BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

The modular nature of protein evolution: domain rearrangement rates across eukaryotic life

Elias Dohmen[1,2†], Steffen Klasberg[1†], Erich Bornberg-Bauer[1], Sören Perrey[2] and Carsten Kemena[1*]

Abstract

Background: Modularity is important for evolutionary innovation. The recombination of existing units to form larger complexes with new functionalities spares the need to create novel elements from scratch. In proteins, this principle can be observed at the level of protein domains, functional subunits which are regularly rearranged to acquire new functions.

Results: In this study we analyse the mechanisms leading to new domain arrangements in five major eukaryotic clades (vertebrates, insects, fungi, monocots and eudicots) at unprecedented depth and breadth. This allows, for the first time, to directly compare rates of rearrangements between different clades and identify both lineage specific and general patterns of evolution in the context of domain rearrangements. We analyse arrangement changes along phylogenetic trees by reconstructing ancestral domain content in combination with feasible single step events, such as fusion or fission. Using this approach we explain up to 70% of all rearrangements by tracing them back to their precursors. We find that rates in general and the ratio between these rates for a given clade in particular, are highly consistent across all clades. In agreement with previous studies, fusions are the most frequent event leading to new domain arrangements. A lineage specific pattern in fungi reveals exceptionally high loss rates compared to other clades, supporting recent studies highlighting the importance of loss for evolutionary innovation. Furthermore, our methodology allows us to link domain emergences at specific nodes in the phylogenetic tree to important functional developments, such as the origin of hair in mammals.

Conclusions: Our results demonstrate that domain rearrangements are based on a canonical set of mutational events with rates which lie within a relatively narrow and consistent range. In addition, gained knowledge about these rates provides a basis for advanced domain-based methodologies for phylogenetics and homology analysis which complement current sequence-based methods.

Keywords: Protein domain, Rearrangement rates, Proteome analysis, Evolutionary history, Ancestral reconstruction

Background

Functional adaptations of proteins have often been observed to be caused by point mutations changing amino acids at crucial positions. These mutations typically result in altered specificity or stability of a protein. Although this process is important for evolutionary adaptations, point mutations often result in only minor changes of a protein. For greater functional changes or innovation, more drastic modifications are necessary that do not rely on numerous mutations.

Molecular mechanisms like crossing over, alternative splicing and transposition through mobile elements can cause mutational events that rearrange larger DNA fragments and therefore also alter larger regions at the protein level. Examples of such mutational events, which rearrange gene content, are for example fusion and fission. All these events lead to rearrangements that can be easily tracked at the level of protein domains, since domains are well characterised in many databases (e.g. in the *Pfam*

*Correspondence: c.kemena@wwu.de
†Elias Dohmen and Steffen Klasberg contributed equally to this work.
[1]Institute for Evolution and Biodiversity, University of Münster, Hüfferstrasse 1, 48149 Münster, Germany
Full list of author information is available at the end of the article

[1] or *Superfamily* [2] database) and represent reusable structural and functional units.

The total number of defined domains is relatively small and is growing only slowly. For example, the *Pfam* domain database [1] defines about 18,000 domains in its current version (version 32). On the other hand, the number of known unique domain arrangements - defined by the linear order of domains in an amino acid sequence [3] - is much larger and growing rapidly [4]. Accordingly, rearrangements of existing domains can help explain the vast protein diversity we observe in nature [4–9].

Several studies have shown that domain rearrangements are essential in the evolution of pathways, signalling networks and cellular components. The evolution of the extracellular matrix in metazoans [10] as well as the blood coagulation cascade [11] are examples in which the reuse of domains in different contexts are considered crucial steps. Additionally, domains have been identified to play an important role in signalling networks [12] or their recombination to new arrangements in T-Cell development [13]. Lees et al. [14] showed the importance of domain arrangement changes in cancer genome evolution. Therefore, it is crucial to analyse domain changes when studying both genome evolution and specific protein families.

First attempts to study general evolutionary domain patterns focused mainly on emergence and loss of single domains [15, 16] or domain repeats [17, 18]. Later, quantitative analyses in plants and insects [19, 20] over time-scales of several hundred million years revealed hotspots of rearrangement events at specific nodes in the phylogenetic tree. Both these studies took into account four different types of rearrangement events: fusion, fission, terminal addition and terminal loss. Together, these events are sufficient to explain a large proportion (60%-70%) of the new domain arrangements considered in those studies.

Based on these four single step events, rearrangement rates for a set of 29 plant species (dating back as far as 800 my [19]) and 20 Pancrustacean species (dating back 430 my [20]) were determined in previous studies.

In this study we use expanded species sets (up to 72 species per phylogenetic clade) to detect common patterns of domain evolution and consider several thousand more arrangements per clade compared to the two previously mentioned studies. In total, domain arrangements in five different eukaryotic clades (vertebrates, insects, fungi, monocots and eudicots) are analysed. For the first time, the results can be directly compared between these clades, since exactly the same methodology was applied to all of them.

Previously, methods were applied that had used either overlapping definitions for rearrangement events, or that analysed domain loss and emergence (e.g. [16]) separately from rearrangement events (e.g. [20]). In this study, we combine these methodologies in one consistent model, allowing us to distinguish six different single step events, thereby analysing the molecular mechanisms leading to protein innovation at unprecedented accuracy. The incorporation of additional clades and a higher number of species ensures the integrity of the observed events, for example by minimising annotation biases. The resulting rearrangement frequencies are directly comparable across the different eukaryotic clades and thus reveal the fundamental mechanisms of functional rearrangements in eukaryotes, in addition to lineage specific trends.

Furthermore, we infer functional implications of the new arrangements via *Gene Ontology* (GO) [21] term enrichment. Finally, we discuss how our methodology can be used to complement existing methods for example in phylogenetic reconstruction, by incorporating data on domain rearrangements.

Results

To be able to draw reliable conclusions about universally valid mechanisms in protein evolution, it is necessary to ensure that a sufficient number of observable rearrangements can be explained by the six different rearrangement events defined in this manuscript (*fusion*, *fission*, *terminal loss/emergence* and *single domain loss/emergence*; see Methods). For this purpose we reconstructed the ancestral domain content and arrangements at all inner nodes of the phylogenetic trees of five eukaryotic clades (vertebrates, insects, fungi, monocots and eudicots). For all domain arrangements that differ from the parental node, we examined whether the change could be explained uniquely by one of the six events.

Unique solutions are either *exact solutions*, where only a single event can explain the arrangement change, or *non-ambiguous solutions*, where multiple events of the same type can explain a new arrangement (e.g. ABC: A+BC / AB+C). Only unique solutions were further analysed in detail to focus on changes which can be explained with certainty (Additional file 2). Unique solutions can explain 50% to 70% of all observed new arrangements, depending on the analysed phylogenetic clade (Fig. 1).

However, there is a small percentage of new arrangements which can be explained by multiple different event types, i.e. *ambiguous solutions* (e.g. ABC: ABC-D / AB+C). Beside these ambiguous solutions, some new arrangements cannot be explained by the defined single step events. These so-called *complex solutions* (25%-50%), would require several successive single step events.

Comparison between clades

One major goal of this study is to find, beside clade-specific differences, universally valid evolutionary mechanisms of protein innovation that are present in all

