EvoProDom: Evolutionary model of protein families by means of translocations of protein domains

Gon Carmi[1 †]

Alessandro Gorohovski [1 †]

Milana Frenkel-Morgenstern [1,*]

[1]Cancer Genomics and BioComputing of Complex Diseases Lab,

The Azrieli Faculty of Medicine, Bar-Ilan University,

8 Henrietta Szold St, Safed 13195, ISRAEL

*Corresponding Author:

e-mail: milana.morgenstern@biu.ac.il

[†] These authors contributed equally to this work.

Abstract

Here, we developed a novel "evolution of protein domains" (EvoProDom) model for evolution of proteins, which was based on "mix and merge" of protein domains. We collected and integrated genomic and proteome data for 109 organisms. These data include protein domain content and orthologous protein families. In EvoProDom, we defined evolutionary events, such as translocations, as reciprocal exchanges of protein domains between orthologous proteins of different organisms. We found that protein domains, which frequently appear in translocation events, were enriched in *trans-splicing* events, *i.e.*, producing novel transcripts fused from two distinct genes. We presented in EvoProDom, a general method to obtain protein domain content and orthologous protein annotation, by predicting these data from protein sequences using the Pfam search tool and KoFamKOALA, respectively. This method can be implemented in other research such as proteomics, protein design and host-virus interactions.

Keywords: protein evolution, protein domains, translocations.

Introduction

Proteins are built from a set of domains that correspond to conserved regions with distinct functional and structural characteristics (1). In accordance with the domain-oriented view of proteins, domains group together to form domain architectures (DAs), i.e., ordered sequences of domains. Some domains participate in specific DAs, while others are found in diverse DAs. This later property is termed “domain promiscuity” or “domain mobility”. Analyses of domain promiscuity can reveal the mechanisms by which domains are gained or lost (2). Marsh and Teichmann in 2010 (1) described five mechanisms by which proteins gain domains: (i) gene fusion, the fusion of a pair of adjacent genes via alternative splicing in non-coding intergenic regions; (ii) exon extension, the expansion of exon regions into adjacent introns, which encode a new domain; (iii) exon recombination, direct merging of two exons from two different genes; (iv) intron recombination or exon shuffling, insertion of an exon into an intron of a different gene; and (v) retroposition, a sequence located within one gene, which is transposed into a different gene along with a flanking genetic sequence. The mechanism that added a given domain to a protein can be identified from the properties of the gained domain, e.g., the position of the domain in a protein sequence and the number of exons. For example, multi-exon domain gain at the c-termini is due to gene fusion. Additionally, during metazoan evolution, new physical protein-protein interactions (PPIs) emerged consequent to exon shuffling of domains that mediated the interaction (3). An additional work, by Bornberg-Bauer and Mar Albà in 2013 (4), refined and expanded these mechanisms and introduced new concepts such as intrinsically disordered regions, implied links between the emergence of de-novo domains, and de-novo genes (4).

We developed a novel "evolution of protein domains" (EvoProDom) model for the evolution of proteins, which was based on "mix and merge" of protein domains. We collected and integrated genomic and proteome data for 109 organisms. These data included protein domain content and orthologous protein content. In EvoProDom, we defined evolutionary events, such as translocations, as the reciprocal exchanges of protein domains between orthologous proteins of different organisms. We found that protein domains, which frequently appear in translocation events, were enriched in *trans-splicing events*, *i.e.*, producing transcripts as a slippage of two distinct genes (5). We presented in EvoProDom, a general method to obtain protein domain content and orthologous protein annotation, by predicting these data from protein sequences using the Pfam search tool (6, 7) and KoFamKOALA (8), respectively. This method can be implemented in other research fields such as proteomics (9), protein design (10) and host-virus interactions (11).

Materials and Methods

The EvoProDom model was based on the data of full genome and annotated proteomes. In addition, the model utilized orthologous protein annotation and protein domain content. Orthologous protein groups were used to group proteins (Refseqs) from different organisms, thereby linking protein domain changes among orthologous proteins with their corresponding groups of organisms. Orthologous proteins were realized as Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs, namely, KEGG ortholog (KO) (12, 13). Protein domain content was realized as Pfam domains, and this content was associated with proteins. Accordingly, proteins were modeled as a group or list of

Pfam domains, and orthologous proteins were modeled as a group of proteins with the same KO number. Both Pfam domains and KO assignments to proteins were predicted from protein sequences alone, using the Pfam search tool (6, 7) and KoFamKOALA (8), respectively. By applying these protein sequence-based methods to obtain protein domain content and orthologous protein annotation, new organisms were easily added to EvoProDom. Moreover, this method is unique to EvoProDom and is applicable to other protein domain -based applications such as proteomics (9), protein design (10) and host-virus interactions (11).

Statistical analysis was performed using R (R: A language and environment for statistical, 3.3.2,2016).

Data resources

In total, the EvoProDom model was tested on a collection of 109 organisms with full genome and annotated proteomes (Entrez/NCBI (14)). These organisms were grouped according to six categories: (i) 15 fish; (ii) 4 subterranean, 8 fossorial and 21 aboveground animals (15, 16); (iii) 65 organisms with known PPIs (BioGrid version 3.5.173, (17, 18)); (vi) 17 organisms with HiC datasets; (v) 4 cats; and (iv) 15 pathogenic organisms (19). Organisms with HiC datasets were obtained by searching 'HiC' in the NCBI GEO database (**Table 1**).

Orthologous protein annotation

Orthologous annotation was based on Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs, namely, KEGG ortholog (KO) (12, 13). Proteins were assigned to KO groups, by utilizing KoFamKOALA, a Hidden Markov Model (HMM) profile based search tool (8). Since this tool assigned proteins to KO groups based on protein

sequences, in-house script was written to automatically assign proteins to KO groups. In this study, only proteins with KO annotation were used for analysis. Additionally, organism code was generated by selecting 3-4 letters from an organism name in uppercase format (lower case code represents organisms from the KEGG database) (**Table 1**).

