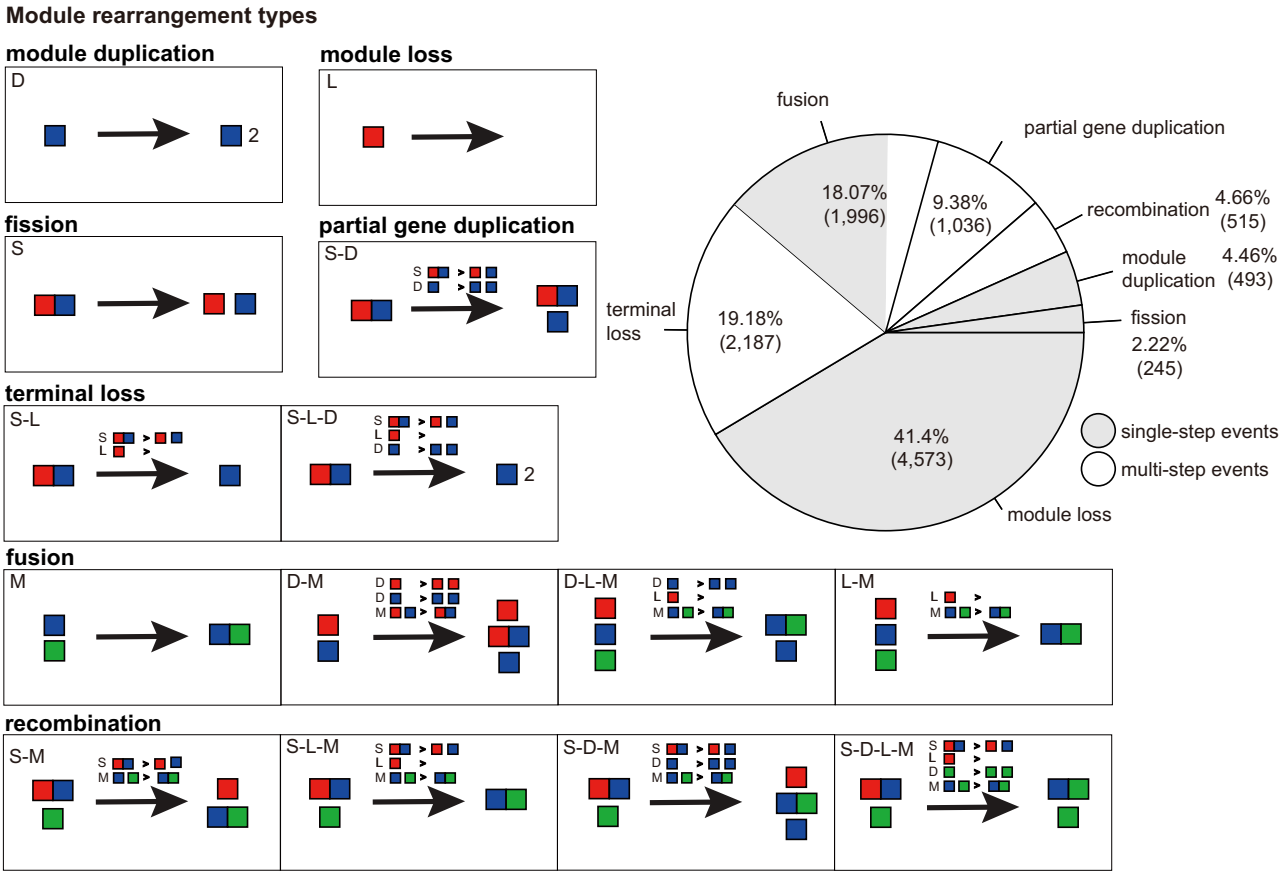

FIG. 5.—Module rearrangement types of ACC-derived families whose size foldchange was >=2 and that generated only novel proteins in corresponding species. (*A*) Examples showing how evolutionary events occur in combination at the subgene level. D, duplication; S, split; M, merge; L, loss. (*B*) Distribution of module rearrangement types in each lineage. The values represent the percentage of expanded families of each type within a species.

duplication in *D. yakuba* were not detected in *D. melanogaster*, and those in *D. grimshawi* had only eight instances in *D. melanogaster*. Even without enough data to perform GO enrichment analysis, we assumed that the single-module architectures that formed along these two independent lineages were related to the species-specific innovative potential of protein. This indicates that module rearrangement can generate diverse architectures in evolutionarily short time periods and have a biased pattern when forming single-module or multimodule architectures.