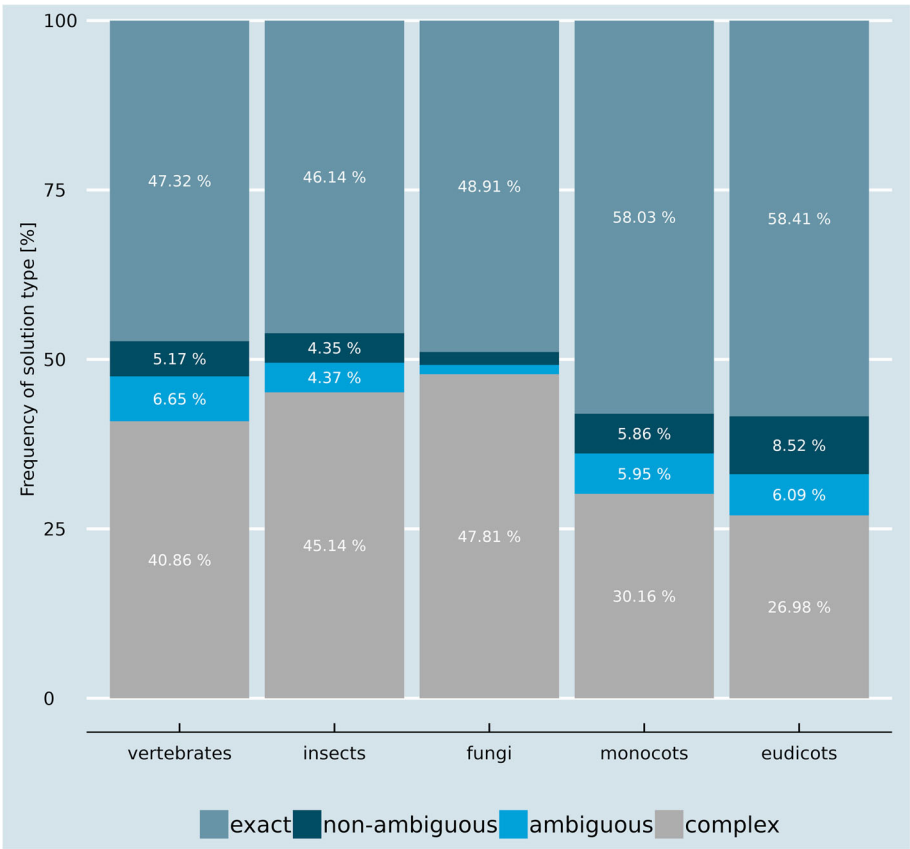

Fig. 1 Frequency of the different solution types. Exact and non-ambiguous solutions can be found in about 50% of the cases

clades. Therefore, we analyse whether common patterns in domain rearrangements can be observed by measuring the relative contributions of each rearrangement event and compare them between the different clades (see Table 1 and Additional file 4).

The percentage of fusion events in our study ranges from 29% in fungi to 64% of all observed events in monocots. Only in fungi, fusions represent not the most frequent event type, but single domain loss is most frequent. Furthermore, in all clades except fungi, fissions and terminal losses account for a similar percentage of all domain rearrangements. In fungi, loss of terminal domains accounts for twice as many rearrangements as fissions. The exceptional distribution of event frequencies in fungi compared to the other clades is discussed below.

The very low contributions of the two emergence categories, terminal and single domain emergence, of only 0.13% to 3.89% show that domain emergence is indeed rare compared to a much higher number of domain rearrangements and losses.

We observed three general patterns of the ranks of rearrangement events corresponding to the taxonomic kingdoms of animals, fungi, and plants. In the first pattern, observed in animals (i.e. vertebrates and insects), the most frequent domain rearrangement event is domain fusion (32% and 42% of rearrangements respectively), followed

Table 1 Frequencies of the six rearrangement events (in %)

	Vertebrates	Insects	Fungi	Monocots	Eudicots
Fusion	32.45	41.52	29.35	64.43	58.22
Fission	19.57	17.21	8.80	12.21	16.28
Terminal loss	20.52	19.21	16.46	10.59	13.00
Terminal emergence	0.13	0.36	0.76	1.01	0.48
Single loss	26.71	19.99	40.74	9.83	10.20
Single emergence	0.61	1.71	3.89	1.93	1.82

by single domain loss (27% and 20%) and terminal domain loss (21% and 19%). Arrangement gain by fission is slightly less common (20% and 17%), but still more frequent than the very low rates of single domain emergence (0.6% and 1.7%) and terminal emergence (0.1% and 0.4%).

The functional analysis of gained arrangements in insects (Additional file 5) using GO term enrichment reveals olfaction related adaptations (represented by GO terms of 'sensory perception of smell', 'olfactory receptor activity' and 'odorant binding') are overrepresented in insects. Other overrepresented GO terms include 'sensory perception of taste' and 'structural constituent of cuticle'.

We did not find expansions of vertebrate specific GO terms at the root of vertebrates. However, we found overrepresented GO terms related to binding (e.g. 'protein binding', 'nucleic acid binding') and terms related to signal transduction (Additional file 6).

The distribution and rank of rearrangement rates in Fungi (Additional file 7) resemble those of animals, with the only qualitative difference being that single domain losses were more frequent than fusions. A more detailed analysis of this phenomenon can be found below.

The third pattern of arrangement changes is observed in plants, i.e. monocots and eudicots. As in metazoans, but with an even higher percentage, the majority of new arrangements is explained by fusion (64% and 58%). The fission of one arrangement into two new arrangements is the second most frequent mechanism (12% and 16%) followed by slightly smaller numbers of terminal (11% and 13%) and single domain loss (10% and 10%).

Some GO terms are enriched in gained arrangements at the root of both plant clades that might be related to plant development and evolution, i.e. 'recognition of pollen' in both plant clades or 'plant-type cell wall organization' in eudicots (Fig. 2 and Additional file 8).

Domain loss in fungi

We analysed the distribution of domain arrangement sizes in the five clades (see Additional file 9) to find possible explanations for the different patterns of event frequencies mentioned above. The results show that a strikingly high number of fungal domain arrangements consists of just a single domain and their arrangements are generally much shorter compared to vertebrates or insects. Both plant clades, monocots and eudicots, also have much shorter domain arrangements than the metazoan clades.

We found that both plant clades show the highest copy number of domain arrangements. Eudicots have 5.79 copies on average per single domain arrangement per species, while monocots have 5.64. This high number of duplications of the same domain arrangement could be explained by multiple whole genome duplications in these clades. Vertebrates follow with 1.93 copies per single domain arrangement and finally insects (1.27), while fungi show the lowest duplication count (1.15).

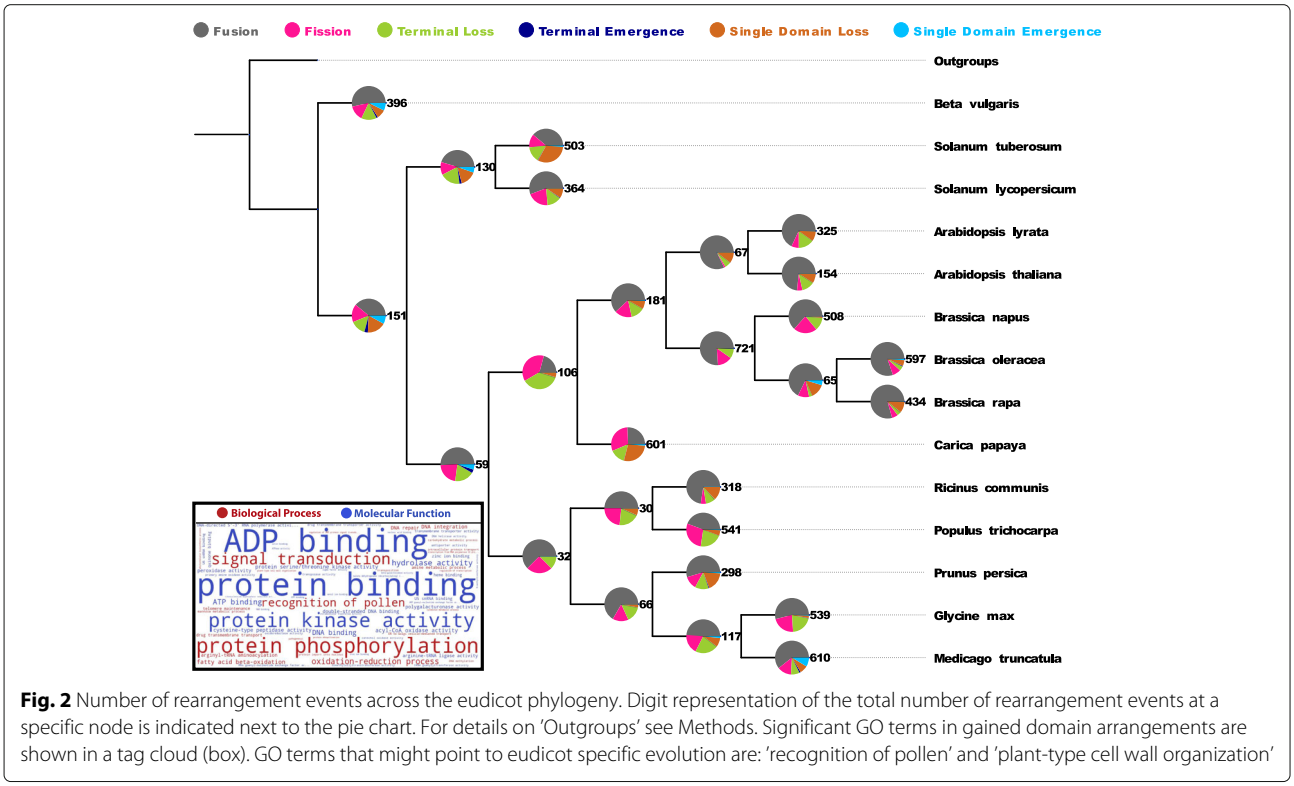

Fig. 2 Number of rearrangement events across the eudicot phylogeny. Digit representation of the total number of rearrangement events at a specific node is indicated next to the pie chart. For details on 'Outgroups' see Methods. Significant GO terms in gained domain arrangements are shown in a tag cloud (box). GO terms that might point to eudicot specific evolution are: 'recognition of pollen' and 'plant-type cell wall organization'