Protein domain detection

Pfam (release 32.0) domains were predicted from protein sequences, using a dedicated HMM-based search tool (6, 7). Therefore, in-house script was written and protein domain content was predicted from protein sequences. Additionally, Pfam domains were members of super families (clan, pfam nomenclature), which we used for our classification. These data were supplemented to the protein domain content.

EvoProDomDB

Genomic and proteomic data, along with orthologous protein and protein domain content data, were related by shared data. Thus, a relational database, EvoProDomDB, was written in MySQL on MariaDB (10.0.26) for the efficient search engine. The EvoProDom model was implemented and tested on EvoProDomDB. EvoProDomDB was organized with orthologous protein and protein content for the 1,835,600 protein products that were distributed among 23,147 KO groups, containing 17,929 unique Pfam domains. Pfam domains were distributed among 629 super families, and EvoProDomDB integrated data for 109 organisms from diverse taxa. EvoProDomDB was built from six relational tables with common features, e.g., organism id and other features, as shown in the scheme (Figure 1). Relational tables, taxonomy, ko_annotation, clan_domain,

pfam_domain provided the annotation data. For example, taxonomy ranks, *e.g.*, species and genus; KO, domain and super family descriptions, respectively.

The main relational tables were included for protein genomic and proteomic data (org_protein_annotation), along with protein domain content (Pfam data). Standard genomic and proteomic data were included, for example, gene_symbol, chromosome, strand, refseq_id, a protein length, and a protein description. To these data, the KO number was added (ko_number). Genomic and proteomic data were linked by the longest isoform identification (isoform). Protein domain content included standard Pfam domains as retrieved from Pfam search tool output (6, 7), and calculated data that identified non-overlapping Pfam domains with maximal score (active) delimited by 'envfrom' and 'envto' coordinates. These corresponded to the largest region where a Pfam domain was predicted to be located within a protein sequence. Among multiple copies of active domains, the highest scoring domain was identified (unique active). To achieve these data, both standard and custom, as well as a construction of EvoProDomDB, in-house bash and perl scripts were written and combined to form a pipeline. The EvoProDom model was implemented as Perl with MySQL quires, to retrieve data from EvoProDomDB and bash scripts. These data sources and databases were summarized in the study workflow (Figure 2).

Results

The EvoProDom model

We hypothesized that proteins evolve by means of "mix and merge" or "shuffling" of protein domains, which are distinct functional units (1, 20). The evolutionary model that

described protein evolution by means of protein domain dynamics was termed EvoProDom. The EvoProDom model defined and formulated standard evolutionary mechanisms such as translocations, duplications and indel (insertion and deletion) events, which acted on protein domains that were realized as Pfam domains (6, 7). Therefore, proteins, under the EvoProDom model, gained or lost their function based on the presence or absence of respective function-conferring domains. Accordingly, proteins were modeled as sets of protein domains; and evolutionary events, such as translocations, were defined. These describe the gain and loss of particular domains between domain sets or DAs. The KEGG database catalogs diverse taxa and forms groups of orthologous proteins (KO) based on shared function. Thus, all members of the KO group were orthologous proteins (8, 12, 13). Additionally, in the EvoProDom model, proteins were assigned to KO groups (see Materials and Methods). Consequently, translocation events were mapped to groups of organisms according to underlying changes in DA. Thus, evolutionary events, which acted upon domains and manifested as changes in DA were reflected at the organism level. A link between changes at these two levels was therefore established. The EvoProDom model was implemented with and tested on EvoProDomDB (see Materials and Methods). In total, 5,548 translocation events, involving 94 protein super families, excluding an “unknown” super family, were found (*Table 2*). This result indicates the existence of multiple evolutionary translocation events as defined by the model.

Mapping of genes to proteins and alternative splicing

EvoProDom combined genomic information (genes) with proteins, and in turn, proteins with Pfam domain composition. In addition, proteins belonging to KEGG ortholog (KO)

groups were also included (8, 12, 13). Genes map to more than one mRNA transcript and, in turn, to more than one protein product, i.e., Refseq id. These correspond to isoforms of a gene, which result from alternative splicing, i.e., the inclusion of gene exons. Since, protein domains mostly coincide with exons (1, 3, 5, 20), changes in protein domain content can account for changes in DA as a result of translocation events. Therefore, to avoid confounding effects of alternative splicing, only a longest isoform was used (see Materials and Methods). Therefore, each gene was associated with a single protein product.

Protein domain content

Proteins are composed of protein domains, which overlap with other protein domains or which have multiple copies. Overlapping domains were undesired as they did not confirm to the linear structure of protein domains within proteins. To address this issue, a single domain (active domain) with a maximal score, as obtained from the Pfam search tool, was determined for each overlapping group of protein domains. This procedure ensured non-overlapping protein domain composition for each protein, consistent with the linear nature of protein domains along a protein sequence. However, multiple copies of such active domains were still present and for translocation events, a unique set of non-overlapping active domains was required. To this end, a similar procedure was applied to multiple copies of active domains, which were referred to as unique active domains. That is, a unique active domain was considered as an active domain with a maximal score among multiple copies of the same domain.

DA as a basic unit in EvoProDom

According to the EvoProDom model, evolutionary events, e.g., translocations, deletions and insertions, acted on protein domains and involved the DAs, orthologous groups, i.e., KO, and organisms. Hence, EvoProDomDB facilitated in organizing these data according to DA. Briefly, each orthologous group (KO) was partitioned into distinct sets (items), i.e., a list of domains (DA), and matched lists of organisms and proteins. Notably, within these matched lists, duplicated organisms correspond to paralogous proteins. Additionally, for each DA, missing and gained domains were determined from all DAs within a particular KO. From these data, mobile domains and translocation domains, *i.e.*, domains that had undergone all the translocation events, were determined. As a result, we found in total, 5,548 translocation events, involving 94 protein super families, excluding an "unknown" super family (*Table 2*). We identified 2,041 mobile domains of them, 259 had undergone translocation events and 1,782 were involved in indel events (Table *3*).