As we found pervasive module rearrangements in expanded gene families, we wanted to determine how often different rearrangement types occurred when the gene family expanded within the *Drosophila* clade. For 1,383 nonoverlapping ACC-derived families whose foldchange was >1 along the corresponding terminal branch (1,698 expansions in total), we calculated the distribution of rearrangement events (supplementary table S8, Supplementary Material online). Based on whether the formed architecture was of a single module or multiple modules, the expansions were classified into three categories, namely, only single-module architectures formed (1,180, 69.49%), only multimodule architectures formed (281, 16.55%), and both types of architectures formed (237, 13.96%). Interestingly, most of the expansions formed only single-module architectures and mainly occurred through partial gene duplication (χ^2 test, χ^2 = 1,590, df = 13, $P < 2.2e\text{-}16$), and all of these single-module architectures were also novel architectures. However, expansions that formed only multimodule architectures were driven by gene duplication (χ^2 test, χ^2 = 6,939.2, df = 13, $P < 2.2e\text{-}16$), which increased gene copies without architectural changes. In conclusion, partial gene duplication might be an important driver of gene family expansions at the subgene level, with novel single-module protein formation.

Discussion

We have presented a phylogenetic workflow that can be used to comprehensively reconstruct protein-coding gene evolution and that captures gene duplication and loss at the gene level and module duplication, loss, fusion, fission, and emergence events at the subgene level. We have also developed a new phylogeny-based approach considering partial homology for gene family construction and demonstrated

