The Dynamics and Evolutionary Potential of Domain Loss and Emergence

Andrew D. Moore [1] and Erich Bornberg-Bauer[1,*]

[1]Evolutionary Bioinformatics Group, Institute for Evolution and Biodiversity, University of Muenster, Germany

***Corresponding author:** E-mail: ebb@uni-muenster.de.

Associate editor: John Parsch

Research article

Abstract

The wealth of available genomic data presents an unrivaled opportunity to study the molecular basis of evolution. Studies on gene family expansions and site-dependent analyses have already helped establish important insights into how proteins facilitate adaptation. However, efforts to conduct full-scale cross-genomic comparisons between species are challenged by both growing amounts of data and the inherent difficulty in accurately inferring homology between deeply rooted species. Proteins, in comparison, evolve by means of domain rearrangements, a process more amenable to study given the strength of profile-based homology inference and the lower rates with which rearrangements occur. However, adapting to a constantly changing environment can require molecular modulations beyond reach of rearrangement alone. Here, we explore rates and functional implications of novel domain emergence in contrast to domain gain and loss in 20 arthropod species of the pancrustacean clade. Emerging domains are more likely disordered in structure and spread more rapidly within their genomes than established domains. Furthermore, although domain turnover occurs at lower rates than gene family turnover, we find strong evidence that the emergence of novel domains is foremost associated with environmental adaptation such as abiotic stress response. The results presented here illustrate the simplicity with which domain-based analyses can unravel key players of nature's adaptational machinery, complementing the classical site-based analyses of adaptation.

Key words: modular protein evolution, molecular innovation, protein domains, genome evolution, Drosophila.

Introduction

Since eukaryotic genomes are sequenced at an ever-increasing pace, comparative genomics has become an indispensable approach in many areas of molecular biosciences. One major goal is to understand, from a molecular perspective, how adaptation, development, and speciation have come about. However, automated functional interpretation of evolutionary traits in molecular terms is still a daunting task: accurate genome-scale de novo predictions of gene and protein structure as well as function are far from feasible. Moreover, many predicted protein-coding genes are "orphans" that lack detectable homology to known proteins, yet may likely be key players in the process of adaptation (Khalturin et al. 2009; Johnson and Tsutsui 2011).

However, by considering the modularity of protein evolution, valuable insights into the evolutionary forces shaping the functional make up of genomes have been obtained (Chothia et al. 2003; Pasek et al. 2005; Moore et al. 2008; Buljan et al. 2010). A key insight to start with is the observation that the overall number of novel, that is, of previously unreported, domains seems to converge, whereas the number of known modular arrangements of these domains is still rapidly expanding (Levitt 2009). Domains are the functional and structural constituents of proteins. They are evolutionary well conserved across taxa (Elofsson and Sonnhammer 1999; Finn et al. 2010) but frequently rearranged between and within proteins and genomes (Moore et al. 2008). These rearrangements can be observed independently of whether domains are defined from a structural perspective (see, e.g., Apic et al. 2001; Wang and Caetano-Anollés 2009) or an "implicit" evolutionary perspective, that is, by comparing sequence fragments that are conserved across many taxa (Björklund et al. 2005; Ekman et al. 2005). The events underlying domain rearrangements are duplication, fusion, and fission (Kummerfeld and Teichmann 2005; Pasek et al. 2005) as well as terminal domain loss (Björklund et al. 2005; Weiner et al. 2006; Buljan et al. 2010). These events are likely fueled by a series of underlying genetic events such as nonallelic homologous recombination, nonhomologous end joining, transposition events, or combinations thereof. Eukaryotic proteomes contain a larger proportion of multidomain proteins than bacteria and archeae (Apic et al. 2001; Ekman et al. 2005), and some studies concentrating on smaller clades found that rearrangement rates differ between kingdoms (Ekman et al. 2007).

The ability to reuse is a hallmark of modular design, and the rearrangement of existing domains is more frequent than the formation of novel domains (Apic et al. 2001). Ergo, it seems likely that functional novelty, such as required in the wake of environmental shifts, can be generated by modular rearrangements as opposed to the formation of novel domains. However, there is evidence that rearrangements of intact domains do not strongly alter arrangement functionality (Tjoelker et al. 2000; Koide 2009), whereas effects such as modified binding affinity or substrate specificity may result (Yu and Lutz 2011). Consequently, certain molecular innovation, such as required for

the adaptation to new environments, may be out of reach by rearrangement alone and may be instead facilitated by the emergence of novel domains. Indeed, change can be observed not only in the arrangements present in a genome but also in domain content (Itoh et al. 2007). For example, more than half of the domains present in *Homo sapiens* originate before the metazoan era; only ~2% originate in *H. sapiens* (Pal and Guda 2006). Although these turnover rates of domains across proteomes seem low, they nonetheless allow for comparative analyses from which phylogenies can be reconstructed (Björklund et al. 2005; Yang et al. 2005; Wang and Caetano-Anollés 2006) and can be qualitatively related to functional classes (Pal and Guda 2006; Itoh et al. 2007; Zmasek and Godzik 2011).

These findings suggest that, albeit rare, novel domains may emerge as a result of functional challenges not met by modular rearrangements; such novel domains may confer a high adaptive potential. Accordingly, we here ask how frequently domain families are gained and lost and, in particular, how frequently novel domain families emerge and whether such new families confer new functionalities. We address these questions in the pancrustacean clade as it is densely covered with well-annotated genomes representing species splits ranging from 1.2 to ~450 My. Furthermore, the pancrustacean clade incorporates species with a wide range of adaptational diversity including both cosmopolitan generalists and geographically restricted specialists. Given that evolutionary analyses are not confounded by whole-genome duplications and that the overall topology of the species tree is well established (Meusemann et al. 2010), the pancrustacean clade provides an excellent data set to study the dynamics of domain turnover across proteomes.

The approach taken here may aid the functional analysis of future genome and proteome projects as it exploits the high precision of profile-based domain detection and is complementary to methods using site-based sequence analysis and turnover of gene families.