Evolutionary mechanism in EvoProDom

Implementation of DA

DA, uniquely identified by (ko,item) pair, data were generated from EvoProDomDB, while filtering for active and unique active domains (see Materials and Methods). Each DA included: (i) ko:item; (ii) Pfam domain list; (ii) a list of organisms (org_id); (iii) a list of refseq_ids; (iv) a list of missing domains; (v) and a list of gained domains. Importantly, the list of organisms (ii) and the list of refseq_ids (iii) were matched lists, i.e., the first refseq belonged to the first organism and the second refseq belonged to the second organism, *etc*. Moreover, all other DA information was shared by all organisms and corresponding refseqs, namely, all refseqs were members of the ko group and had similar domain content (item). Gained and lost patterns, (iv) and (v) above, were

computed for each KO group among all DAs, i.e., as items. Of note, a minimal number of DA, i.e, items, was two.

Formally, DA, the active domain and unique active domain were defined as follows:

Definition: Domain Architecture (DA)

Algorithm: Let $p_1, p_2, \dots, p_n \subseteq D$, were $D = \{d_1, d_2, \dots, d_m\}$ is a set of protein domains and p_i are DA. Grouping of DAs into distinct groups is a partition of $p_1, p_2, \dots, p_n$

Definition: Active domains and unique active domains

Assumptions: protein, $p = \{d_1, d_2, \dots d_m\}$, be DA, $c(d) \in \mathbb{R}$ be a score

Algorithm: domain $d \in p$ is an active domain if $c(d)$ is maximal among overlapping or nested domains.

A unique active domain is the highest scoring active domain among multiple copies of the same domain within p.

Translocation and indel events of a mobile domain.

Informally, translocations of mobile domains involve gain/loss from/to orthologous proteins from two KO groups, in which the mobile domains were determined by gain/loss patterns within a single KO group. Therefore, a mobile domain was described and defined formally. The main objective of the EvoProDom model was to reflect changes in domain content, namely, at the protein level, with the organism level; and that this would highlight groups of organisms with orthologous proteins, sharing similar patterns of protein domain gain/ loss. Protein domain composition was coupled with organisms by defining mobile and translocation domains. This was based on groups of organisms, and their sizes, with orthologous proteins sharing the same protein domain composition.

Protein domains were contained within orthologous proteins, or the domain missing from a protein, which was based on a number of organisms in each group, *i.e.*, orthologous proteins with and without a particular domain. The mobile domain was defined as follows:

Assumptions: Let $A,\ B,\ T$ be sets of organisms with proteins in KO group, k, such that $T = A \cup B,\ A \cap B = \emptyset, O \in A\,\{p \in O\ |d_x \in p\}$, $O \in B\,\{p \in O\ |d_x \notin p\}$. Organisms in A contain domain d_x whereas organisms in B lack domain d_x.

Algorithm: Unique active domain d_x is mobile between organisms in A and in B if $4 \leq |A| < |T| - 4$.

Next, translocations and indel events of mobile domains were defined. Translocations and indel events were mutually exclusive events. Translocation events of mobile domains expanded the patterns of the domain gain and/or domain loss in a single orthologous group to two orthologous groups, such that these patterns generated a reciprocal event. A reciprocal event was characterized by a mobile domain that was gained and lost in the first and second orthologous group, and vice versa (Figure 3). Similar to a mobile domain definition, translocation event criterion was defined for groups of organisms with four or more members. For example, a translocation event of the Pfam domain, PAS_11, is shown in Figure 3. In this translocation event, PAS_11 was present in the KEGG orthologous group number 09095, with correspondence to HIF2A. PAS_11 was absent from the orthologous protein group number 15589 (ARNT2). This gain and loss pattern of PAS_11 was observed among ten orthologous proteins from five different organisms (A^*), namely, DIO to ptr, such that each organism had a protein from HIF2A and one from the ARNT2 orthologous group. Reciprocal gain and loss patterns of PAS_11 were

observed in orthologous proteins from a second group of organisms (B^*), namely, ORO to gaa. A reciprocal gain and loss pattern meant that PAS_11 was absent from the HIF2A orthologous group and gained in the ARNT2 orthologous group. Since these reciprocal patterns of domain gain/loss, between two orthologous proteins, involved two groups of organisms (A^* and B^*) with more than four organisms each, a translocation event occurred for PAS_11. Note that an additional translocation event occurred for PAS_3 among same organisms and orthologous proteins (Figure 3).

Translocations and indel events were formally defined as follows:

Assumptions: Let d_x be a mobile domain between A_i and B_i in k_i, where $i = 1,2$, A_i, B_i are sets of organisms and k_i, are KO groups. Let $A^* = A_1 \cap B_2$ and $B^* = A_2 \cap B_1$.

Algorithm: Mobile domain d_x undergoes translocation if $|A^*|, |B^*| \geq 4$. Otherwise, the indel event has occurred.

Duplication of domains.