Effects of domain rearrangements

The general rates of rearrangement events and their distribution in a given phylogenetic tree can provide an insight into the evolutionary history of a whole clade as well as general adaptational processes in certain lineages. However, by taking a more detailed look at the specific domains involved in the rearrangement events at specific time points, we can trace back some major steps in the evolutionary history of the studied species. Here, we show three examples of new or outstanding functions at specific nodes in the evolution of vertebrates, plants and insects which can be related to the emergence of new domains or domain arrangements.

The origin of hair and adaptations of the immune system in mammals

One remarkable pattern in the distribution of rearrangement events in the vertebrate phylogeny is the high rate (33%) of single domain emergences at the root of all mammals. This represents the highest percentage of single domain emergences at any node in the vertebrate tree. A closer investigation of the function of these emerged domains shows that ∼30% of the emerged domains (domains of unknown function excluded) are associated with hair. This finding is a strong signal for the origin of hair or fur, respectively, in the common ancestor of all mammals.

One of the most important structural protein families of mammalian hair is the keratin-associated protein family (KRTAPs). Hair keratins are embedded in an inter-filamentous matrix consisting of KRTAPs located in the hair cortex. Two major types of KRTAPs can be distinguished: high-sulfur/ultra-high-sulfur and high-glycine/tyrosine KRTAPs [22]. Three of these high-sulfur proteins can be found in the set of emerged domains as 'Keratin, high sulfur B2 protein' (Pfam-ID: PF01500), 'Keratin-associated matrix' (PF11759) and 'Keratin, high-sulphur matrix protein' (PF04579). The proteins are synthesised during the hair matrix cell differentiation and form hair fibres in association with hair keratin intermediate filaments. Another domain that can be found in this set is the 'PMG protein' (PF05287) domain, which occurs in two genes in mice (PMG1 and PMG2) that are known to be expressed in growing hair follicles and are members of a KRTAP gene family [23]. PMG1 and PMG2 are additionally involved in epithelial cell differentiation, while a further member of the emerged domains - 'KRTDAP' (PF15200) - is a keratinocyte differentiation-associated protein. Keratinocytes are a cell type of the epidermis, the layer of the skin closest to the surface [24]. The KRTDAP related gene was isolated in rats between skin of prehair-germ stage embryos and hair-germ stage embryos, and shows high expression in regions of the hair follicle [25]. We can infer that the emergence of hair and fur also involved adaptation and restructuring of the skin, resulting in novel skin cell types and cell differentiation regulation mechanisms. Furthermore, the skin, and keratinocytes in particular, act as a first barrier against environmental damage and pathogen infestation and are therefore related to the second barrier, the immune system. Indeed, immune system related domains are the second biggest group in these emerged domains (>20% of domains with known function). As an example, the 'Interleukin' domain (PF03487) emerged at the root of mammals and is associated with a group of secreted proteins and signalling molecules. The mammalian immune system is highly dependent on interleukins with certain deficiencies linked to autoimmune diseases and other immune system defects [26]. 'Lymphocyte activation family X' is a domain also found in this set (PF15681), which is membrane-associated and expressed in B- and T-cells in addition to other lymphoid-specific cell types [27]. Additionally, out of all events occurring at the root of mammals, 'regulation of lymphocyte activation' is an overrepresented term in the GO term enrichment analysis (see Additional file 10). These results reinforce the importance of the immune system for the early evolution of mammals.

Resistance to fungi in wheat

The functional analysis of gained domain arrangements using GO terms revealed an interesting pattern for the node leading to *Triticeae* which includes the two wheat species *Triticum urartu* and *Triticum aestivum* as well as the grass species *Aegilops tauschii*. Five out of the 15 enriched GO terms in *Triticeae* can be related to resistance to fungal pathogens via three different mechanisms. Chitinases are enzymes, which are known to be involved in plants' fungal resistance and have been extensively studied in wheat species [28, 29]. The ability of these enzymes to degrade chitin, a primary component of fungal cell walls, can lead to the lysis of fungal cells and therefore provide resistance against them. We found the three significant GO terms 'chitin catabolic process', 'cell wall macromolecular catabolic process' and 'protein phosphorylation' related to chitinases, which explain the innate fungal resistance of wheat and can also be utilized in genetic engineering to enhance fungal resistance in other crop plants [30]. The GO term 'protein kinase activity' and the underlying Serine Threonine kinase has also been shown to be used in plants' defense to fungi [31]. Another mechanism of fungal resistance is based on an ATP-binding cassette transporter, which is used in many crop plants [32]. We relate the GO term of 'ATP binding' to this function of fungal resistance. Overall, the gained arrangements in *Triticeae* can be linked to the increased resistance of this clade to fungal pathogens.

Eusociality in bees

We found an example of interesting GO terms enriched at a node in *Apidae*, i.e. in the last common ancestor of the honey bee *Apis mellifera* and the bumblebee *Bombus terrestris*. This node marks one of the transitions of solitary bees to eusocial bees [33]. The overrepresented GO terms that relate to the evolution of eusociality comprise 'embryonic morphogenesis', 'insulin-like growth factor binding' and 'regulation of cell growth' [33] and are additionally expanded in the species *Bombus terrestris* and *Apis cerana*. Insulin and insulin-like signalling (IIS) pathways have been shown to be differently expressed between castes in the honeybee and play a role in caste differentiation [34, 35]. Additionally, IIS modifies the behaviour of honey bee workers in foraging [36]. Functions of some domains that are associated with overrepresented GO terms can possibly be related to the emergence of eusociality, either by being involved in development or have been shown to be differentially expressed in different castes. Two domains are associated with growth factors, 'Insulin-like growth factor binding' (PF00219) [34, 35] and 'EGF-like domain' (PF00008). Epidermal growth factor (EGF) has been shown to be involved in caste differentiation in the honey bee by knockdown experiments [37, 38]. Several domains have been found to be differentially expressed in queens and workers in the honey bee and might be related to eusociality [39], i.e. 'Fibronectin type III domain' (PF00041), 'Protein kinase domain' (PF00069), 'Myb-like DNA-binding domain' (PF00249) and 'Insect cuticle protein' (PF00379). 'Insect cuticle protein' is also suspected to play a role in the transition from solitary to eusocial bees [40].

Discussion

In comparison to previous studies we can verify some of the key findings like fusions being the most common event type accounting for new domain arrangements [19, 20, 41]. At the same time we can show to what extent these findings also apply to other phylogenetic clades or where differences exist (e.g. single domain loss being the most common event type in fungi). Comparing the data basis of this study to previous ones reveals that the total number of events with a unique solution (Additional file 3) is much higher than in any previous study, while the proportion of considered solutions in other studies is similar to ours. The underlying total numbers in previous studies sum up to only a few thousand unique solutions (~5200 in Moore's pancrustacean set [20]) compared to ten thousands in this study (~24250 in the insect set, which also contains 18 out of 20 of Moore's pancrustacean species).

This increasing total number of resolvable events, while representing constant proportions over time, suggests that with increasing quality of sequences, annotations and motifs in databases we are able to explain more of the evolutionary history, but at the same time add more unknown or complex cases. However, the ambiguous and complex solutions we find in this study can be resolved to some extent with further investigation and approaches specific for this problem. In some cases, the ambiguity of ambiguous solutions might be resolved by computing domain trees based on the primary sequences. This is, though, outside the scope of this study and the information gain would be minimal as only a very low percentage (~5%) of all solutions are ambiguous ones.