Methods

Proteomes and Annotation

Due to the high density of available genomes within the clade, we chose to analyze domain emergence within pancrustacea. We used the predicted peptides of the 12 Drosophila species (Drosophila Genome Consortium , 2007): *Drosophila simulans* (r1.3), *D. sechellia* (r1.3), *D. melanogaster* (r5.11), *D. yakuba* (r1.3), *D. erecta* (r1.3), *D. ananassae* (r1.3), *D. pseudoobscura* (r2.3), *D. persimilis* (r1.3), *D. willistoni* (r1.3), *D. mojavensis* (r1.3), *D. virilis* (r1.2), and *D. grimshawi* (r1.3). The proteomes were obtained from FlyBase. We complimented the Drosophila data set with the proteomes of the three mosquitoes *Anopheles gambiae* (P3.49), *Culex pipiens* (1.2), and *Aedes aegypti* (L1.49) (obtained from VectorBase); the moth *Bombyx mori* (1.0, obtained from the Silkworm Genome Database); the beetle *Tribolium castaneum* (51,906, obtained from BeetleBase); the two hymenoptera *Nasonia vitripennis* (1.2, obtained from the Baylor College of Medicine/Human Genome Sequencing Center (BCM/HGSC)) and *Apis mellifera* (4.0, obtained from BCS/HGSC); and the coleoptera *Daphnia pulex* (060905, obtained from the Joint Genome Institute). As outgroups we used the proteomes of *H. sapiens* (NCBI36.51, obtained from GenBank) and *Caenorhabditis elegans* (WS 206, obtained from WormBase). We chose these outgroups in order to identify old domains that are common to a wide range of taxa and hence cannot be specific to the pancrustacean clade; the use of outgroups that are only distantly related to the species considered reduces the number of pancrustacea-specific domain candidates. For the complete tree including outgroups, see supplementary figure 3, Supplementary Material online.

Proteomes were scanned using the pfamscan utility and HMMER 3.0 against Pfam-A and B domain models obtained from Pfam (v.24) (Finn et al. 2010). For Pfam-A, we employed the curated, model-defined gathering threshold as bit score cutoff. For Pfam-B, we chose an *E* value cutoff of 10^{-3}, similar to previous studies (Ekman et al. 2007). If multiple transcripts were present, we removed all but the longest splice variant. The domain residue coverage is roughly 50% for each proteome; roughly 76% of all proteins had at least one domain. Due to the domain centric view employed in this study, we discarded proteins that lack domain annotation.

Ancestral Domain Contents: Domain Gain, Loss, and Emergence

We used Dollo parsimony (Farris 1977) for prediction of ancestral domain contents. The assumption underlying the use of Dollo parsimony is that domains are gained only once and that number of losses required to explain domain contents at nodes is minimized. Under Dollo, domain gain events will tend to occur early and will be offset by a large number of domain loss events. However, we consider Dollo parsimony as used here sufficiently robust. First, in this study we do not consider copy number variation; we consider only the binary state, presence or absence, of a given domain in any given node. Hence, a domain can only be lost along a branch if 1) it has been gained at an ancestral node to the branch considered and 2) not a single copy is present in the descendant node (or its subtree). Second, in most cases, domains represent the functional unit within a given protein. As horizontal transfer of genetic material within eukaryotes can at least be considered rare, gain events of such functional modules would imply de novo formation. Finally, the danger of overestimating loss events is larger, the more deeply the tree is rooted. Here, we use a shallow and densely populated tree dating back only 430 My. Other studies have successfully employed Dollo to considerably larger data sets (Zmasek and Godzik 2011). Hence, we feel that the assumptions underlying Dollo parsimony are reasonable within the framework of this study.

After ancestral reconstruction, we measured domain gain and loss events along each branch in the tree. Two correction steps were undertaken to distinguish between domain "gain" events (where domains that can be found

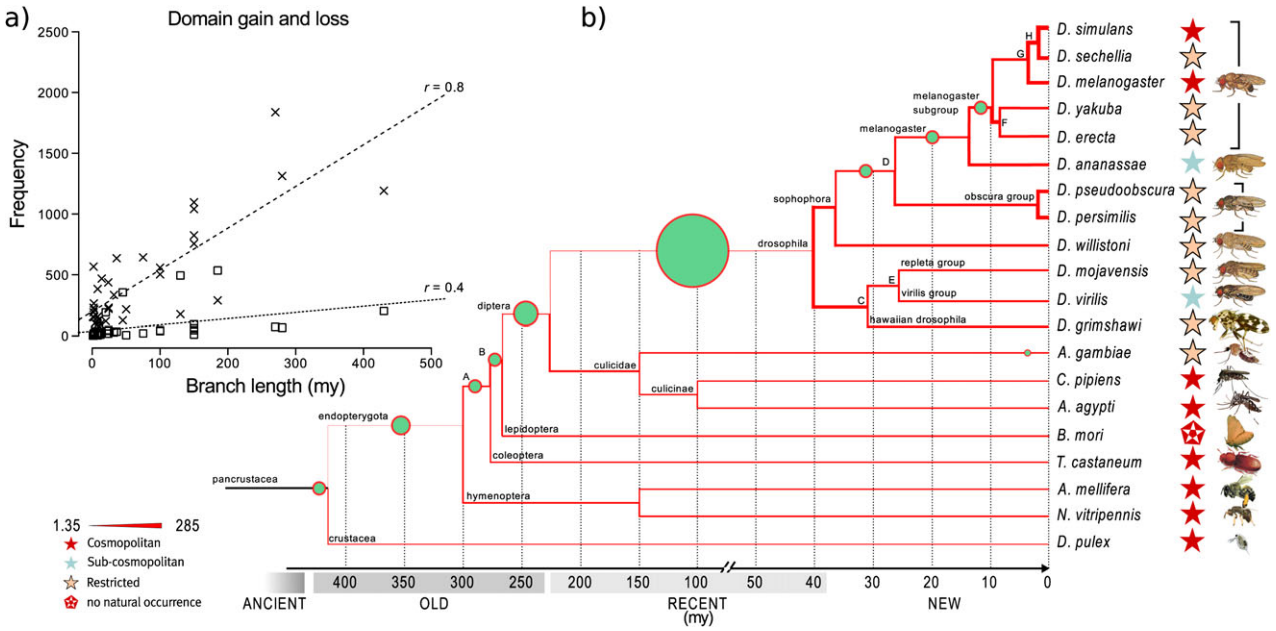

FIG. 1. Domain loss, gain, and emergence across 20 species of pancrustacea. (*a*) Domain gain (squares) and loss (crosses) against branch length. Ancestral domain content was reconstructed using a parsimony-based approach. Events were inferred along each branch of the tree. Domain loss correlates well with branch length (Pearson $r = 0.808$, $P \ll 0.001$). (*b*) Domain loss and emergence along branches. Nonclassified common ancestors are labeled A–H. Line strength corresponds to rate of domain loss per My along the respective branch. Domain emergence is indicated by green circles scaled to the number of emergence events along the respective branch (see also Table 1). Tree and approximate divergence times are based on Honeybee Genome Sequencing Consortium (2006) and Hedges et al. (2006).