Translocation and indel events were defined for mobile domains; and mobile domains, in turn, were derived from unique active Pfam domains. For duplication events, active domains were used, in order to retain non-overlapping duplicates of Pfam domains (see Materials and Methods). These active domains were calculated for each orthologous protein group, i.e., KO group. Therefore, the duplicate status, “duplicated” or “non-duplicated”, was determined for a particular orthologous KO group and, consequently, varied among orthologous KO groups. Accordingly, the overall duplicate status of a particular Pfam domain was determined by a majority vote from individual KO groups on

duplicate status. For example, a Pfam domain's overall duplicate status was "duplicated", if the difference between the number of KOs with "duplicated" to the number of KOs with "non-duplicated" status was significant, i.e., 99% percentile of the cumulative sum of the differences. Similarly, overall "non-duplicated" status was determined when considering "non-duplicated" to "duplicated" differences. The duplicate status, namely, "duplicated" or "non-duplicated", for a particular KO group, was determined based on whether the copy number of particular domains varied or was constant across all DAs respectively. For example, if two copies of a particular domain appeared in one KO while three copies of a domain appeared in a second KO, the domain was determined "duplicated". However, if a particular domain had the same number of copies within all DA, e.g., two copies, then the domain was considered "non-duplicated". Duplication was formally defined as follows:

Assumptions: Let d_x be an active domain, ko be the KO group with $da_1, da_2, \ldots., da_m$ DAs of active domains. Then d_x is "non-duplicated" in ko, if the copy number of d_x is the same in each da, otherwise d_x is "duplicated".

Algorithm: d_x is duplicated if the difference between the number of KO groups where d_x is "duplicated" and the number of KO groups where it is "non-duplicated" is significant (above 99% of the cumulative sum of the differences). Similarly, a non-duplicated domain is defined.

Translocation domains were enriched in chimeric transcript

Previously, Frenkel-Morgenstern and Valencia 2012 (5), analyzed enrichment of the domain content for chimeric transcripts combined from two distinct genes. They found

that domains were enriched within chimeric transcripts that belonged to the following super families (Super family name): ANK (Ank), EFh (EF_hand), EGF-like (EFG), GTP_EFTU (P-loop_NTPase), IG-like (E-set), LRR (LRR), PH (zf-FYVE-PHD), Pkinase (PKinase), RING (RING), RRM (RRM), SH2 (SH2-like), SH3 (SH3), WD40 (Beta_propeller) and ZnF (C2H2-zf) (5). Of these, EFh (EF_hand), EGF-like (EFG), GTP_EFTU (P-loop_NTPase), IG-like(E-set), Pkinase (PKinase), RRM (RRM), SH2 (SH2-like), SH3 (SH3), WD40 (Beta_propeller) and ZnF (C2H2-zf) were confirmed by RNA-seq data (5). These domains appeared in high copy numbers within proteins, e.g., Ank (21-23) and WD40 (24), as repeats or highly abundant within proteins, e.g., SH3 (25, 26). Therefore, we hypothesized that highly abundant domains might experience a high number of translocation events. To this end, we applied EvoProDom to the assembly of organisms (EvoProDomDB) and identified, in total, 2,041 mobile domains. Of these, 259 had undergone translocation events and 1,782 were involved in indel events (Table *2*, Table *3*). The Pfam domains were classified into super families (6, 7), and translocation events and indel event frequencies were grouped by super families (Table 2, Table 3, respectively). Among the ten most frequent domain super families were SH3 (Src homology-3 domain), Ig (Immunoglobulin superfamily) and Ank (*Table 2*). Likewise, the most frequent super families of mobile domains that were involved in indel events were "Unknown", P-loop_NTPase and TPR (*Table 3*).

The SH3 super family contains SH3_2 (239 translocations), SH3_1 (198 translocations) and SH3_9 (193 translocations). SH3 (src Homology-3) domains were small protein domains, of approximately 50 amino acids in length (27, 28). They were found in diverse intracellular or membrane-associated proteins (29-31), for example, fodrin and yeast actin

binding protein ABP-1. SH3 domains mediate PPIs to facilitate a protein complex assembly (25). The Ig super family contains Ig (219 translocations), I-set (135 translocations), V-set (117 translocations) and Ig_2 (116 translocations). These domains were found in the vertebrate immune system (V-set), cell surface proteins and in intracellular muscle proteins (I-set) (32, 33). Ank repeats super family is comprised of Ank_2 (231 translocations), Ank_4 (184 translocations) and Ank_5 (94 translocations). These domains were found in copies, which form arrays. These repeats were involved in PPIs that regulate cell cycle transition from G1 to S (21-23). This regulation is achieved by the complex formation of INK4 proteins (inhibitors of cyclin-dependent kinase 4) and inhibition of cyclin-dependent kinases 4 and 6 (CDK4/6) proteins (23). These findings show that protein domains enriched in chimeric transcripts undergone many translocations. This supports a connection between chimeric transcripts and EvoProDom translocations. In addition, translocation events for protein domains such as the Pkinase and ubiquitin were found in multiple events and formed a new fusion. Moreover, from each novel transcript, one domain underwent a translocation event (5). Note that both super families with the most and the least number of translocations, SH3 (630) and SH2-like (6), were enriched in chimeric transcripts (Table *2*).

Discussion

Here, we presented a novel protein evolution model, EvoProDom, which was developed according to the “mix and merge” of protein domains. The EvoProDom model was implemented with and tested on EvoProDomDB. EvoProDomDB consists of genomic and proteome data, along with orthologous protein and protein domain data, for 109

organisms from diverse taxa. In the EvoProDom model, translocations, indel and duplication events were defined to reflect changes in a protein's domain content among orthologous groups. Moreover, in this model, these changes in protein domain composition were manifested at an organism level. Thus, SH3 was observed as a highly abundant protein domain in translocations, which binds ligands (25, 26) and mediates PPIs (34). Notably, the repetitive domains, such as Ank (21-23) and WD40 (24) appeared in many copies in proteins. These domains fold into 3D structures and generally mediate PPIs (21-24) by means of novel chimeric proteins that form novel PPIs and alter PPI networks of parent proteins (35). Accordingly, these domains, e.g., SH3_2, Ig and Ank_2 and others (see Results), were enriched in multiple fusion event chimeric transcripts (5). As hypothesized, these domains participated in a high number of evolutionary translocation events. The repetition of the domains provides a probable explanation for a high frequency of their translocation events. Fusions were produced as a slippage of two parent genes. These parent genes were truncated at a junction point, resulting in loss of domains. Consequently, the proper function of a chimeric protein is impaired (35). As an example, fusion within a catalytic domain would render the protein non-functional, and such fusion would be selected against. Naturally, due to selection, repetitive domains, which appear in a high copy number, would appear in chimeras at higher frequencies than expected from their shear amount alone, albeit, with lower repeats. Indeed, an average copy number of these domains was reduced in chimeric transcripts (5). In EvoProDom, repetitive domains or abundant domains, e.g., SH3, within KO groups, resulted in a higher number of distinct DAs. This translates into a higher number of (ko,item) pairs (see Materials and Methods). Consequently, repetitive and