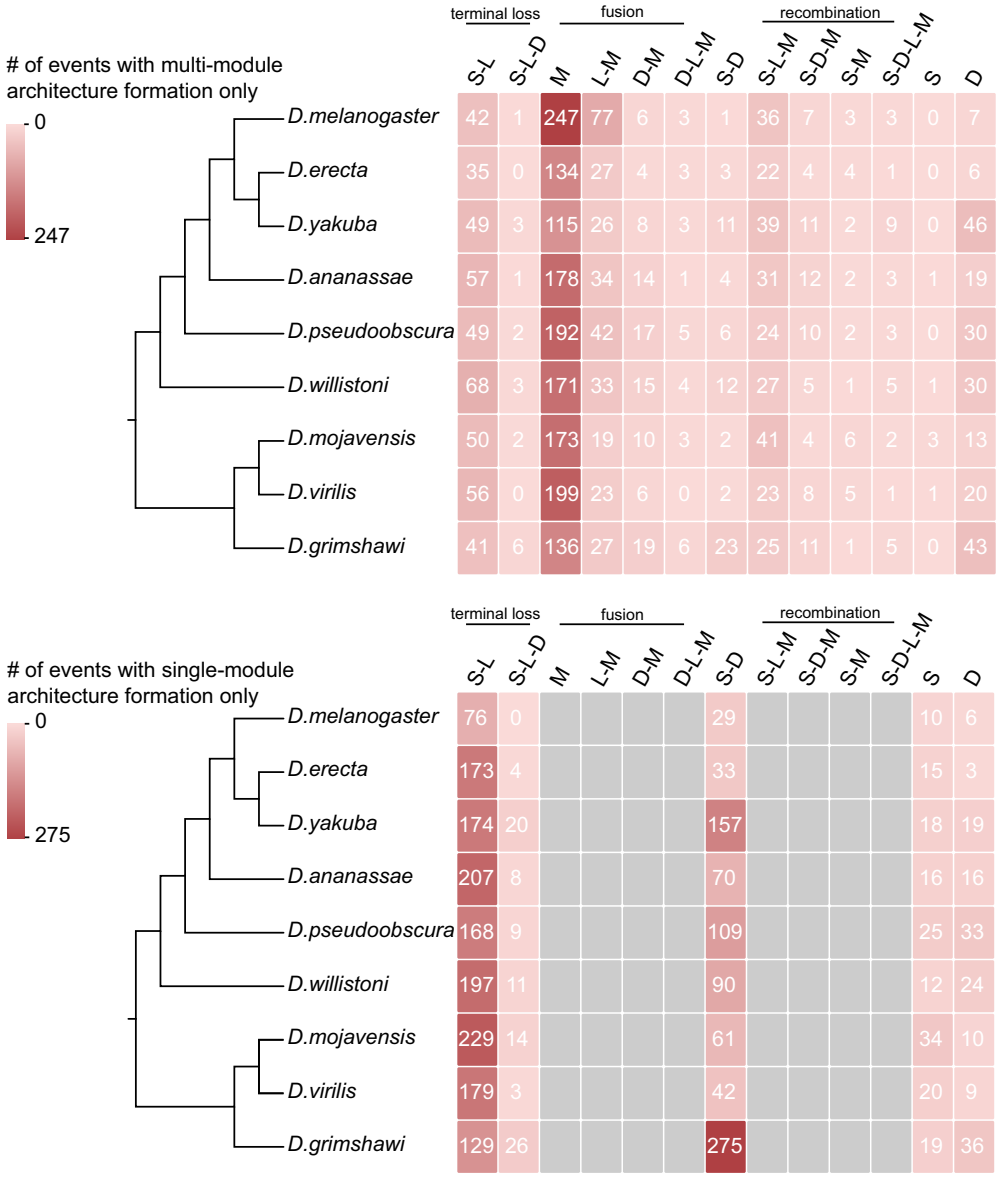

Fig. 6.—Distributions of module rearrangement events among the nine species. The upper portion shows the distribution when only multimodule architectures are formed, and the bottom portion shows the distribution when only single-module architectures are formed. In the bottom part, the columns are colored in gray when the corresponding module rearrangement event cannot form single-module architectures.

some of its advantages when compared with similarity-based approaches. First, our phylogeny-based approach is more realistic regarding the biological notion of homology. This strategy can comprehensively depict sequence evolutionary processes such as when and which evolutionary events might have occurred along the genome phylogeny that have led to partial homologs. For our constructed gene families, according to comparison with OrthoFinder and CompositeSearch, among genes that evolved only at the gene level, the constructed families were clearly almost identical. Interestingly, for genes that evolved at the module level, more reliable partial homologous relationships were revealed within each family constructed via our method. Second, our approach provides the detail investigation of gene family evolution. When gene family expansion and contraction analysis is performed on those families with partial homologs, the reconstructed evolutionary process may reflect genetic mechanisms including module rearrangement and not be limited to gene-level duplication and loss. Significantly, with both gene and module evolutionary events, we first provided a complete landscape of gene family evolution within a set of closely related genomes and found that the gene family expanded mainly through module rearrangements (fig. 4).

By adopting a more pluralistic account of homology, we methodologically defined homologous gene families that included full-length homologs as well as partial homologs by

analyzing the module and protein evolutionary processes through which proteins have frequently originated. The biological notion of homology was first raised by Owen (1868), who summarized homologies of the vertebrate skeleton and interpreted homologies in some archetypes as variants without an evolutionary view. After almost 140 years, the meaning of homology has been clarified with molecular sequence data (Reeck et al. 1987). Although the sentiments "homology is indivisible" (Fitch 2000) and "homology should mean 'possessing a common evolutionary origin'" (Reeck et al. 1987) are quoted in the vast majority of reports, Fitch (2000) and Hillis (1994) suggested that partial homology can be used with care to describe chimeric gene relationships. As our studies revealed and Haggerty et al. (2014) emphasized, homology should refer to proteins that possess at least one ancestor in common with other proteins, especially those that have evolved at the subgene level. With our observations on the architecture evolutionary process, we found that some proteins formed from different genetic parts, which can be expected to have different origins. Assuming that the number of ancestors of some proteins was >1, we constructed the more complex gene families and reflected gene family evolutionary histories more realistically. Therefore, the scope of evolutionary analyses may further expand. For example, because two genes from different gene families may be related through a multimodule gene that belongs to both gene families, these families related through intermediate sequences will have a family resemblance relationship (Haggerty et al. 2014). Some gene families may display family resemblances through intermediate sequences; thus, we can ask whether sequences are similar in function when they have a close family resemblance relationship and how the functional variation is related to family resemblance. These homologous gene families can be seen as complements to conventional similarity-based gene families. Obviously, such a classification of gene families will incur a greater computational cost, although the rationale of such an approach is appealing. Moreover, the reconstruction of architecture scenarios reveals the timing and form of architecture rearrangement but not the evolutionary process of sequences. Further analyses supporting the phylogenetic reconstruction of proteins or genes, such as the development of *N*-rooted fusion networks, are still needed.