outside of pancrustacea are gained along a branch within the pancrustacean clade) and domain "emergence" events (where domains are gained that are only found "within" the pancrustacean clade). First, we only considered domains that are gained within pancrustacea and discarded those domains that were gained at ancestral nodes of pancrustacea and either of the outgroups. Of the initial 11,735 domain gain events in the whole tree including outgroups (4,987 Pfam-A; 6,748 Pfam-B), a total of 8,492 (4,558 Pfam-A; 3,934 Pfam-B) domains are either ancient, that is, are shared by at least one pancrustacean species and one outgroup species, or are gained along a branch to an outgroup. In both cases, the domains in question cannot be specific to the pancrustacean clade. Next, we constructed a database containing the hidden Markov models of all the remaining 3,243 domains that are gained within pancrustacea and used HMMER 3.0 to scan these models against a sequence database consisting of NCBIs NR and Integr8 (Kersey et al. 2005); gained domains with hits to sequences of species outside of the pancrustacean clade were removed, facilitating a set of 30 (29 Pfam-A and 1 Pfam-B) domains that emerge within pancrustacea.

Emergence Bins and Disorder in Emerging Domains

Emerging domains were grouped into three bins according to their age. The OLD bin contains domains that emerge at the root of the tree 430 Ma up until the diptera node, 225 Ma and spans ~200 My. The RECENT bin spans 185 My from diptera to Drosophila, the last common ancestor of all Drosophila species. The NEW bin incorporates all domains that are younger than 40 My (see fig. 1). We also constructed an ANCIENT bin, which contains domains that likely emerged before our root node. We did not ensure that domains from the ANCIENT bin actually emerge at ancestral nodes; we required a set of domains that are gained before pancrustacea. Such domains have a hit in at least one of the outgroups and one pancrustacean species and hence must be considerably older than 430 My. For disorder prediction, we chose randomly 100 domains from the ANCIENT bin while maintaining the fraction of Pfam-A and Pfam-B domains within the selection. Finally, we created a RANDOM bin containing 100 randomly selected domains, irrespective of the time point of their emergence.

Domain Arrangements with Emerging Domains

A domain arrangement is defined as the linear combination of domains in a protein. To avoid overestimating the number of unique arrangements an emerging domain can be found in, we collapsed repeats to a single instance as copy number variation in repeats can occur between even closely related species (Ekman et al. 2007). Our analysis pipeline utilizes both custom implementations and existing software. The pipeline consists of software for domain annotation, RUBY libraries for managing domain annotation and ancestral domain contents reconstruction, and software for assessing and visualizing overrepresentation of gene ontology (GO) terms. A description of the pipeline with links can be found online at http://iebservices.uni-muenster.de/radmoore/emergence.

Functional Analysis of Emerging Domains

To analyze the functional impact of domain gains, we conducted an overrepresentation analysis of GO (Reference

Genome Group of the Gene Ontology Consortium 2009) terms. As only 6 of the 30 emerging domains are directly annotated with GO terms, we employed a, to our knowledge, novel indirect GO analysis. First, we annotated all 20 proteomes using Blast2GO (Conesa and Götz 2008) with default settings. We then extracted all proteins that contain a gained domain (1,291). Using the entire functional annotation of pancrustacean proteins as universe, we sought to find functional terms that are associated with domain emergence using R and Bioconductors TopGO (Alexa et al. 2006) package. We used the weighted algorithm of TopGO, which eliminates local similarities and dependencies between GO terms by utilizing the topology of the GO graph during the analysis. After correction for multiple testing using Bonferroni, we found 43 significantly overrepresented terms in the ontology `biological_process` and 6 in the ontology `molecular_function`. Inspired by sequence logos, which are frequently used to represent the frequency of a nucleotide or amino acid in an alignment column, we visualized the significant terms from the `biological_process` ontology using a tag cloud–like representation, which we call a TermLogo (see fig. 3). Tag clouds typically represent the importance of a given word or phrase within a text document by scaling them according to their frequency. We used a tag cloud representation of the GO terms and transformed the *P* value obtained from the TopGO analysis using a scaling factor τ defined as

$$\tau = |\log_{10}(p)|$$

such that the size of the font within the cloud does not represent term frequency but the significance of the respective term in the overrepresentation analysis. Hence, in our TermLogo, the larger the font, the smaller the associated *P* value.

Results and Discussion

Rates of Domain Loss, Gain, and Emergence

We annotated the proteomes of 20 pancrustacean species and two outgroups using Pfam-A and Pfam-B (Finn et al. 2010) and reconstructed the ancestral domain content at each node of the species tree using a parsimony-based approach (see Methods). We then measured, along each branch of the tree, the number of gained, lost, and novel domains. The results are summarized in figure 1 and table 1.