abundant domains contribute more to the pool of mobile domains from which translocation events were determined, and were thus highly abundant in translocation events. Together, these results indicate that translocation events involving repetitive domains and highly abundant domains rewire PPI networks to achieve adaptive evolution.

The inclusion of new organisms into EvoProDomDB required only full genomes and annotated proteomes. Protein domain content and orthologous protein data were identified from protein sequences using the Pfam search tool (6, 7) and KoFamKOALA(8), respectively. The use of these tools enabled the applicability of EvoProDom to any new organism with a full genome and annotated proteome. Moreover, the usage of these tools in EvoProDom provides a general method to obtain protein domain content and orthologous protein annotation. This method may be applicable to other research fields such as proteomics (9), protein design (10) and host-virus interactions (11). Thus, EvoProDom presents a novel model for protein evolution based on “mix and merge” of protein domains rather than DNA based models. This confers the advantage of chromosomal alterations in evolutionary events.

List of abbreviations used

DA	Domain Architecture
EvoProDom	Evolution of Protein Domains
PPI	Protein-Protein Interaction
KEGG	Kyoto Encyclopedia of Genes and Genomes

KO	KEGG Ortholog

Declarations

Competing interests

None.

Consent for Publication

All co-authors consent to the publication of this manuscript.

Acknowledgments:

M.F.M. is a member of the Dangoor Center for Personalized Medicine and the Data Science Institute (DSI), Bar-Ilan University, Israel.

We would like to thank Dr. Eivatar Nevo for his expertise and helpful comments on the manuscript.

Funding

This work was supported by Grant for Biomarkers for treatment of Arthritis patients (Israel Innovation Authority, 66824, 1.7.2019-30.6.2020) and The Roland and Dawn Arnall Foundation, Research Grant, 205227 1.9.2018-31.8.2019

Author contributions

M.F.M designed the study, supervised it and wrote the paper; G.C, A.G produced the study, verified results and wrote the papers.

References

1. Marsh JA, Teichmann SA. How do proteins gain new domains? Genome Biology. 2010;11(7):126.

2. Basu MK, Poliakov E, Rogozin IB. Domain mobility in proteins: functional and evolutionary implications. Briefings in Bioinformatics. 2009;10(3):205-16.
3. Cancherini DV, França GS, de Souza SJ. The role of exon shuffling in shaping protein-protein interaction networks. BMC Genomics. 2010;11 Suppl 5:S11.
4. Bornberg-Bauer EA, M Mar. Dynamics and adaptive benefits of modular protein evolution. Current Opinion in Structural Biology. 2013;23(3):459 - 66.
5. Frenkel-Morgenstern M, Valencia A. Novel domain combinations in proteins encoded by chimeric transcripts. Bioinformatics. 2012;28(12):i67.
6. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, et al. The Pfam protein families database. Nucleic Acids Res. 2008;36(Database issue):D281-8.
7. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44(D1):D279-85.
8. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. bioRxiv. 2019:602110.
9. Jadwin JA, Ogiue-Ikeda M, Machida K. The application of modular protein domains in proteomics. FEBS Lett. 2012;586(17):2586-96.
10. Setiawan D, Brender J, Zhang Y. Recent advances in automated protein design and its future challenges. Expert Opin Drug Discov. 2018;13(7):587-604.
11. Brito AF, Pinney JW. Protein-Protein Interactions in Virus-Host Systems. Front Microbiol. 2017;8:1557.
12. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44(D1):D457-62.
13. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research. 2017;45(D1):D353.
14. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, et al. GenBank. Nucleic Acids Res. 2018;46(D1):D41-D7.
15. Davies KTJ, Bennett NC, Faulkes CG, Rossiter SJ. Limited evidence for parallel molecular adaptations associated with the subterranean niche in mammals: a comparative study of three superorders. Mol Biol Evol. 2018.
16. Tavares WC, Seuánez HN. Changes in selection intensity on the mitogenome of subterranean and fossorial rodents respective to aboveground species. Mamm Genome. 2018;29(5-6):353-63.
17. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, et al. The BioGRID interaction database: 2015 update. Nucleic Acids Res. 2015;43(Database issue):D470-8.
18. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, et al. The BioGRID interaction database: 2017 update. Nucleic Acids Res. 2017;45(D1):D369-D79.
19. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. Nature. 2008;451:990.
20. Marsh JA, Teichmann SA. Structure, Dynamics, Assembly, and Evolution of Protein Complexes. Annual Review of Biochemistry. 2015;84(1):551-75.