A major issue in our study is the definition of subgene units. Identifying modules is computationally intensive. To delineate all possible evolutionary units of protein-coding genes across a set of closely related genomes, we could not use any existing domain databases because they not only are usually limited to some domains but also contain domains that are sparsely distributed along sequences. Thus, we need to identify modules based on sequence similarity across the set of selected genomes, currently using the approach proposed by Wu et al. (2012). This means that all possible modules must be reidentified for a new set of genomes, resulting in high computational costs. We can imagine that even if we add or remove one genome from the set of genomes in which the modules have been identified, we will need to recompute all modules again.

In addition, phylogenetic reconstruction should be improved. The reconstruction of ACC evolutionary scenarios is achieved by using the maximum parsimony method STAR-MP. Given extant architectures and inferred ancestral module counts, STAR-MP can generate a set of possible ancestral architectures at each internal node and determine the module events that would be necessary to transform the ancestral architectures into the extant architectures. It then applies a dynamic programing algorithm to find the minimum total cost along all internal branches and finally determines the most parsimonious ancestral architectures and module events at each node. Although the number of possible ancestral architectures can be intractably large, STAR-MP relies on heuristics to limit the architecture space and uses a maximum parsimony algorithm to determine the architectures. However, a better understanding of module rearrangements may help us better sample architecture space and even build a more biologically relevant model for accurate architecture reconstruction and module event inference. For example, determining how often module events occur may provide insight into event cost assignment. In this study, we set equal costs for the five types of module events and did not incorporate duplication-to-loss or merge-to-split ratios to avoid circular dependencies. In addition, although the parsimonious reconstructions of evolutionary histories are rapid and efficient, using other methods that propagate sequence information across all reconstructions, similarly to existing maximum likelihood and Bayesian methods, we may model both sequence and architecture evolution and better capture their evolutionary histories. Future studies may estimate module event rates independently and incorporate them in a probabilistic framework.

Our findings about the distribution of evolutionary events are comparable to those of previous studies. For whole-gene evolution, most gene duplication events (5,795 of 7,443, 77.86%) occurred along internal branches, whereas most gene loss events (18,379 of 20,859, 88.11%) occurred along terminal branches ($\chi^2 = 11{,}535$, df $= 1$, $P < 2.2\text{e-}16$). Gene losses occurred 2.8 times more often than gene duplications, which was in line with the findings of previous studies (Rasmussen and Kellis 2011; Koskiniemi et al. 2012; Puigbo et al. 2014; Nelsonsathi et al. 2015). In particular, the gene loss-to-gene duplication ratio of terminal branches was 11.15. The large number of gene losses relative to gene duplications might be due to gene dispensability. As many studies revealed, only a few hundred genes are essential, and nearly 90% of genes in bacteria (Baba et al. 2006; De Berardinis et al. 2008), 80% in yeast (Giaever et al. 2002; Kim et al. 2010), and 65–85% in *Caenorhabditis elegans* (Kamath et al. 2003; Sonnichsen et al. 2005) and *D. melanogaster* (Kamath