A domain is considered to be lost at a node if it does not occur in any of its child nodes and gained if absent at a nodes' parent (which follows a well-established approach; see also Fong et al. 2007; Rogers et al. 2010; Zmasek and Godzik 2011). A domain that is both gained within and taxonomically restricted to the pancrustacean clade is considered a novel "emerging" domain (see Methods). Domain loss rates correlate well with branch length (see fig. 1*a* and supplementary table 1, Supplementary Material online) but are lineage dependent. In total, there are 5,375 loss events within the Drosophila clade (1,313 Pfam-A and 4,062 Pfam-B), with an average loss rate of 3.41 $\pm$ 0.31 domains per My along Drosophila lineages. In comparison, the non-Drosophila lineages within the pancrustacean clade see a total of 10,818 loss events (3,180 Pfam-A and 7,638 Pfam-B) and exhibit an average loss rate of 4.43 $\pm$ 0.84 domains per My. The highest loss rates within pancrustacea can be found along short branches within the Drosophila clade, in particular within the subtrees of the *melanogaster* subgroup and *obscura* group. This is in line with the previous studies focusing on gene family turnover rates (Hahn et al. 2007). For many of the lost domains, multiple instances can be found in sister taxa. The TB domain (PF00683), for example, is found in fibrillins and Transforming Growth Factor-binding proteins and is localized in the extracellular matrix. The TB domain is likely quite old; instances can be found in the outgroup *H. sapiens* and in the pancrustacea *D. pulex*, *B. mori*, *A. mellifera*, and *T. castaneum*. TB seems to have been lost along the branches to *N. vitripennis* and the last common ancestor of lepidoptera and diptera; it cannot be found within the Dropsophila clade. By loosening the *E* value threshold to 0.1, weak traces of TB can be found in *N. vitripennis* and some Drosophila species suggesting either ectopic decay at the sequence level or functional divergence beyond detection by the current model.

The average domain gain rate along all pancrustacean lineages is 1.9 $\pm$ 0.84 events per My. In comparison, the Drosophila lineages exhibit an average domain gain rate of 4 $\pm$ 0.03 per My. It should be noted that inferred gain and loss rates are partially dependent on the chosen *E* value cutoff used during initial domain annotation. A domain may diverge beyond detection, either as the result of functional divergence or as the result of mutations that render it nonfunctional. If the *E* value cutoff used for detecting domains is lowered, domains previously absent may become visible to our analysis. Supplementary figure 2 (Supplementary Material online) illustrates the effect of different thresholds on gain and loss rates. It demonstrates that domain loss is particularly sensitive to variation in *E* value threshold; loss rates decrease with more stringent cutoffs, likely as the total number of detected domains decreases. Domain gain is less affected as gain is restricted under Dollo's law. To ensure robust rate estimation, we chose the model-defined gathering threshold for Pfam-A to minimize the number of falsely annotated domains. For Pfam-B, we chose a cutoff of 10^{-3} that offers a fair balance between sensitivity and selectivity (Ekman et al. 2007).

Among the ~3,000 domains gained across the whole pancrustacean tree, a tiny fraction of only 30 domains are evolutionarily novel, that is, they are not detectable anywhere outside of pancrustacea (see fig. 1*b* and table 1). In non-Drosophila arthropods, these novel emerging domains amount for 0.02 of the approximately two domains gained per My. The Drosophila clade features the largest number of emerging domains with more than 50% of all events dated to Drosophila or a descendant node of Drosophila. Within the Drosophila clade, the average emergence rate is roughly 0.06 domains per My. Ergo, the Drosophila lineages see a 3-fold increase in domain emergence in comparison to the rest of the pancrustacean species. Since emergent domains are a potential resource of evolutionary innovation, we draw

Table 1. Domains Emerging Within the Pancrustacean Clade .

Bin	Pfam ID	Node	$P(d)$	d_f	$d_{\max}$	$\bar{x}_d$	U_d	NCO_d
	Anophelin	Agam	1	1	1	1	1	0
	Turandot	D	0.8	47	9	6.7	1	0
NEW (6)	Sex_peptide	mel_subgrp	1	9	2	1.8	1	0
	DUF3629	mel_grp	1	9	3	1.5	3	2
	Acp26Ab	D	0.8	7	1	1	1	0
	MAGSP	mel_subgrp	1	6	2	1.2	1	0
	Bin average		0.93(0.1)	13.16(16.4)	n/a	2.2(2.3)	1.3(0.8)	0.3(0.8)
	GYR	Drosophila	1	390	39	32.5	7	6
	DUF733	Drosophila	1	111	12	9.2	1	0
	L71	Drosophila	1	87	20	7.2	2	1
	Dec-1	Drosophila	0.6	52	10	6.5	2	3
	Vitelline_membr	Drosophila	1	63	6	5.2	2	1
RECENT (11)	ACP53EA	Drosophila	0.83	43	7	4.3	1	0
	DUF2967	Drosophila	1	14	3	1.1	1	0
	Roughex	Drosophila	1	13	2	1	1	0
	DEC-1_C	Drosophila	0.9	13	2	1.8	3	3
	P53_C	Drosophila	1	12	1	1	1	1
	Antimicrobial10	Drosophila	1	12	1	1	1	0
	Bin average		0.94(0.1)	73.64(110.3)	n/a	6.44(9.1)	2(1.8)	1.36(1.9)
	DUF1213	Diptera	0.8	436	58	36.3	3	2
	Retinin_C	Endopterygota	0.89	165	22	9.7	3	2
	DUF1431	A	0.94	154	18	9.6	1	0
	DUF1091	Diptera	1	150	16	10	3	2
	DIM	Pancrustacea	0.65	99	10	7.6	1	0
	DUF1074	A	0.94	72	10	4.5	7	6
OLD (13)	Dscam_C	Pancrustacea	0.95	19	1	1	4	3
	DEC-1_N	B	0.81	18	3	1.3	5	4
	DUF3610	Endopterygota	0.47	15	5	1.6	5	5
	Pfam-B_3809	Diptera	0.4	15	4	2.5	2	1
	OMB	Endopterygota	0.68	14	2	1	7	9
	MSSP	B	0.56	12	4	1.3	1	0
	FTZ	Diptera	0.8	12	1	1	1	1
	Bin average		0.76(0.2)	90.84(120)	n/a	6.72(9.6)	3.31(2.2)	2.69(2.7)
	ANCIENT domains		0.40(0.5)	61.27(275.4)	n/a	4.35(14.1)	22.51(94)	3.95(11.4)

NOTE.— The Bin signifies the age of the emergence event (see fig. 1); Pfam ID is the ID of the emerging domains; Node represents the node at which the respective domain emerges (labeled as in fig. 1, mel_grp and mel_subgrp represent the melanogaster group and subgroup, respectively; Agam represented *A. gambiae*); $P(d)$ denotes the prevalence of emerging domains (see text); d_f denotes the total number of domain instances after resolving overlaps; $d_{\max}$ represents the maximum count of the emerging domain d in any one proteome; $\bar{x}_d$ signifies the average count of the emerging domain; U_d signifies the number of unique arrangements with the emerging domain; NCO_d shows the number of co-occurring domains. The bin average is indicated below each bin section, with standard deviation indicated in parentheses. Average properties of ANCIENT domains, while not emergent, are indicated for comparison.

attention to their possible origins, evolutionary dynamics, molecular properties, and adaptive potential.