21. Sedgwick SG, Smerdon SJ. The ankyrin repeat: a diversity of interactions on a common structural framework. Trends Biochem Sci. 1999;24(8):311-6.
22. Mosavi LK, Cammett TJ, Desrosiers DC, Peng ZY. The ankyrin repeat as molecular architecture for protein recognition. Protein Sci. 2004;13(6):1435-48.
23. Li J, Mahajan A, Tsai MD. Ankyrin repeat: a unique motif mediating protein-protein interactions. Biochemistry. 2006;45(51):15168-78.
24. Xu C, Min J. Structure and function of WD40 domain proteins. Protein Cell. 2011;2(3):202-14.
25. Morton CJ, Campbell ID. SH3 domains. Molecular 'Velcro'. Curr Biol. 1994;4(7):615-7.
26. Kaneko T, Li L, Li SS. The SH3 domain--a family of versatile peptide- and protein-recognition module. Front Biosci. 2008;13:4938-52.
27. Pawson T, Schlessingert J. SH2 and SH3 domains. Curr Biol. 1993;3(7):434-42.
28. Mayer BJ. SH3 domains: complexity in moderation. J Cell Sci. 2001;114(Pt 7):1253-63.
29. Musacchio A, Gibson T, Lehto VP, Saraste M. SH3--an abundant protein domain in search of a function. FEBS Lett. 1992;307(1):55-61.
30. Mayer BJ, Baltimore D. Signalling through SH2 and SH3 domains. Trends Cell Biol. 1993;3(1):8-13.
31. Pawson T. Protein modules and signalling networks. Nature. 1995;373(6515):573-80.
32. Williams AF, Barclay AN. The immunoglobulin superfamily--domains for cell surface recognition. Annu Rev Immunol. 1988;6:381-405.
33. Harpaz Y, Chothia C. Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. J Mol Biol. 1994;238(4):528-39.
34. Malagrinò F, Troilo F, Bonetti D, Toto A, Gianni S. Mapping the allosteric network within a SH3 domain. Sci Rep. 2019;9(1):8279.
35. Frenkel-Morgenstern M, Gorohovski A, Tagore S, Sekar V, Vazquez M, Valencia A. ChiPPI: a novel method for mapping chimeric protein–protein interactions uncovers selection principles of protein fusion events in cancer. Nucleic Acids Research. 2017;45(12):7094-105.

Figures

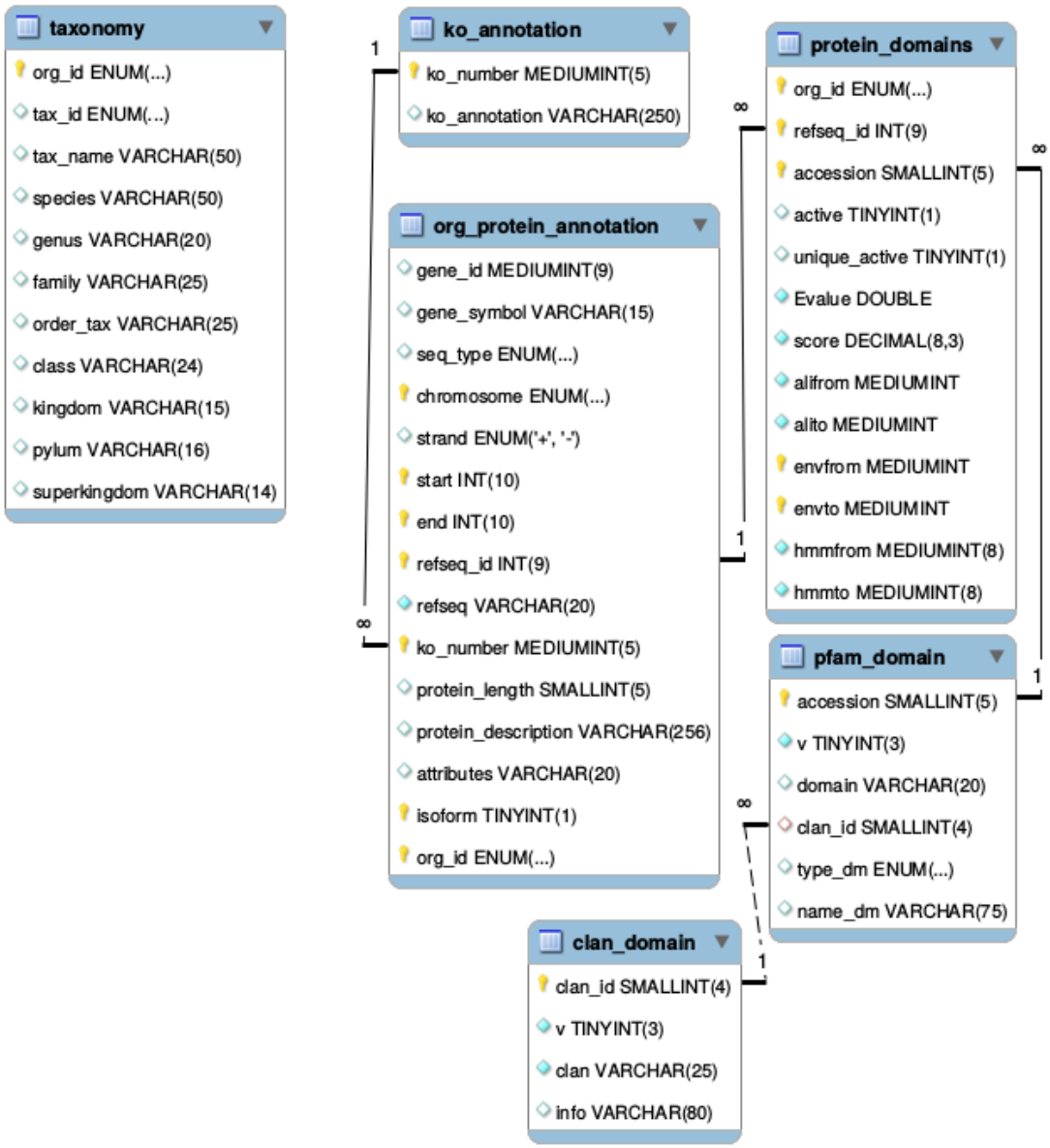

Figure 1 MySQL scheme for EvoProDomDB. Six relation tables were included. Of them, four contained data regarding annotation for taxonomy (taxonomy), KO (ko_annotation,), super families (clan_domain), pfam domain (pfam_domain), such as taxonomy ranks, e.g, species and genus; KO, domain and super family descriptions, respectively. The main relational tables comprise protein, genomic and proteomic data (org_protein_annotation), as well as protein domain content (pfam data) (see the main text for details).