Complex solutions might be resolved with the use of a deeper and denser phylogeny. Such a phylogeny might provide additional inner nodes which are required to be able to track the arrangement changes using single steps. Another potential way to resolve the underlying molecular rearrangement events of complex gains could be to consider not only single step events, but also solutions with two or more steps. However, the latter approach would strongly increase the complexity of the calculations, while at the same time introducing uncertainty by introducing multiple additional ambiguous solution possibilities.

The GO term enrichment analysis based on domain changes during evolution can give additionally useful insights into major functional adaptations of a clade. In insects for example all described enriched GO terms ('sensory perception of smell', 'olfactory receptor activity', 'odorant binding', 'sensory perception of taste' and 'structural constituent of cuticle') are essential for communication between individuals, for example to find mating partners by sensing pheromones over long distances or to tell nest mates from potential enemies in social insects [42–44]. For the fungi clade enriched terms are 'carbohydrate metabolic process' and 'cellulose binding', which can be seen as important adaptations for the lifestyle of some fungal species. Many fungal species (e.g. Serpula lacrymans) are wood-decaying, for which both metabolic functions are crucial. Another hint for the wood-decay related background of these adaptations could be the enriched GO term 'oxidation-reduction process', which can be associated to lignin deconstruction as well as to cellulose/xylan degradation.

One evolutionary mechanism of specific interest is loss of function as a process of adaptation. In this study especially the different signals for losses in plants and fungi are worth a more detailed investigation. In plants the high rates of fusion and fission and low rates of losses can be related to plant specific genome properties. Transposable elements play a major role in plants by the frequent creation of retrocopies and thus contribute to a high number of observable gene duplications in plants [45–47]. Additionally, many whole genome duplications have been observed in plants, leading to large genomes as a basis for

rearrangements while maintaining the original gene and function [47–49].

A possible explanation for the high frequency of single domain loss in fungi could be the generally high fraction of single domain arrangements in their proteomes. Such a high fraction of single domain loss is however not observed in plants, although eudicots also have a high fraction of short domain arrangements, comparable to that of fungi (Additional file 9). The difference between eudicots and fungi regarding single domain losses can be explained via the average copy number of single domain arrangements in both clades. The results of the duplication count analysis imply that fungi possess by average just one copy (1.15) of every single domain arrangement, which can explain the high amount of single domain losses observed in this clade, while eudicots possess by average 5-6 copies (5.79). From a functional perspective there is evidence that gene loss plays a particularly important role in fungi. In fungi, massive gene loss as a major evolutionary mechanism has been linked to biotrophy to discard dispensable genomic components [50] and to adaptations to new hosts [51]. In addition to some biotrophic species in our fungi dataset, such as *Puccinia graminis* [52] or *Ustilago maydis* [53], there are other species for which host adaptations or biotrophy cannot be the explanation for large-scale gene loss, since they are not biotrophic, like *Saccharomyces cerevisiae*. However, for *Saccharomyces* species there is evidence for an ancient whole genome duplication event followed by massive gene loss (an estimated 85%) of the duplicated genes [54]. Next to whole genome duplication, other studies also linked polyploidy in fungi and plants to high loss rates [55]. In contrast to plants, where whole genome duplication events appear to lead to a high copy number of domains, fungi seem to possess mechanisms to rapidly reduce their genome size and throw out redundant or unnecessary information. The examples suggest that the unusually high rate of single domain losses observed in the fungi clade are the result of a fungi-specific evolutionary mechanism of genome evolution involving gene loss as a major driving force. In conclusion, next to genomic properties such as the abundance of duplicates as a basis for subsequent changes other factors likely play important roles for the evolutionary distribution of certain rearrangement events. These factors can be as described differences in lifestyles, but also differences in reproduction patterns are potential candidates, as the presence/absence of sexual reproduction in many plant and fungal species can provide an explanation for the observed differences in these clades.

Conclusions

Robustness of results and methodological limitations

Overall, this study shows that only six different basic event types are sufficient to explain the majority of new domain arrangements contributing to the complex process of protein innovation in major phylogenetic clades. The results are highly consistent across all major clades, i.e. similar proportions of arrangements can be explained by the same events across all clades, suggesting that misannotations do not bias the outcome significantly and the findings can be considered to be universally valid across eukaryotes. Furthermore, the similar distribution of events in insects and eudicots, representing 50% and 70% uniquely resolved events in the corresponding clade, suggests that unresolved events in all clades are likely a matter of resolution of the tree and not changing the distribution of events observed in this study. Additionally, the results of the conducted jackknife test (see Additional file 4) make sampling biases unlikely.

However, this study focuses on phenotypic changes through mutational events, which are observable solely on a domain level. Many of the investigated event types can be caused by different molecular mechanisms on the DNA level, which rates can vary compared to each other and be influenced by lifestyles or reproduction patterns. For a more complete picture of the evolutionary history, domain-based methods such as the here presented one, should be therefore complemented with primary sequence-based methods to answer specific biological questions.

Future implications and perspectives

Domain-based approaches have some special properties compared to primary sequence-based ones, making them particularly suited for different types of analyses. A general difference of domain-based approaches is the use of a larger alphabet with fewer letters per sequence. Additionally, changes on the domain level are less frequent than mutations of amino acids or nucleotides, why domains are especially suited for long time scales. The high conservation of domains and a high sensitivity in detection via their underlying Hidden Markov Models enable the accurate detection of homologous sequence fragments even in highly diverged sequences. Therefore, domain-based approaches avoid problems of primary sequence-based methods as in homology detection. Also, for phylogenetic analyses there are certain advantages such as reduced biases through saturation or long branch attraction.

Still, multiple parameters and properties for domain rearrangements are unknown, limiting the possibilities for practical implementations of domain-based approaches. Unfortunately, no general rates and transition probabilities for domain rearrangement events were known before this study that could be applied to diverse and bigger data sets. Also time depths for all phylogenies and branches are not resolvable by now. Despite these limitations, the parsimony approach used in this study can map the changes across different speciation events in the tree and shows

no significant bias introduced by the method. In fact, as demonstrated in this study, domain rearrangement rates hardly depend on depths of single nodes in the phylogenetic tree, suggesting the here used parsimony approach seems to be accurate and resulting in feasible and substantiated basic rearrangement rates. In a next step these estimated rates can lay the foundation for more advanced domain-based methods, while this further step cannot be provided by this study on its own already. It should be noticed that the here estimated rates and frequency of events are the raw descriptive numbers to provide an unbiased data basis, but for advanced methods these should be carefully normalised dependent on the scope of application. The available number of proteins in a proteome as well as the frequency of duplication events and therefore active mobile elements in a genome are for example influencing factors for domain rearrangements and should be taken into account. Additionally, emergence and loss events in this study are seen from a functional perspective and the presence or absence of an arrangement in the protein repertoire is of main interest, while we do not consider expansions or contractions of the same arrangement through copy number.

Summarising, this study is meant to elucidate the dynamics of domain rearrangements in different taxonomic groups and by doing so providing a data basis for more advanced methods. Analyses from a domain point of view could complement other methods and make it easier to estimate biases of other studies or overcome certain limitations. In conclusion, the results of this study demonstrate the high potential of domain-based approaches, while at the same time providing a basis for further development in this field.

Methods

Data set preparation

Five data sets are analysed in this study, each representing a different phylogenetic clade: vertebrates (61 species), insects (72), fungi (36) monocots (19) and eudicots (14) (see Additional file 11). Only proteomes are included that have a DOGMA [56] quality score $\geq$75%, to ensure that all proteomes used are of high and similar quality. This prevents the calculation of unduly high number of rearrangement events due to poor genome and gene prediction quality. To assure better comparability between the clades and the species within a clade, the corresponding ensembl database [57] as a widely used source for comparative genomics, was screened primarily for proteomes when available (fungi, plants(eudicots and monocots) and vertebrates).

As outgroups, a set of five well-annotated species (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens* and *Saccharomyces cerevisiae*) is chosen. For each clade members of the clade itself are not used as outgroups, for example *Drosophila melanogaster* is not used as an outgroup for the insects. *Strigamia maritima* is additionally added as outgroup for the insect clade to make sure insect specific rearrangements are studied and not general arthropod rearrangements. In a first step all but the longest isoform of each gene is removed from the data set to prevent a bias in event rate detection by their influence on the analysis. Proteomes are annotated with *Pfam* domain models [58] (version 30) using the pfam_scan.pl script (version 1.5) provided by *Pfam*. We used default parameters so that the script applies the thresholds specified in the Pfam database for annotating and filtering of the domains. Consecutive domain repeats in arrangements are collapsed to one instance of the domain (A-B-B-B-C $\rightarrow$ A-B-C), as it has been shown that even between closely related species copy number of repeated domains can vary a lot [59] and also to avoid miscalculations due to split domains caused by annotation/gene model errors.