et al. 2003; Sonnichsen et al. 2005) are dispensable. It is plausible to assume that redundant genes—paralogs arising via duplication—functionally overlap with (Mcclintock et al. 2001; Gitelman 2007; Canestro et al. 2009) or supplement alternative pathways (Danchin et al. 2006), accounting for ancient gene duplications followed by many gene losses. For module evolution, we observed that module losses (9,181) outnumbered module duplications (3,494) on terminal branches by a factor of 2.63, which was similar to the pattern for whole-gene evolution and in accordance with previously reported results at the domain level (Zmasek and Godzik 2011). Emergence events all occurred along internal branches. Note that we focused on the well-known *Drosophila* clade and did not use outgroup genomes to search for de novo-created genes. Thus, some inferred emerged modules might have existed prior to the formation of the *Drosophila* clade. The module merge-to-split ratio along terminal branches was 0.58, which seemed to be inconsistent with previous findings that fusion is more frequent than fission (Kummerfeld and Teichmann 2005; Fong et al. 2007; Kersting et al. 2012). However, as Wu et al. emphasized (Wu et al. 2012), their method measured individual events to describe the detailed process of architecture formation. For example, partial gene duplication (architecture AB to architecture AB and A) and partial gene loss (architecture AB to architecture A) all require a split event prior to module duplication/loss. If we consider "simple" merges and splits that are unaccompanied by other module evolutionary events, as shown in figure 5, the merge-to-split ratio becomes 6.31, which is comparable to that obtained in previous studies.

Investigating the evolution of gene families across species may provide insight into the evolutionary forces that have shaped gene family diversity and adaptation. Most cases of gene family expansion and contraction reflect the pervasiveness of gene duplication and gene loss. Complementarily, we show that module rearrangement is also prevalent in gene family evolution. For a long time, gene duplication was thought to be the major driver of increased gene family size and to contribute to adaptive evolution (Demuth and Hahn 2009; Kondrashov 2012). Additionally, the duplicated genes that accumulate mutations may further diverge and increase the diversity of gene family members (Gao et al. 2014). The formation of proteins with new functions from preexisting ones is thought to be easier than the de novo formation of genes (Kondrashov 2012). In our study, we found that gene family expansions mainly generated single-module genes instead of multimodule genes. Although family expansions frequently result from gene duplication when forming multimodule genes, partial gene duplication is more frequent when forming single-module genes. One explanation is that the specific functional part of proteins might be expanded because of selective pressure. The monkey-king (*mkg*) gene family discovered in *D. melanogaster* shows that genes can originate from different domains of an ancestral protein through a fission process and that the underlying mechanism is duplication followed by complementary partial degeneration (Wang et al. 2004), leading to the specification of protein functions in gene duplicates. That is, the mechanism of partial gene duplication is gene duplication with subsequent partial degeneration. It is conceivable that gene duplicates might be redundant in the early stage and then subsequent degenerations might lead to functional specification during the process of species adaptation. Overall, diversity among gene family members can result from not only divergent gene duplicates but also new single-module proteins from partial gene duplication, which implies the formation of specific functional proteins. For gene family contractions, whether gene loss is adaptive or neutral is sometimes controversial. The less-is-more hypothesis proposes that adaptive loss-of-function mutations are due to changes in environmental conditions (Olson 1999; Olson and Varki 2003; Howes et al. 2011; Hottes et al. 2013), and regression evolution offers many examples of the loss of useless genes and characteristics with neutral effects on fitness (Moreau and Dabrowski 1998; Protas et al. 2006). The neutralism–selectionism debate questions whether neutral variants are related to the emergence of evolutionary innovations (Wagner 2008). We demonstrated that for all 6,472 branches with gene family contractions, although the largest percentage (80.18%, 5,189) was due to pure gene loss, 19.22% (4,053) resulted from gene loss accompanied by novel gene emergence. The novel genes formed through module rearrangement may have potential functional innovations and be adaptive and selection driven, contributing to adaptation of the species, and the ancestral genes might be neutral or slightly deleterious and finally lost by genetic drift. These gene families might contract in size but increase in diversity with different members, which implies functional transformation through the evolution of gene families.

In summary, reconsidering exactly how genes evolve, particularly at the subgene level, we adopt a revised homology model to methodologically resolve the problem of homology detection. In the case study, we found pervasive module rearrangement during gene family evolution and characterized more details as complementary to our previous understanding of gene family expansion and contraction. Further analyses incorporating partial homology into gene family construction may provide new insights into the complex relationships between gene function, species phylogenetic relationships, and evolutionary processes.