Radiation of Emerging Domains

After domains first emerge, they may spread rapidly among all descendants or remain invariant along some lineages while expanding along others. For each emerged domain, we extracted all instances and examined their properties in the extant species.

The 30 domains that emerge within the pancrustacean clade affect a total of 1,291 proteins (∼0.36% of all proteins), to which they either are fused or form single-domain proteins. The distribution of domains in proteins affected by emerging domains suggests that older domains have more cooccurring domains within arrangements, whereas younger domains more likely form single-domain proteins. In order to estimate the "evolutionary success" of domains after they emerge, we calculated the prevalence P of a domain d defined as $P(d) = n_d/n_N$, where n_d is the number of child nodes that contain d and n_N the total number of leaves a given node has. Domains that emerge in the ANCIENT bin, that is, which are older than 430 My have the lowest average prevalence and the strongest deviation with 0.4 ± 0.5 (see table 1).

Roughly 80% of domains that emerge in the OLD bin form multidomain proteins with an average number of approximately seven neighbors per protein. In contrast, only roughly 50% of the domains in the RECENT bin form multidomain proteins and have on average less cooccurring domains with only ∼1.3 neighbors on average. Domains that have recently emerged and are younger than 40 My old mostly form single-domain proteins, with only one-sixth of the emerging domains found in multidomain proteins.

If novel domains are the result of recruitment from noncoding regions, they might display a higher content of residues in disorder than, for example, ancient domains; recent evidence indicates that disorder may be evolutionarily difficult to maintain (Schaefer et al. 2010) and that gained domains contain a high proportion of disorder (Buljan et al. 2010). We extracted all sequences of emerging

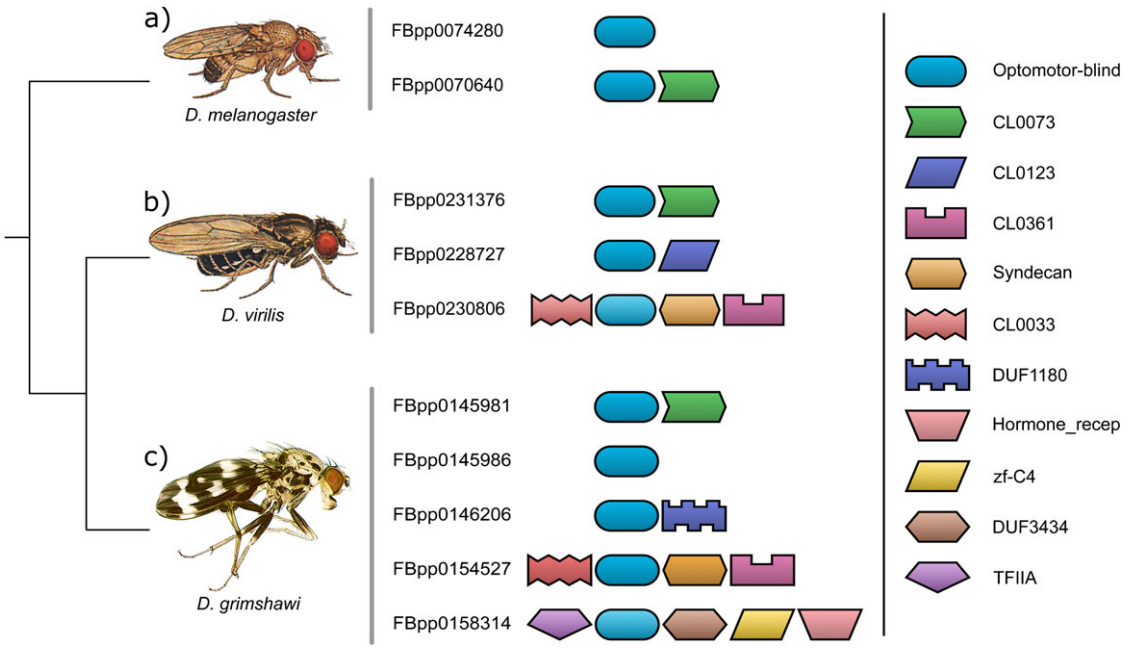

FIG. 2. Arrangements of OMB domains in three species of Drosophila. Domains are represented as shapes; OMB is shown as oval box. The *E* value cutoff for the presented arrangements is $\leqslant 0.01$. (*a*) *Drosophila melanogaster* has two different arrangements with OMB, one of which includes the T-box domain (arrow-shaped polygon, member of Pfam clan CL0073). The majority of species share the latter arrangement. (*b*) *Drosophila virilis* has a slightly different morphology and has three arrangements with OMB, where one instance is found in a region of domain overlap. (*c*) *Drosophila grimshawi* exhibits a strikingly different morphology and has, as the only species of Drosophila, a total of five arrangements that contain traces of omb where it cooccurs or overlaps with domains that have been implicated in growth, development, and transcriptional regulation.

domains from extant species and calculated the proportion of disorder in the sequence using VSL2 (Peng et al. 2006). We indeed find that domains that emerge within pancrustacea show a significantly higher proportion of disorder than ancient domains (Kruskal–Wallis $P \ll 0.001$, see also supplementary fig. 1, Supplementary Material online). There were, however, no conclusive differences between the age bins (data not shown) that may be due to the small sample size. Furthermore, no significant differences between domains in age bins could be found with respect to average sequence length of domains and average sequence similarities between instances of a domain (data not shown).

Finally, we looked into the position of emergent domains within the *D. melanogaster* genome, as it has, to date, the most complete assembly. The majority of *D. melanogaster* chromosomes harbor less than 1% emergent domains, with two exceptions. On the X chromosome, 2% of the domains (72 of 3,738) are emergent; on the 3L chromosome, 1.5% domains (67 of 4,327) are emergent. Although insufficient for statistical inference, this could hint that novel domains result from increased evolutionary rates on the X chromosome, for which some evidence has been obtained (Baines et al. 2008).