The phylogenetic tree for the vertebrate clade is taken from ensembl [57]. The fungi tree is built using *NCBI Taxonomy database* [60] and *Superfamily* [2] as basis and resolving unknown branches from literature [61, 62]. The insect tree is built according to the *NCBI Taxonomy database*, while multifurcating branches of the genera Papilio, Apis, Bombus and Dufourea are transformed to bifurcating solutions according to literature [63–66]. Plant phylogenies are initially inferred using *NCBI Taxonomy* and refined using literature [67–69]. Next to the quality criterion mentioned above the resolvability of the phylogenetic relationship to other species was the second crucial criterion for the sampling process. The effect of subsampling replicates on the analysis is discussed based on a jackknife test.

Reconstruction of ancestral domain arrangements

The reconstruction of ancestral domain arrangements and calculation of rates of domain rearrangement events is carried out using the in-house developed program 'DomRates' (http://domainworld.uni-muenster.de/programs/domrates/).

Reconstruction of ancestral states of domains and domain arrangements is based on a parsimony principle. While single domain presence/absence states are usually better modelled by a Dollo parsimony, multi-domain arrangements with their modular nature are better modelled by a Fitch parsimony. The assumption underlying the use of Dollo parsimony is that novel domains are gained only once [16], while arrangements can be formed and broken several times. For this reason, 'DomRates' reconstructs the ancestral states of the whole tree twice: First with Fitch parsimony for all domain arrangements (including single domain arrangements) and a second time with Dollo parsimony for all single domains

included in any arrangement (see Fig. 3). The inferred single domain states with Dollo parsimony are used to verify all terminal emergence events and single domain loss/emergence events found by the Fitch parsimony reconstruction.

The copy number of certain domain arrangements is not considered in DomRates, which means only the presence/absence of a given arrangement is reconstructed and taken into account, but not the number of appearances in the proteome. This means that emergence and loss are seen from a functional perspective in this study based on if an arrangement is available in the functional repertoire of a proteome. Expansions and contractions of the same arrangement regarding the numbers of its copies are not described as emergence or loss.

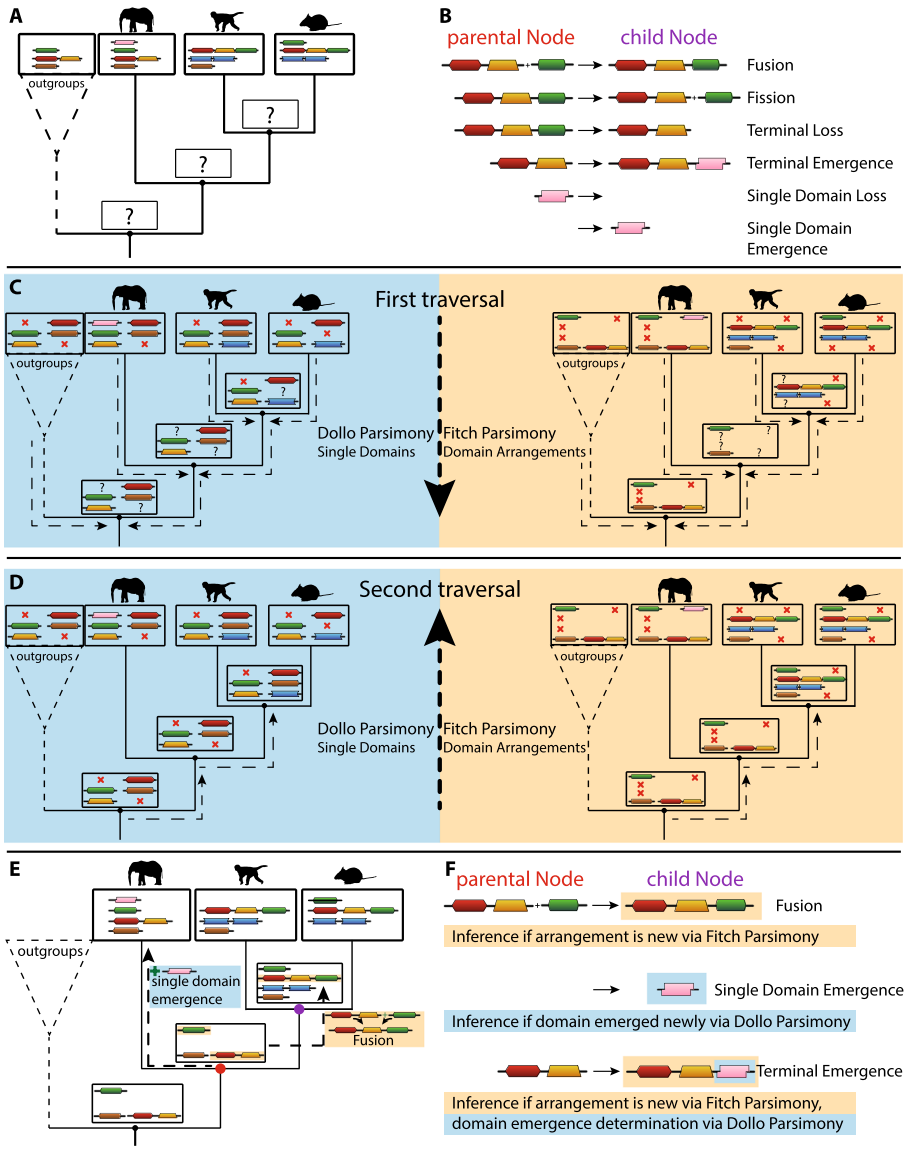

Fig. 3 Reconstruction of ancestral domain content and rearrangement events. Given a known phylogeny and domain annotations of all included species (**a**) , it becomes possible to infer six event types leading to new domain contents over time (**b**). First, the ancestral domain content of all inner nodes is inferred by two different parsimony approaches: for all single domains using a Dollo parsimony approach (light blue background), and for all arrangements, using a Fitch parsimony approach (light orange background). In a first traversal from the leaves to the root of the tree, all inner node states are annotated as present, absent or unknown according to the regarding parsimony rules (**c**) (see Additional file 1). In a second traversal from the root to the leaves, the unknown states at the root are first resolved according to the parsimony rules (see Additional file 1) and subsequently all following unknown states set to the parental state (**d**). In the reconstructed tree it becomes possible to infer the different event types at any node by comparison with the parental node (**e**). In this way emergences/losses of domains are inferred from the Dollo tree, while arrangements are inferred from the Fitch tree (**f**)

Terms and definitions - event and solution types

Since previous research in the field of protein domains focused mainly either on emergence and loss of single domains or on the evolutionary history of whole arrangements, sometimes postulating concepts such as recombination or domain-shuffling, it is necessary to specify the rearrangement events considered in this study (see Fig. 3b). In fact, just four biological events can explain the formation of virtually all domain arrangements: *fusion* of existing (ancestral) arrangements (also of single domain proteins which amounts to gene fusion), *fission* of existing (ancestral) domain arrangements, *loss* of one or more domains (i.e. there are no traces left as the underlying DNA sequence is for example no longer transcribed) and *emergence* of one domain. The latter two biological events of *loss* and *emergence* can be divided into two different conceptual ones each. We distinguish in our study terminal loss/emergence and single domain loss/emergence, which can be both explained by the underlying mechanisms for loss and emergence. Terminal events describe the loss or emergence of domains at the ends of arrangements, while single domain events describe the complete loss or the first emergence of a single domain as a discrete arrangement. Terminal loss allows for more than one domain to be lost in contrast to just one domain considered for terminal emergence, since terminal loss can easily be caused by an introduced stop codon, which affects dependent on the position all following domains in the protein and not just the next or last domain. With this conceptual differentiation we make it possible to combine the two different approaches of previous studies (loss and emergence of single domains vs. reshuffling of domain arrangements).