Supplementary Data

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We cordially thank two anonymous reviewers for their invaluable comments. We also thank all members of the Lin lab for their valuable suggestions. This work was supported by the State Key Basic Research and Development Plan (2017YFA0605104) and a key project of the State Key Laboratory of Earth Surface Processes and Resource Ecology.

Literature Cited

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 38(Suppl 2):W7–W13.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389–3402.

Andreeva A, et al. 2014. SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res. 42(D1):D310–D314.

Ane C, et al. 2006. Bayesian estimation of concordance among gene trees. Mol Biol Evol. 24(2):412–426.

Armisén D, et al. 2018. The genome of the water strider *Gerris buenoi* reveals expansions of gene repertoires associated with adaptations to life on the water. BMC Genomics. 19(1):832.

Baba T, et al. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2:2006.0008.

Benton R. 2015. Multigene family evolution: perspectives from insect chemoreceptors. Trends Ecol Evol. 30:590–600.

Bjorklund AK, et al. 2005. Domain rearrangements in protein evolution. J Mol Biol. 353:911–923.

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. J Stat Mech. 10:P10008.

Bornberg-Bauer E, Alba MM. 2013. Dynamics and adaptive benefits of modular protein evolution. Curr Opin Struct Biol. 23(3):459–466.

Bornberg-Bauer E, et al. 2005. The evolution of domain arrangements in proteins and interaction networks. Cell Mol Life Sci. 62(4):435–445.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12(1):59–60.

Buljan M, Frankish A, Bateman A. 2010. Quantifying the mechanisms of domain gain in animal proteins. Genome Biol. 11(7):R74.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1):421.

Canestro C, et al. 2009. Consequences of lineage-specific gene loss on functional evolution of surviving paralogs: ALDH1A and retinoic acid signaling in vertebrate genomes. PLoS Genet. 5:e1000496.

Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol. 7(3–4):429–447.

Corel E, Lopez P, Meheust R, Bapteste E. 2016. Network-thinking: graphs to analyze microbial complexity and evolution. Trends Microbiol. 24(3):224–237.

Cortesi F, et al. 2015. Ancestral duplications and highly dynamic opsin gene evolution in percomorph fishes. Proc Natl Acad Sci U S A. 112(5):1493–1498.

Danchin EGJ, Gouret P, Pontarotti P. 2006. Eleven ancestral gene families lost in mammals and vertebrates while otherwise universally conserved in animals. BMC Evol Biol. 6(1):5–5.

De Berardinis V, et al. 2008. A complete collection of single-gene deletion mutants of *Acinetobacter baylyi* ADP1. Mol Syst Biol. 4(1):174.

De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22(10):1269–1271.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol (Amst). 24(6):332–340.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6(5):361–375.

Demuth JP, Hahn MW. 2009. The life and death of gene families. BioEssays 31(1):29–39.

Ding Y, Zhou Q, Wang W. 2012. Origins of new genes and evolution of their novel functions. Annu Rev Ecol Evol Syst. 43(1):345–363.

Doyon JP, Ranwez V, Daubin V, Berry V. 2011. Models, algorithms and programs for phylogeny reconciliation. Briefings Bioinf. 12(5):392–400.

Drosophila 12 Genomes Consortium, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450(7167):203–218.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14(9):755–763.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Ekman D, Bjorklund AK, Elofsson A. 2007. Quantification of the elevated rate of domain rearrangements in metazoa. J Mol Biol. 372(5):1337–1348.

El-Gebali S, et al. 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47(D1):D427–D432.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20(1):238.

Enright AJ, Kunin V, Ouzounis CA. 2003. Protein families and TRIBES in genome sequence space. Nucleic Acids Res. 31(15):4632–4638.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30(7):1575–1584.

Fitch WM. 2000. Homology: a personal view on some of the problems. Trends Genet. 16(5):227–231.

Fong JH, Geer LY, Panchenko AR, Bryant SH. 2007. Modeling the evolution of protein domain architectures using maximum parsimony. J Mol Biol. 366(1):307–315.

Francino MP. 2005. An adaptive radiation model for the origin of new gene functions. Nat Genet. 37(6):573–577.