Functional Impact of Novel Domains

Recently emerging domains are, by definition, restricted to a relatively small clade and therefore not widely distributed. Accordingly, they are not always functionally and structurally well characterized. Twenty-nine of the 30 emerging domains are Pfam-A, 20 of which have been previously characterized. Only 6 of the 20 Pfam-A domains are functionally classified by the GO (Reference Genome Group of the Gene Ontology Consortium 2009). Five of the emerging domains are defined as "Domain of unknown function" (DUF), and only one of the emerging domains (DUF1074) is a member of a Pfam clan. Nonetheless, some of the proteins that gain emergent domains have been studied extensively.

The optomotor-blind (OMB) domain, for example, occurs N-terminal of the OMB protein that plays manifold regulatory roles in development (Pflugfelder 2009). The OMB domain frequently co-occurs with members of the T-box family, an ancient family of transcriptional regulators thought to be a key player in animal development (Wilson and Conlon 2002). In *D. melanogaster*, the OMB domain has been identified as a key element in the establishment of wing and abdominal pigmentation patterns (Brisson et al. 2004). Furthermore, the OMB proteins have been linked to a diverse array of morphological traits including structure of the head and external genitalia (Pflugfelder 2009) and are thought to impact transcription of a number of basal developmental genes such as *tkv*, *mtv*, *vg*, and *sal* (del Alamo Rodríguez et al. 2004). Some of these genes are targets of decapentaplegic (*dpp*), a morphogen of prime importance in Drosophila development (Nellen et al. 1994). The OMB domain emerges along the branch of endopterygota and has subsequently been lost along some lineages while maintained along the others. By loosening the *E* value threshold up to ≤ 0.1, we can detect traces of the OMB domain in all other child nodes of endopterygota, with the exception of *B. mori* and *A. gambiae*. Furthermore, we find additional copies in species that

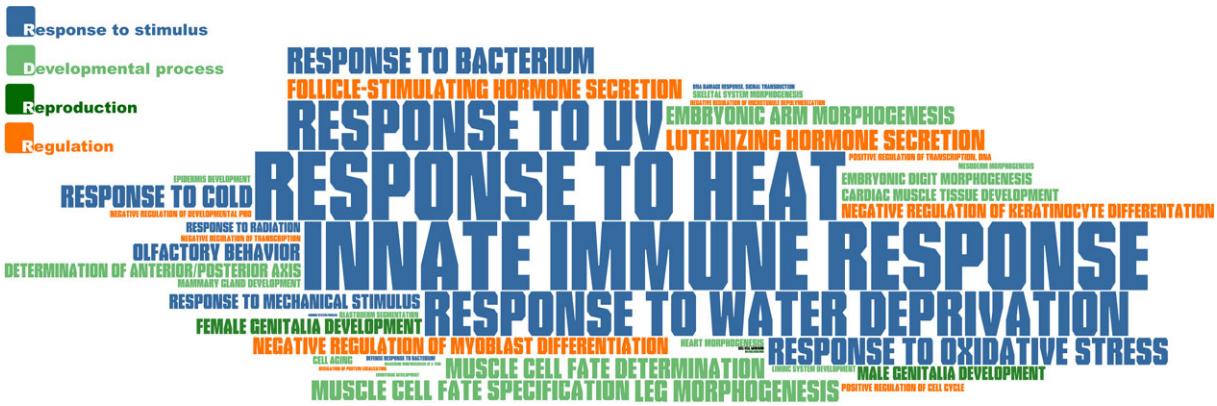

FIG. 3. *P* value transformed TermLogo of functional groups with emerging domains. GO terms effected by emergence were subject to an overrepresentation analysis with the weighted algorithm of the TopGO package and using all GO terms present in pancrustacea as universe (see Methods). The size of the font corresponds to the strength of significance obtained for the term. Significance was established after correction for multiple testing using Bonferroni at $P < 0.01$. The color coding corresponds to parental nodes in the GO graph. The majority of the significant terms are related to stimulus response. Only the term "cell adhesion," displayed in black, is not included in one of the four categories displayed in the top left as parent node.

already bear a copy of the OMB domain. For example, after loosening the match requirements, we detect traces of additional four copies of OMB within *D. grimshawi*, where they occur in arrangements absent in all other pancrustacean species (see fig. 2). *D. grimshawi* is endemic to the island of Hawaii and is known for its strikingly different morphology in comparison to other Drosophila, including the diverse array of wing pigmentation patterns (Edwards et al. 2007).

In order to globally assess the functional effect of domain emergence, and to overcome the weak links to GO categories that emerging domains exhibit, we analyzed the GO annotations of proteins that recruited emergent domains using Blast2GO (Conesa and Götz 2008).

From the `biological_process` ontology, a strikingly high number of the statistically most significant terms correspond to environmental adaptation such as response to heat, drought, UV, and other abiotic stresses (see fig. 3 and supplementary table 2, Supplementary Material online). This is followed by response to biotic stress and terms relating to sex differentiation and further to development and morphogenesis.

The pancrustacean species considered here contain a number of highly specialized, geographically restricted species. *D. sechellia*, for example, habituates an archipelago of 115 islands in the Indian Ocean and feeds off a fruit found toxic to most other Drosophila species (Farine et al. 1996). Similarly, *D. erecta* and *D. mojavensis* are highly specialized species with restricted geographic distributions (Singh et al. 2009). The Drosophila clade also contains cosmopolitan species such as *D. melanogaster* or *D. simulans*. Some Drosophila species find optimal conditions in high temperature areas, such as *D. mojavensis*, which is found in North American deserts where it feeds off rotting cactus, or species of the *obscura* group, which seek near-desert habitats during winter (Markow and O'Grady 2007). The differences among the Drosophila species affect courtship, developmental time from egg to adult, as well as morphological traits (see Markow and O'Grady (2007) and references therein).

The protein functionalities affected by emerging domains reflect these differences and illustrate the diverse life history and the outstanding success of the pancrustacea, in particular the cosmopolitan species of Drosophila, in adapting to new environments.

Within the other two ontologies, we find the `cellular_component` term, `extracellular_space`, as well as terms from the `molecular_function` ontology related to DNA binding and transcriptional regulation to be overrepresented.