It is important to note that all mutational events described here are defined purely on a domain level. On a DNA level different molecular mechanisms and mutations can lead to the same mutational event described here (e.g. fusion of two arrangements by fusion of neighboring genes through stop codon loss or through transposition of a second gene through mobile elements). For this reason we just define events we can infer explicitly on a domain level, while other potential molecular mechanisms leading to additional (less common) mutational events are not considered. An example for this would be the insertion of a domain/arrangement in the middle of an existing domain arrangement, which can happen through crossing over or transposition through smaller mobile elements, but cannot be distinguished on a domain level between insertion in the middle of an arrangement or two subsequent fusion events of independent arrangements. The possibility of multi step events or multiple possible solutions makes the definition of different solution types necessary.

One can differentiate between four different solution types (see Additional file 2): exact solution, non-ambiguous solution, and ambiguous solution can all be explained by one instance of the single step event types above, while a complex solution can only be explained by a chain of the above mentioned events. Exact solutions represent new arrangements that can be explained by a single event and just this one solution exists. In contrast, non-ambiguous solutions describe the case that a new arrangement can just be explained by one out of several single events, all of the same type. Ambiguous solutions involve more than one event type as a possible explanation for a new arrangement. If there does not exist a solution in a single step, it is defined as a complex solution.

Domain rearrangement rates calculation

For the rate determination only exact and non-ambiguous solutions are considered, ambiguous and complex solutions are ignored. To avoid bias introduced by outgroup-specific arrangements, we exclude the nodes of the outgroup, the root of the complete tree and the root of each clade (first node after root) from the rate calculation. A jackknife test with 100 repetitions is carried out by randomly removing 3 species from every clade and rerunning DomRates on the altered phylogeny to ensure robustness of the found rates and to identify possible sampling biases within clades. Means and standard deviation for every event type frequency in the jackknife test are shown in Additional file 4.

Enriched gene ontology terms

A Gene Ontology (GO) term enrichment is carried out with *topGO* package [70] in R. The GO universe is composed of all domain arrangements that are present in all species in a clade as well as the reconstructed domain arrangements set in the ancestral nodes. Domains in new domain arrangements that can be explained by an exact or non-ambiguous solution are annotated with the 'pfam2go' mapping of *Pfam* domains to GO terms [71]. The enrichment analysis is done using the ontologies of 'Molecular function' and 'Biological process' and *topGO*s 'weight01' algorithm. Significantly enriched (P-value$\leq$0.05) GO terms are visualized as tag clouds.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12862-020-1591-0.

Additional file 1: Rules of inference for both parsimony approaches. The middle panel shows which two parental states (present, absent or unknown) for a domain or arrangement lead to which inference in the child node according to Dollo parsimony (left) or Fitch parsimony (right). The last line shows to what state an unknown state at the root is resolved.

Additional file 2: Solution types. There are four different solution types by which a new arrangement can be explained. Exact and non-ambiguous solutions involve each just one event type (see Fig. 3b) and are called unique solutions. Ambiguous and complex solutions cannot be explained by a single event type and are called manifold solutions. Just unique solutions are considered for the rate calculation in this study.

Additional file 3: Total number of events per solution type for all five clades.

Additional file 4: Jackknife test. Mean and standard deviation for all event type frequencies of a jackknife test with 100 replicates. For the jackknife test 3 species per clade were randomly removed and the resulting phylogeny tested with DomRates (100 repetitions).

Additional file 5: Number of rearrangement events across the insect phylogeny. Digit representation of the total number of rearrangement events at a specific node is indicated next to the pie chart. For details on 'Outgroups' see Methods. Significant GO terms in gained domain arrangements are shown in a tag cloud (box). GO terms that might point to insect specific evolution are: chitin metabolic process, sensory perception of taste.

Additional file 6: Number of rearrangement events across the vertebrate phylogeny. Digit representation of the total number of rearrangement events at a specific node is indicated next to the pie chart. For details on 'Outgroups' see Methods. Significant GO terms in gained domain arrangements are shown in a tag cloud (box). GO terms related to vertebrate evolution are strongly associated with regulation and signal transduction.

Additional file 7: Number of rearrangement events across the fungi phylogeny. Digit representation of the total number of rearrangement events at a specific node is indicated next to the pie chart. For details on 'Outgroups' see Methods. Significant GO terms in gained domain arrangements are shown in a tag cloud (box).

Additional file 8: Number of rearrangement events across the monocot phylogeny. Digit representation of the total number of rearrangement events at a specific node is indicated next to the pie chart. For details on 'Outgroups' see Methods. Significant GO terms in gained domain arrangements are shown in a tag cloud (box). GO terms that might point to monocot specific evolution are: 'recognition of pollen'.

Additional file 9: Domain arrangement sizes. The size represents the number of domains an arrangement consists of, while the fraction relates to all discriminative domain arrangements in total for the specific clade. The total number of different arrangements considered in the data sets was 22199 (vertebrates), 22346 (insects), 10030 (fungi), 15565 (monocots) and 12097 (eudicots).

Additional file 10: GO term enrichment analysis. Tag cloud for all events at the root of mammals in the vertebrate tree.

Additional file 11: List of all species included in this study. Furthermore, for each species the related DOGMA completeness score of their proteome and version of the used genome assembly is shown.

Abbreviations
EGF: Epidermal growth factor; GO: Gene Ontology; IIS: Insulin and insulin-like signalling; KRTAP: Keratin-associated protein; KRTDAP: Keratinocyte differentiation-associated protein

Acknowledgements
We thank Stephen Richards, Richard Gibbs and the i5K pilot initiative for allowing us to include pre-publication data into our insect data set. For proof reading and linguistic improvement we want to thank Brennen Heames and Daniel Dowling.

Authors' contributions
ED and SK wrote the original draft and performed the formal analysis. ED developed the software. EBB designed the study and acquired funding. ED, SK, EBB, SP and CK interpreted the data and revised the manuscript. CK was responsible for project coordination. All authors approved the final version of the manuscript.

Funding
ED received support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 281125614 / GRK2220. SK was funded by the Leibniz Graduate School on Genomic Biodiversity Research, grant number "SAW-2013-ZFMK-4" to EBB. The funding sources played no role in the design of the study, collection, analysis, and interpretation of data or in writing the manuscript.

Availability of data and materials
The developed software *DomRates* is available via the project homepage http://domainworld.uni-muenster.de/programs/domrates/ and with full source code on our gitlab https://ebbgit.uni-muenster.de/domainWorld/DomRates/. Furthermore, an archived version of the source code with all data related to this study (DomRates result files, GOterm analysis, jackknife test) is available via https://doi.org/10.5281/zenodo.2630419. DomRates is implemented in C++ and platform independent, published under the free GNU GPL licence version 3. For further information please check the UserManual on the website, where you can find also tutorials and example data sets to test the software.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Institute for Evolution and Biodiversity, University of Münster, Hüfferstrasse 1, 48149 Münster, Germany . [2]Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, August-Schmidt-Ring 10, 45665 Recklinghausen, Germany.

Received: 12 April 2019 Accepted: 31 January 2020
Published online: 14 February 2020

References
1. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1):427–32.
2. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J. SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. 2009;37(Database issue):380–6.
3. Forslund K, Sonnhammer ELL. Evolution of protein domain architectures. In: Anisimova M, editor. Evolutionary Genomics: Statistical and Computational Methods, Volume 2. Totowa, NJ: Humana Press; 2012. p. 187–216. https://doi.org/10.1007/978-1-61779-585-5_8.
4. Levitt M. Nature of the protein universe. Proc Natl Acad Sci USA. 2009;106(27):11079–84.
5. Apic G, Gough J, Teichmann Sa. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol. 2001;310(2):311–25. https://doi.org/10.1006/jmbi.2001.4776.
6. Ekman D, Björklund Å, Frey-Skött J, Elofsson A. Multi-domain proteins in the three kingdoms of life: Orphan domains and other unassigned regions. J Mol Biol. 2005;348(1):231–43. https://doi.org/10.1016/j.jmb.2005.02.007.
7. Yang X, Jawdy S, Tschaplinski TJ, Tuskan GA. Genome-wide identification of lineage-specific genes in Arabidopsis, Oryza and Populus. Genomics. 2009;93(5):473–80. https://doi.org/10.1016/j.ygeno.2009.01.002.
8. Kummerfeld SK, Teichmann SA. Protein domain organisation: adding order. BMC Bioinformatics. 2009;10:. https://doi.org/10.1186/1471-2105-10-39.
9. Zmasek CM, Godzik A. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol. 2011;12(1):4. https://doi.org/10.1186/gb-2011-12-1-r4.