Gao F, Song W, Katz LA. 2014. Genome structure drives patterns of gene family evolution in ciliates, a case study using *Chilodonella uncinata* (Protista, Ciliophora, Phyllopharyngea). Evolution 68(8):2287–2295.

Giaever G, et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418(6896):387–391.

Gitelman I. 2007. Evolution of the vertebrate twist family and synfunctionalization: a mechanism for differential gene loss through merging of expression domains. Mol Biol Evol. 24(9):1912–1925.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59(3):307–321.

Haggerty LS, et al. 2014. A pluralistic account of homology: adapting the models to the data. Mol Biol Evol. 31(3):501–516.

Hahn MW, Han MV, Han S. 2007. Gene family evolution across 12 *Drosophila* genomes. PLoS Genet. 3(11):e197.

Hahn MW, et al. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. 15(8):1153–1160.

Heger A, Holm L. 2003. Exhaustive enumeration of protein domain families. J Mol Biol. 328(3):749–767.

Hillis DM. 1994. Homology in molecular biology. In: Hall B, editor. Homology, the hierarchical basis of comparative biology. San Diego (CA): Academic Press. p. 483.

Hottes AK, et al. 2013. Bacterial adaptation through loss of function. PLoS Genet. 9(7):e1003617.

Howes RE, et al. 2011. The global distribution of the Duffy blood group. Nat Commun. 2(1):266.
Huang X, Miller W. 1991. A time-efficient, linear-space local similarity algorithm. Adv Appl Math. 12(3):337–357.
Huson DH, Scornavacca C. 2011. A survey of combinatorial methods for phylogenetic networks. Genome Biol Evol. 3:23–35.
Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 11(2):97–108.
Jachiet P-A, et al. 2013. MosaicFinder: identification of fused gene families in sequence similarity networks. Bioinformatics 29(7):837–844.
Kamath RS, et al. 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature 421(6920):231–237.
Kazazian HH. 2004. Mobile elements: drivers of genome evolution. Science 303(5664):1626–1632.
Kersting AR, Bornberg-Bauer E, Moore AD, Grath S. 2012. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. Genome Biol Evol. 4(3):316–329.
Kim D-U, et al. 2010. Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*. Nat Biotechnol. 28(6):617–623.
Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc R Soc B 279(1749):5048–5057.
Koskiniemi S, Sun S, Berg OG, Andersson DI. 2012. Selection-driven gene loss in bacteria. PLoS Genet. 8(6):e1002787.
Kummerfeld SK, Teichmann SA. 2005. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet. 21(1):25–30.
Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9):2178–2189.
Liu L, Pearl DK. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol. 56(3):504–514.
Maddison WP. 1997. Gene trees in species trees. Syst Biol. 46(3):523–536.
Markow TA, Ogrady PM. 2007. *Drosophila* biology in the genomic age. Genetics 177(3):1269–1276.
Mcclintock JM, Carlson R, Mann DM, Prince VE. 2001. Consequences of Hox gene duplication in the vertebrates: an investigation of the zebrafish Hox paralogue group 1 genes. Development 128(13):2471–2484.
McInerney JO, Pisani D, Bapteste E, O'Connell MJ. 2011. The public goods hypothesis for the evolution of life on Earth. Biol Direct 6(1):41.
Meheust R, et al. 2016. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. Proc Natl Acad Sci U S A. 113:3579–3584.
Moore AD, Bornberg-Bauer E. 2012. The dynamics and evolutionary potential of domain loss and emergence. Mol Biol Evol. 29(2):787–796.
Moore AD, et al. 2008. Arrangements in the modular evolution of proteins. Trends Biochem Sci. 33(9):444–451.
Moore AD, et al. 2013. Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree. Biochim Biophys Acta 1834(5):898–907.
Moreau RF, Dabrowski K. 1998. Body pool and synthesis of ascorbic acid in adult sea lamprey (*Petromyzon marinus*): an agnathan fish with gulonolactone oxidase activity. Proc Natl Acad Sci U S A. 95(17):10279–10282.
Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 247(4):536–540.
Nakhleh L. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol Evol. 28:719–728.
Nelsonsathi S, et al. 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature 517:77–80.