Conclusion

Previous studies have estimated genome-wide gene turnover rates, that is, gene gain and loss, within the Drosophila clade (Hahn et al. 2007; Rogers et al. 2010). We find lower domain turnover rates for domains than for genes. This is in line with our expectations since the average domain copy number across a given proteome is 4 ± 15. Accordingly, a gene gain or loss event will, on average, only affect few domains, many of which will retain copies in other genes. Ergo, although the copy number of domains will be subject to fluctuation, the presence or absence of domains, such as is considered here, will not be affected. A potentially confounding factor in the analysis of domain gain and loss is erroneous domain annotations. Accelerated rates of evolution or sequence bias in domain models may facilitate a signal of domain loss or shift the point of domain gain and hence influence emergence rates in our anaylsis. However, by using the model-defined gathering thresholds for Pfam-A domains and a conservative cutoff for Pfam-B domains, we are confident that the trends in our analysis are robust. In particular, as we find that our results are in agreement with a previous study on gene family

turnover in Drosophila (Hahn et al. 2007), we similarly find increased rates of loss and gain along branches to the *simulans*/*sechellia* subclade, as well as along branches within the *obscura* group.

Our results indicate that thousands of domains are lost along every lineage. High rates of domain loss seem to entail a strong loss of evolutionary potential for further innovation as de novo formation of novel functional domains is likely difficult (Bornberg-Bauer et al. 2010). Just how precisely can this loss of evolutionary potential be compensated, considering the need of species to adapt to an ever-changing environment? First, depletion of the repertoire of functional domains may be offset by the creation of new domain arrangements. Over evolutionary long timescales, domain arrangements become longer and more diverse and assume new functions (Björklund et al. 2005; Ekman et al. 2005; Fong et al. 2007; Wang and Caetano-Anollés 2009). Second, new or strongly divergent proteins without any apparent homology even to closely related species (and accordingly without any domain assignment) can be found in any newly sequenced genome (see, e.g., Drosophila Genome Consortium 2007; Werren et al. 2010). Such orphan genes can make up to 30–40% of the gene repertoire and seem to be of particular importance for adaptation; their spatiotemporal expression profiles can be very specific for tissues, developmental stages, and reproductive division of labor (Colbourne et al. 2011; Johnson and Tsutsui 2011). Third, as is shown in this study, the emergence of new domains is of great adaptive value and, accordingly, emerging domains spread rapidly across genomes.

Finally, given the use of Dollo parsimony, loss rates should be considered an upper boundary (see also Methods). However, given the comparably shallow tree employed here, results that are in agreement with studies that employed an alternative model (Hahn et al. 2007) and similar signals found among other taxonomic groups (Zmasek and Godzik 2011), we are confident that the overall trend should prevail.

The emergence and rapid spread of novel domains are particularly striking. Domains emerge frequently in the context of abiotic stress, biotic defense, reproduction, and development. The former two categories have not been reported by studies focusing on gene families (Hahn et al. 2007). A possible explanation is that domains affect only small parts of proteins and may thus be overlooked if they are incorporated in one protein out of many of a family. Furthermore, the rates of emergence reported here must be seen as a lower boundary. A novel domain can, almost by definition, not be reported by current bioinformatic techniques. Hidden Markov models (HMMs), a technique on which, for example, Pfam builds, first require several instances of a domain to build a profile. Accordingly, very recent domains that may still be strongly diverging or have just a single instance, for example, in orphan proteins, will be overlooked.

The origin of new proteins remains generally elusive (Levine et al. 2006; Bornberg-Bauer et al. 2010) and only very rarely can be accurately reconstructed. Here, it was found that novel domains mostly form single-domain proteins and are significantly enriched in disordered regions. Both facts indicate that novel domains are either the result of de novo formation from DNA, possibly via intermediate RNA genes (Zhou et al. 2008), or structurally very flexible in choosing novel ligands or binding partners or both. In contrast, older domains have more neighbors, form a larger variety of arrangements, and less frequently form single-domain proteins than newer domains. This is in line with previous studies (Vogel et al. 2005) that indicate that the process of modular rearrangement is at least partly fueled by random attachment.

To our knowledge, the study presented here is the first to date to assess the amount of domain gain, loss, and emergence within a dense and exceptionally well-studied clade. Furthermore, since potentially confounding effects such as whole-genome duplications are absent, the derived rates of loss and emergence will help set a framework to push further the limits of phylogenetic inferences and sequence comparison based on domain arrangements (Björklund et al. 2005; Yang et al. 2005; Fong et al. 2007; Song et al. 2008). The greater accuracy of HMMs in identifying homologous sequences and the relatively low rates of domain turnover (as opposed to amino acid replacements) help capture functional shifts at a rather coarse-grained level and across evolutionary long timescales of tens to hundreds of My. The combination of indirect functional inference of GO terms (by analyzing proteins that acquire novel domains) and graphical representation of the statistical analysis as illustrated here provide an intuitive representation of adaptive signals. Accordingly, our method should be applicable to most genome projects for which it offers a valuable complement to other more established methods such as site-based statistical analysis or studies of gene families.

Supplementary Material

Supplementary figures 1–3 and tables 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the DFG (Deutsche Forschungs Gemeinschaft) grant BO 2455/4-1 to E.B.B.; A.D.M. acknowledges support by the WWU-PAS stipend. The authors would like to thank two anonymous reviewers for critical reading and helpful comments on the manuscript. Author contributions: E.B.B. and A.D.M. conceived the study; A.D.M. conducted all analyses; E.B.B. and A.D.M. wrote the manuscript.

References

Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics* 22:1600–1607.

Apic G, Gough J, Teichmann SA. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol.* 310: 311–325.

Baines JF, Sawyer SA, Hartl DL, Parsch J. 2008. Effects of x-linkage and sex-biased gene expression on the rate of adaptive protein evolution in drosophila. *Mol Biol Evol*. 25:1639–1650.

Björklund AK, Ekman D, Light S, Frey-Skött J, Elofsson A. 2005. Domain rearrangements in protein evolution. *J Mol Biol*. 353:911–923.

Bornberg-Bauer E, Huylmans AK, Sikosek T. 2010. How do new proteins arise? *Curr Opin Struct Biol*. 20:390–396.