Figure 2 Study workflow: A collection of organisms, 109 in total, was used to implement and test the EvoProDom model. The collection included six categories: (i) 15 fish; (ii) 4 subterranean, 8 fossorial and 21 aboveground animals (15, 16); (iii) 65 organisms with known PPIs (BioGrid version 3.5.173, (17, 18)); (vi) 17 organisms with HiC datasets; (v) 4 cats; and (iv) 15 pathogenic organisms (19).
Protein domains were predicted by utilizing Pfam (release 32.0) database along with search tool (6, 7). Orthologous proteins were defined as belonging to Kyoto Encyclopedia of Genes and Genomes (KEGG) (12, 13) ortholog (KO, KEGG ortholog) and assignment to KO group was obtained using KofamKOALA (8).

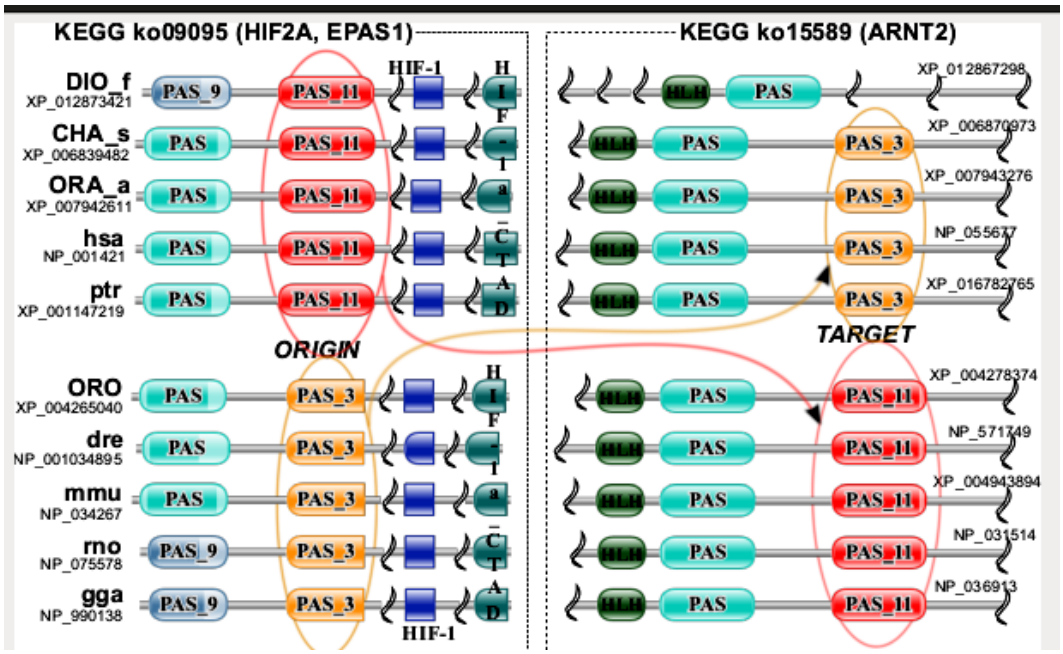

Figure 3 Illustration of translocation event for PAS_11. Red domain (PAS_11) undergone reciprocal a translocation event between two orthologous protein groups K09095 (HIF2A) and K15589 (ARNT2). Accordingly, the red domain (PAS_11) is present in HIF2A and absent from ARNT2 for organisms DIO, etc. while for organisms ORO etc., PAS_11 is present in ARNT2 and missing from HIF2A.

TABLES

Table 1 EvoProDom model is applied to an assembly of organisms from diverse taxa. In total, 109 organisms were included in the ensemble and grouped according to six categories: (i) 15 fish; (ii) 4 subterranean (S), 8 fossorial (F) and 21 aboveground (A) animals (SFA) (15, 16); (iii) 65 organisms with known PPIs (BioGrid version 3.5.173, (17, 18)); (vi) 17 organisms with HiC datasets (GEO_hic); (v) 4 cats; and (iv) 15 pathogenic organisms (Jones, et al. 2008) (19). Organisms with HiC datasets were obtained by searching 'HiC' in the NCBI GEO database. Taxonomy ID, organism ID, organism name and common name were provided. Additionally, assembly and group classification are indicated. Organism ID is 3-4 letter code were lowercase letter code correspond to KEGG organisms while uppercase letter to organisms not included in KEGG database.

NCBI taxid	Organism ID	Organism Name	Ecology	Common Name	Source	Assembly
32536	aju	*Acinonyx jubatus*	na	cheetah	cats	GCF_001443585.1_aciJub1
180454	aga	*Anopheles gambiae PEST*	na	African malaria	biogrid_3.5.173	GCF_000005575.2_AgamP3