10. Cromar G, Wong K-C, Loughran N, On T, Song H, Xiong X, Zhang Z, Parkinson J. New Tricks for "Old" Domains: How Novel Architectures and Promiscuous Hubs Contributed to the Organization and Evolution of the ECM. Genome Biol Evol. 2014;6(10):2897–917. https://doi.org/10.1093/gbe/evu228.
11. Patthy L. Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. Cell. 1985;41(3):657–63.
12. Pawson T. Protein modules and signalling networks. Nature. 1995;373(6515):573–80.
13. Sardar AJ, Oates ME, Fang H, Forrest AR, Kawaji H, Gough J, Rackham OJ. The evolution of human cells in terms of protein innovation. Mol Biol Evol. 2014;31(6):1364–74.
14. Lees JG, Dawson NL, Sillitoe I, Orengo CA. Functional innovation from changes in protein domains and their combinations. Curr Opin Struct Biol. 2016;38(Supplement C):44–52. https://doi.org/10.1016/j.sbi.2016.05.016. New constructs and expression of proteins ● Sequences and topology.
15. Weiner J, Beaussart F, Bornberg-Bauer E. Domain deletions and substitutions in the modular protein evolution. FEBS J. 2006;273(9):2037–47. https://doi.org/10.1111/j.1742-4658.2006.05220.x.
16. Moore AD, Bornberg-Bauer E. The dynamics and evolutionary potential of domain loss and emergence. Mol Biol Evol. 2012;29(2):787–96. https://doi.org/10.1093/molbev/msr250.
17. Björklund ÅK, Light S, Sagit R, Elofsson A. Nebulin: A Study of Protein Repeat Evolution. J Mol Biol. 2010;402(1):38–51. https://doi.org/10.1016/j.jmb.2010.07.011.
18. Schüler A, Bornberg-Bauer E. Evolution of Protein Domain Repeats in Metazoa,. Mol Biol Evol. 2016;33(12):3170–82. https://doi.org/10.1093/molbev/msw194.
19. Kersting AR, Bornberg-Bauer E, Moore AD, Grath S. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. Genome Biol Evol. 2012;4(3):316–29. https://doi.org/10.1093/gbe/evs004.
20. Moore AD, Grath S, Schüler A, Huylmans AK, Bornberg-Bauer E. Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree. Biochim Biophys Acta. 2013;1834(5):898–907. https://doi.org/10.1016/j.bbapap.2013.01.007.
21. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
22. Wu D-D, Irwin DM, Zhang Y-P. Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair. BMC Evol Biol. 2008;8(1):241. https://doi.org/10.1186/1471-2148-8-241.
23. Kuhn F, Lassing C, Range A, Mueller M, Hunziker T, Ziemiecki A, Andres A-C. Pmg-1 and Pmg-2 constitute a novel family of KAP genes differentially expressed during skin and mammary gland development. Mech Dev. 1999;86(1–2):193–6. https://doi.org/10.1016/S0925-4773(99)00115-X.
24. McGrath JA, Eady RAJ, Pope FM. Anatomy and organization of human skin. In: Rook's Textbook of Dermatology, Chap. 3. Wiley; 2008. p. 45–128. https://doi.org/10.1002/9780470750520.ch3.
25. Oomizu S, Sahuc F, Asahina K, Inamatsu M, Matsuzaki T, Sasaki M, Obara M, Yoshizato K. Kdap, a novel gene associated with the stratification of the epithelium. Gene. 2000;256(1–2):19–27. https://doi.org/10.1016/S0378-1119(00)00357-7.
26. Brocker C, Thompson D, Matsumoto A, Nebert DW, Vasiliou V. Evolutionary divergence and functions of the human interleukin (IL) gene family. Human Genom. 2010;5(1):30. https://doi.org/10.1186/1479-7364-5-1-30.
27. Zhu M, Janssen E, Leung K, Zhang W. Molecular Cloning of a Novel Gene Encoding a Membrane-associated Adaptor Protein (LAX) in Lymphocyte Signaling. J Biol Chem. 2002;277(48):46151–58. https://doi.org/10.1074/jbc.M208946200.
28. Liu ZH, Yang CP, Qi XT, Xiu LL, Wang YC. Cloning, heterologous expression, and functional characterization of a chitinase gene, Lbchi32, from Limonium bicolor. Biochem Genet. 2010;48(7-8):669–79. https://doi.org/10.1007/s10528-010-9348-x.
29. Punja ZK, Zhang YY. Plant Chitinases and Their Roles in Resistance To Fungal Diseases. J Nematol. 1993;25(4):526–40. https://doi.org/10.5943/mycosphere/3/4/14.
30. Singh A, Kirubakaran SI, Sakthivel N. Heterologous expression of new antifungal chitinase from wheat. Protein Expr Purif. 2007;56(1):100–9. https://doi.org/10.1016/j.pep.2007.06.013.
31. Afzal AJ, Wood AJ, Lightfoot DA. Plant receptor-like serine threonine kinases: roles in signaling and plant defense. Mol Plant-microbe Interact MPMI. 2008;21(5):507–17. https://doi.org/10.1094/MPMI-21-5-0507.
32. Krattinger SG, Lagudah ES, Spielmeyer W, Singh RP, Huerta-espino J, McFadden H, Bossolini E, Selter LL, Keller B. Pathogens in Wheat. Science. 2009;323(MARCH):1360–63.
33. Kapheim KM, Pan H, Li C, Salzberg SL, Puiu D, Magoc T, Robertson HM, Hudson ME, Venkat A. Social evolution. Genomic signatures of evolutionary transitions from solitary to group living. Science. 2015;348(6239):1139–44. https://doi.org/10.1126/science.aaa4788.
34. Wheeler DE, Buck N, Evans JD. Expression of insulin pathway genes during the period of caste determination in the honey bee, Apis mellifera. Insect Mol Biol. 2006;15(5):597–602. https://doi.org/10.1111/j.1365-2583.2006.00681.x.
35. de Azevedo SV, Hartfelder K. The insulin signaling pathway in honey bee (Apis mellifera) caste development - differential expression of insulin-like peptides and insulin receptors in queen and worker larvae. J Insect Physiol. 2008;54(6):1064–71. https://doi.org/10.1016/j.jinsphys.2008.04.009.
36. Mott CM, Breed MD. Insulin modifies honeybee worker behavior. Insects. 2012;3(4):1084–92. https://doi.org/10.3390/insects3041084.
37. Formesyn EM, Cardoen D, Ernst UR, Danneels EL, Van Vaerenbergh M, De Koker D, Verleyen P, Wenseleers T, Schoofs L, de Graaf DC. Reproduction of honeybee workers is regulated by epidermal growth factor receptor signaling. Gen Comp Endocrinol. 2014;197:1–4. https://doi.org/10.1016/j.ygcen.2013.12.001.
38. Kamakura M. Royalactin induces queen differentiation in honeybees. Nature. 2011;473(7348):478–83.
39. Barchuk AR, Cristino AS, Kucharski R, Costa LF, Simões ZLP, Maleszka R. Molecular determinants of caste differentiation in the highly eusocial honeybee Apis mellifera. BMC Dev Biol. 2007;7(1):70. https://doi.org/10.1186/1471-213X-7-70.
40. Elias-Neto M, Nascimento ALO, Bonetti AM, Nascimento FS, Mateus S, Garófalo CA, Bitondi MMG. Heterochrony of cuticular differentiation in eusocial corbiculate bees. Apidologie. 2014;45(4):397–408. https://doi.org/10.1007/s13592-013-0254-1.
41. Kummerfeld SK, Teichmann SA. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet. 2005;21(1):25–30.
42. Harrison MC, Jongepier E, Robertson HM, Arning N, Bitard-Feildel T, Chao H, Childers CP, Dinh H, Doddapaneni H, Dugan S, Gowin J, Greiner C, Han Y, Hu H, Hughes DST, Huylmans AK, Kemena C, Kremer LPM, Lee SL, Lopez-Ezquerra A, Mallet L, Monroy-Kuhn JM, Moser A, Murali SC, Muzny DM, Otani S, Piulachs MD, Poelchau M, Qu J, Schaub F, Wada-Katsumata A, Worley KC, Xie Q, Ylla G, Poulsen M, Gibbs RA, Schal C, Richards S, Belles X, Korb J, Bornberg-Bauer E. Hemimetabolous genomes reveal molecular basis of termite eusociality. Nat Ecol Evol. 2018;2(3):557–66.
43. Zhou X, Rokas A, Berger SL, Liebig J, Ray A, Zwiebel LJ. Chemoreceptor Evolution in Hymenoptera and Its Implications for the Evolution of Eusociality. Genome Biol Evol. 2015;7(8):2407–16.
44. Helmkampf M, Cash E, Gadau J. Evolution of the insect desaturase gene family with an emphasis on social Hymenoptera. Mol Biol Evol. 2015;32(2):456–71.
45. Panchy N, Lehti-Shiu M, Shiu S-H. Evolution of gene duplication in plants. Plant Physiology. 2016;171(4):2294–316. https://doi.org/10.1104/pp.16.00523.
46. Lisch D. How important are transposons for plant evolution?. Nat Rev Genet. 2013;14(1):49–61.
47. Soltis PS, Marchant DB, de Peer YV, Soltis DE. Polyploidy and genome evolution in plants. Curr Opin Genet Dev. 2015;35:119–25. https://doi.org/10.1016/j.gde.2015.11.003.
48. Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. Polyploidy and angiosperm diversification. Am J Bot. 2009;96(1):336–48. https://doi.org/10.3732/ajb.0800079.
49. Reineke AR, Bornberg-Bauer E, Gu J. Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. Nucleic Acids Res. 2011;39(14):6029–43.

50. Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, Loren van Themaat EV, Brown JKM, Butcher SA, Gurr SJ, Lebrun M-H, Ridout CJ, Schulze-Lefert P, Talbot NJ, Ahmadinejad N, Ametz C, Barton GR, Benjdia M, Bidzinski P, Bindschedler LV, Both M, Brewer MT, Cadle-Davidson L, Cadle-Davidson MM, Collemare J, Cramer R, Frenkel O, Godfrey D, Harriman J, Hoede C, King BC, Klages S, Kleemann J, Knoll D, Koti PS, Kreplak J, López-Ruiz FJ, Lu X, Maekawa T, Mahanil S, Micali C, Milgroom MG, Montana G, Noir S, O'Connell RJ, Oberhaensli S, Parlange F, Pedersen C, Quesneville H, Reinhardt R, Rott M, Sacristán S, Schmidt SM, Schön M, Skamnioti P, Sommer H, Stephens A, Takahara H, Thordal-Christensen H, Vigouroux M, Weßling R, Wicker T, Panstruga R. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Science. 2010;330(6010):1543–6. https://doi.org/10.1126/science.1194573.
51. Sharma R, Mishra B, Runge F, Thines M. Gene loss rather than gene gain is associated with a host jump from monocots to dicots in the smut fungus melanopsichium pennsylvanicum. Genome Biol Evol. 2014;6(8): 2034–49. https://doi.org/10.1093/gbe/evu148.
52. Kämper J, Kahmann R, Bölker M, Ma L-J, Brefort T, Saville BJ, Banuett F, Kronstad JW, Gold SE, Müller O, et al. Insights from the genome of the biotrophic fungal plant pathogen ustilago maydis. Nature. 2006;444(7115):97.
53. Duplessis S, Cuomo CA, Lin Y-C, Aerts A, Tisserant E, Veneault-Fourrey C, Joly DL, Hacquard S, Amselem J, Cantarel BL, Chiu R, Coutinho PM, Feau N, Field M, Frey P, Gelhaye E, Goldberg J, Grabherr MG, Kodira CD, Kohler A, Kües U, Lindquist EA, Lucas SM, Mago R, Mauceli E, Morin E, Murat C, Pangilinan JL, Park R, Pearson M, Quesneville H, Rouhier N, Sakthikumar S, Salamov AA, Schmutz J, Selles B, Shapiro H, Tanguay P, Tuskan GA, Henrissat B, Van de Peer Y, Rouzé P, Ellis JG, Dodds PN, Schein JE, Zhong S, Hamelin RC, Grigoriev IV, Szabo LJ, Martin F. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proc Natl Acad Sci. 2011;108(22):9166–71. https://doi.org/10.1073/pnas.1019315108.
54. Cliften PF, Fulton RS, Wilson RK, Johnston M. After the duplication: gene loss and adaptation in saccharomyces genomes. Genetics. 2006;172(2): 863–72.
55. Albalat R, Cañestro C. Evolution by gene loss. Nat Rev Genet. 2016;17: 379–91.
56. Dohmen E, Kremer LPM, Bornberg-Bauer E, Kemena C. DOGMA: domain-based transcriptome and proteome quality assessment. Bioinformatics. 2016;32(17):2577. https://doi.org/10.1093/bioinformatics/btw231.
57. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, García Girón C, Hourlier T, Howe K, Kähäri A, Kokocinski F, Martin FJ, Murphy DN, Nag R, Ruffier M, Schuster M, Tang YA, Vogel J-H, White S, Zadissa A, Flicek P, Searle SMJ. The Ensembl gene annotation system. Database. 2016;2016:093. https://doi.org/10.1093/database/baw093.
58. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44(D1):279–85. https://doi.org/10.1093/nar/gkv1344.
59. Ekman D, Björklund AK, Elofsson A. Quantification of the elevated rate of domain rearrangements in metazoa. J Mol Biol. 2007;372(5):1337–48. https://doi.org/10.1016/j.jmb.2007.06.022.
60. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2009;37(Database issue):5–15.
61. Helston RM, Box JA, Tang W, Baumann P. Schizosaccharomyces cryophilus sp. nov., a new species of fission yeast. FEMS Yeast Res. 2010;10(6):779–86. https://doi.org/10.1111/j.1567-1364.2010.00657.x.
62. Ebersberger I, de Matos Simoes R, Kupczok A, Gube M, Kothe E, Voigt K, von Haeseler A. A consistent phylogenetic backbone for the fungi. Mol Biol Evol. 2012;29(5):1319–34. https://doi.org/10.1093/molbev/msr285.
63. Lo N, Gloag RS, Anderson DL, Oldroyd BP. A molecular phylogeny of the genus Apis suggests that the Giant Honey Bee of the Philippines, A. breviligula Maa, and the Plains Honey Bee of southern India, A. indica Fabricius, are valid species. Syst Entomol. 2010;35(2):226–33. https://doi.org/10.1111/j.1365-3113.2009.00504.x.
64. Rehan SM, Glastad KM, Lawson SP, Hunt BG. The Genome and Methylome of a Subsocial Small Carpenter Bee, Ceratina calcarata. Genome Biol Evol. 2016;8(5):1401. https://doi.org/10.1093/gbe/evw079.
65. Zakharov EV, Caterino MS, Sperling FAH, Schultz T. Molecular Phylogeny, Historical Biogeography, and Divergence Time Estimates for Swallowtail Butterflies of the Genus Papilio (Lepidoptera: Papilionidae). Syst Biol. 2004;53(2):193. https://doi.org/10.1080/10635150490423403.
66. Misof B, Liu S, Meusemann K, Peters RS, Al E. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346(6210): 763–67. https://doi.org/10.1017/CBO9781107415324.004.
67. Hatje K, Kollmar M. A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. Front Plant Sci. 2012;3:192.
68. Lei W, Ni D, Wang Y, Shao J, Wang X, Yang D, Wang J, Chen H, Liu C. Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of *Astragalus membranaceus*. Sci Rep. 2016;6: 21669.
69. The Angiosperm Phylogeny Group. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc. 2009;161(2):105–21. https://doi.org/10.1111/j.1095-8339.2009.00996.x.
70. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics (Oxford, England). 2006;22(13):1600–7. https://doi.org/10.1093/bioinformatics/btl140.
71. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, Rato C, Yong SY, Bateman A, Punta M, Attwood TK, Sigrist CJA, Redaschi N, Rivoire C, Xenarios I, Kahn D, Guyot D, Bork P, Letunic I, Gough J, Oates M, Haft D, Huang H, Natale DA, Wu CH, Orengo C, Sillitoe I, Mi H, Thomas PD, Finn RD. The InterPro protein families database: The classification resource after 15 years. Nucleic Acids Res. 2015;43(D1):213–21. https://doi.org/10.1093/nar/gku1243.

Publisher's Note