Brisson JA, Templeton AR, Duncan I. 2004. Population genetics of the developmental gene *optomotor-blind* (*omb*) in *Drosophila polymorpha*: evidence for a role in abdominal pigmentation variation. *Genetics* 168:1999–2010.

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

Chothia C, Gough J, Vogel C, Teichmann SA. 2003. Evolution of the protein repertoire. *Science* 300:1701–1703.

Colbourne JK, Pfrender ME, Gilbert D, et al. (69 co-authors). 2011. The ecoresponsive genome of *Daphnia pulex*. *Science* 331: 555–561.

Conesa A, Götz S. 2008. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics*. 2008:619832.

del Alamo Rodríguez D, Felix JT, Díaz-Benjumea FJ. 2004. The role of the T-box gene *optomotor-blind* in patterning the drosophila wing. *Dev Biol*. 268:481–492.

Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, et al. (418 co-authors). 2007. Evolution of genes and genomes on the drosophila phylogeny. *Nature* 450:203–218.

Edwards KA, Doescher LT, Kaneshiro KY, Yamamoto D. 2007. A database of wing diversity in the hawaiian drosophila. *PLoS One* 2:e487.

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

Ekman D, Björklund AK, Frey-Skött J, Elofsson A. 2005. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J Mol Biol*. 348:231–243.

Elofsson A, Sonnhammer EL. 1999. A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics* 15:480–500.

Farine JP, Legal L, Moreteau B, Quere JLL. 1996. Volatile components of ripe fruits of morinda citrifolia and their effects on drosophila. *Phytochemistry* 41:433–438.

Farris JS. 1977. Phylogenetic analysis under Dollo's law. *Syst Zool*. 26:77–88.

Finn RD, Mistry J, Tate J, et al. (14 co-authors). 2010. The Pfam protein families database. *Nucleic Acids Res*. 38:D211–D222.

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

Hahn MW, Han MV, Han SG. 2007. Gene family evolution across 12 drosophila genomes. *PLoS Genet*. 3:e197.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22:2971–2972.

Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443:931–949.

Itoh M, Nacher J, Kuma KI, Goto S, Kanehisa M. 2007. Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. *Genome Biol*. 8:R121.

Johnson BR, Tsutsui ND. 2011. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. *BMC Genomics* 12:164.

Kersey P, Bower L, Morris L, et al. (20 co-authors). 2005. Integr8 and genome reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res*. 33:D297–D302.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet*. 25:404–413.

Koide S. 2009. Generation of new protein functions by nonhomologous combinations and rearrangements of domains and modules. *Curr Opin Biotechnol*. 20:398–404.

Kummerfeld SK, Teichmann SA. 2005. Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet*. 21:25–30.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently x-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A*. 103:9935–9939.

Levitt M. 2009. Nature of the protein universe. *Proc Natl Acad Sci USA*. 106:11079–11084.

Markow TA, O'Grady PM. 2007. Drosophila biology in the genomic age. *Genetics* 177:1269–1276.

Meusemann K, von Reumont BM, Simon S, et al. (16 co-authors). 2010. A phylogenomic approach to resolve the arthropod tree of life. *Mol Biol Evol*. 27:2451–2464.

Moore AD, Björklund AK, Ekman D, Bornberg-Bauer E, Elofsson A. 2008. Arrangements in the modular evolution of proteins. *Trends Biochem Sci*. 33:444–451.

Nellen D, Affolter M, Basler K. 1994. Receptor serine/threonine kinases implicated in the control of drosophila body pattern by decapentaplegic. *Cell* 78:225–237.

Pal LR, Guda C. 2006. Tracing the origin of functional and conserved domains in the human proteome: implications for protein evolution at the modular level. *BMC Evol Biol*. 6:91.

Pasek S, Bergeron A, Risler JL, Louis A, Ollivier E, Raffinot M. 2005. Identification of genomic features using microsyntenies of domains: domain teams. *Genome Res*. 15:867–874.

Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. 2006. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 7:208.

Pflugfelder GO. 2009. *omb* and circumstance. *J Neurogenet*. 23:15–33.

Reference Genome Group of the Gene Ontology Consortium. 2009. The gene ontology's reference genome project: a unified framework for functional annotation across species. *PLoS Comput Biol*. 5:e1000431.

Rogers HH, Bergman CM, Griffiths-Jones S. 2010. The evolution of tRNA genes in Drosophila. *Genome Biol Evol*. 2:467–477.

Schaefer C, Schlessinger A, Rost B. 2010. Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. *Bioinformatics* 26:625–631.

Singh ND, Larracuente AM, Sackton TB, Clark AG. 2009. Comparative genomics on the drosophila phylogenetic tree. *Annu Rev Ecol Evol Syst*. 40:459–480.

Song N, Joseph JM, Davis GB, Durand D. 2008. Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput Biol*. 4:e1000063.

Tjoelker LW, Gosting L, Frey S, Hunter CL, Trong HL, Steiner B, Brammer H, Gray PW. 2000. Structural and functional definition of the human chitinase chitin-binding domain. *J Biol Chem*. 275: 514–520.

Vogel C, Teichmann SA, Pereira-Leal J. 2005. The relationship between domain duplication and recombination. *J Mol Biol*. 346: 355–365.

Wang M, Caetano-Anollés G. 2006. Global phylogeny determined by the combination of protein domains in proteomes. *Mol Biol Evol*. 23:2444–2454.

Wang M, Caetano-Anollés G. 2009. The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure* 17:66–78.

Weiner J, Beaussart F, Bornberg-Bauer E. 2006. Domain deletions and substitutions in the modular protein evolution. *FEBS J*. 273:2037–2047.

Werren JH, Richards S, Desjardins CA, et al. (164 co-authors). 2010. Functional and evolutionary insights from the genomes of three parasitoid nasonia species. *Science* 327:343–348.

Wilson V, Conlon FL. 2002. The T-box family. *Genome Biol.* 3:REVIEWS3008.

Yang S, Doolittle RF, Bourne PE. 2005. Phylogeny determined by protein domain content. *Proc Natl Acad Sci U S A.* 102:373–378.

Yu Y, Lutz S. 2011. Circular permutation: a different way to engineer enzyme structure and function. *Trends Biotechnol.* 29:18–25.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in Drosophila. *Genome Res.* 18:1446–1455.

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:R4.