Ness RW, Graham SW, Barrett SC. 2011. Reconciling gene and genome duplication events: using multiple nuclear gene families to infer the phylogeny of the aquatic plant family Pontederiaceae. Mol Biol Evol. 28(11):3009–3018.
Nurminsky DI, De Aguiar D, Bustamante C, Hartl DL. 2001. Chromosomal effects of rapid gene evolution in *Drosophila melanogaster*. Science 291(5501):128–130.
Nurminsky DI, Nurminskaya M, De Aguiar D, Hartl DL. 1998. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature 396(6711):572–575.
Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet. 64(1):18–23.
Olson MV, Varki A. 2003. Sequencing the chimpanzee genome: insights into human evolution and disease. Nat Rev Genet. 4(1):20–28.
Omland KE, Cook LG, Crisp MD. 2008. Tree thinking for all biology: the problem with reading phylogenies as ladders of progress. BioEssays 30(9):854–867.
Östlund G, et al. 2009. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38:D196–D203.
Owen R. 1868. On the archetype and homologies of the vertebrate skeleton. London: Richard and John E. Taylor.
Pathmanathan JS, Lopez P, Lapointe FJ, Bapteste E. 2018. CompositeSearch: a generalized network approach for composite gene families detection. Mol Biol Evol. 35(1):252–255.
Posada D. 2008. jModelTest: phylogenetic model averaging. Mol Biol Evol. 25(7):1253–1256.
Protas ME, et al. 2006. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nat Genet. 38(1):107–111.
Puigbo P, et al. 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biol. 12:66–66.
Rasmussen MD, Kellis M. 2011. A Bayesian approach for fast and accurate gene tree reconstruction. Mol Biol Evol. 28(1):273–290.
Reeck GR, et al. 1987. "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. Cell 50(5):667.
Rubin GM, et al. 2000. Comparative genomics of the eukaryotes. Science 287(5461):2204–2215.
Servant F, et al. 2002. ProDom: automated clustering of homologous domains. Briefings Bioinf. 3(3):246–251.
Sibbald SJ, Hopkins JF, Filloramo GV, Archibald JM. 2019. Ubiquitin fusion proteins in algae: implications for cell biology and the spread of photosynthesis. BMC Genomics. 20(1):38.
Sonnhammer ELL, Eddy SR, Durbin R. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28(3):405–420.
Sonnichsen B, et al. 2005. Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. Nature 434:462–469.
Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 35(11):1026–1028.
Supek F, Bosnjak M, Skunca N, Smuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 6(7):e21800.
Szöllősi GJ, Tannier E, Daubin V, Boussau B. 2015. The inference of gene trees with species trees. Syst Biol. 64(1):e42–e62.
Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol Biol Evol. 21(1):36–44.
Wagner A. 2008. Neutralism and selectionism: a network-based reconciliation. Nat Rev Genet. 9(12):965–974.
Wang M, Caetano-Anollés G. 2009. The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. Structure 17(1):66–78.
Wang W, Yu H, Long M. 2004. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. Nat Genet. 36(5):523–527.

Waterhouse RM, et al. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res. 41(Database issue):D358–D365.
Wilson D, et al. 2007. The SUPERFAMILY database in 2007: families and functions. Nucleic Acids Res. 35(Database):D308–D313.
Wu YC, Rasmussen MD, Bansal MS, Kellis M. 2013. TreeFix: statistically informed gene tree error correction using species trees. Syst Biol. 62(1):110–120.
Wu YC, Rasmussen MD, Kellis M. 2012. Evolution at the subgene level: domain rearrangements in the *Drosophila* phylogeny. Mol Biol Evol. 29(2):689–705.
Yeh SD, et al. 2012. Functional evidence that a recently evolved *Drosophila* sperm-specific gene boosts sperm competition. Proc Natl Acad Sci U S A. 109(6):2043–2048.
Zhang H, et al. 2012. EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. Nucleic Acids Res . 40(Web Server issue):W569–W572.
Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18:292–298.
Zmasek CM, Eddy SR. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 17(9):821–828.
Zmasek CM, Godzik A. 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol. 12(1):R4.

Associate editor: Josefa Gonzalez