				mosquito		
7460	ame	*Apis mellifera*	na	honey bee	biogrid_3.5.173	GCF_000002195.4_Amel_4.5
3702	ath	*Arabidopsis thaliana*	na	thale cress	GEO_hic,biogrid_3.5.173	GCF_000001735.4_TAIR10.1
7994	ASM	*Astyanax mexicanus*	na	Mexican tetra	fish	GCF_000372685.2_Astyanax_mexicanus-2.0
224308	bsu	*Bacillus subtilis 168*	na	na	biogrid_3.5.173	GCF_002009135.1_ASM200913v1
1415167	bsp	*Bacillus subtilis PY79*	na	na	GEO_hic	GCF_000497485.1_ASM49748v1
9913	bta	*Bos taurus*	A	cattle	biogrid_3.5.173,SFA	GCF_002263795.1_ARS-UCD1.2
6239	cel	*Caenorhabditis elegans*	na	nematode	GEO_hic,biogrid_3.5.173	GCF_000002985.6_WBcel235
237561	cal	*Candida albicans SC5314*	na	na	biogrid_3.5.173,Jones, et al. 2008	GCF_000182965.3_ASM18296v3
9615	cfa	*Canis familiaris*	na	dog	biogrid_3.5.173	GCF_000002285.3_CanFam3.1
7957	CAA	*Carassius auratus*	na	goldfish	fish	GCF_003368295.1 ASM336829v1
190650	ccr	*Caulobacter vibrioides*	na	na	GEO_hic	GCF_000006905.1_ASM690v1
10141	CAP	*Cavia porcellus*	A	domestic guinea pig	biogrid_3.5.173,SFA	GCF_000151735.1 Cavpor3.0
34839	CHL	*Chinchilla lanigera*	A	long-tailed chinchilla	SFA	GCF_000276665.1_ChiLan1.0
3055	cre	*Chlamydomonas reinhardtii*	na	green algae	biogrid_3.5.173	GCF_000002595.1_v3.0
60711	csab	*Chlorocebus sabaeus*	na	green monkey	biogrid_3.5.173	GCF_000409795.2_Chlorocebus_sabeus_1.1
185453	CHA	*Chrysochloris asiatica*	S	Cape golden mole	SFA	GCF_000296735.1_ChrAsi1.0
143302	COC	*Condylura cristata*	F	star-nosed mole	SFA	GCF_000260355.1_ConCri1.0
56716	COG	*Cottoperca gobio*	na	channel bull blenny	fish	GCF_900634415.1 fCotGob3.1
10029	cge	*Cricetulus griseus*	A	Chinese hamster	biogrid_3.5.173,SFA	GCF_000419365.1_C_griseus_v1.0
7962	ccar	*Cyprinus carpio*	na	common carp	fish	GCF_000951615.1_common_carp_genome
7955	dre	*Danio rerio*	na	zebrafish	fish,biogrid_3.5.173,GEO_hic	GCF_000002035.6_GRCz11
9361	DAN	*Dasypus novemcinctus*	F	nine-banded armadillo	SFA	GCF_000208655.1_Dasnov3.0
352472	ddi	*Dictyostelium discoideum AX4*	na	na	biogrid_3.5.173	GCF_000004695.1_dicty_2.7
10020	DIO	*Dipodomys ordii*	F	Ord's kangaroo rat	SFA	GCF_000151885.1_Dord_2.0
7227	dme	*Drosophila melanogaster*	A	fruit fly	SFA,GEO_hic,biogrid_3.5.173	GCF_000001215.4_Release_6_plus_ISO1_MT
9371	ECTE	*Echinops telfairi*	A	small Madagascar hedgehog	SFA	GCF_000313985.1 EchTel2.0
28737	ELE	*Elephantulus edwardii*	A	Cape elephant shrew	SFA	GCF_000299155.1 EleEdw1.0
227321	ani	*Emericella nidulans FGSC A4*	na	Aspergillus nidulans	biogrid_3.5.173	GCF_000149205.2_ASM14920v2
9796	ecb	*Equus caballus*	na	horse	biogrid_3.5.173	GCF_002863925.

						1_EquCab3.0
9365	ERE	*Erinaceus europaeus*	A	western European hedgehog	SFA	GCF_000296755.1_EriEur2.0
511145	eco	*Escherichia coli str. K-12 substr. MG1655*	na	na	biogrid_3.5.173	GCF_001566335.1_ASM156633v1
9685	fca	*Felis catus*	A	domestic cat	SFA,cats	GCF_000181335.3_Felis_catus_9.0
885580	FUD	*Fukomys damarensis*	S	Damara mole-rat	SFA	GCF_000743615.1_DMR_v1.0
9031	gga	*Gallus gallus*	A	chicken	biogrid_3.5.173,SFA,GEO_hic	GCF_000002315.5_GRCg6a
3847	gmx	*Glycine max*	na	soybean	biogrid_3.5.173	GCF_000004515.4_Glycine_max_v2.0
11103	HCV	*Hepatitis C virus*	na	HCV	biogrid_3.5.173,Jones, et al. 2008	GCF_000861845.1_ViralProj15432
10181	hgl	*Heterocephalus glaber*	S	naked mole-rat	SFA	GCF_000247695.1_HetGla_female_1.0
9606	hsa	*Homo sapiens*	A	human	biogrid_3.5.173,SFA,GEO_hic	GCF_000001405.37_GRCh38.p11
10376	HHV4	*Human gammaherpesvirus 4*	na	EBV	GEO_hic,biogrid_3.5.173	GCF_002402265.1_Decoy
37296	HHV8	*Human gammaherpesvirus 8*	na	KSHV	GEO_hic,biogrid_3.5.173,Jones, et al. 2008	GCF_000838265.1_ViralProj14158
10298	HHV1	*Human Herpesvirus 1*	na	Herpes simplex virus type 1	biogrid_3.5.173,Jones, et al. 2008	GCF_000859985.2_ViralProj15217
10310	HHV2	*Human Herpesvirus 2*	na	HHV2	biogrid_3.5.173	GCF_000858385.2_ViralProj15218
10335	HHV3	*Human Herpesvirus 3*	na	Varicella-zoster virus	biogrid_3.5.173,Jones, et al. 2008	GCF_000858285.1_ViralProj15198
10359	HHV5	*Human Herpesvirus 5*	na	Human cytomegalovirus	biogrid_3.5.173,Jones, et al. 2008	GCF_000845245.1_ViralProj14559
32603	HHV6A	*Human Herpesvirus 6A*	na	HHV6A	biogrid_3.5.173	GCF_000845685.1_ViralProj14462
32604	HHV6B	*Human Herpesvirus 6B*	na	HHV6B	biogrid_3.5.173	GCF_000846365.1_ViralProj14422
10372	HHV7	*Human Herpesvirus 7*	na	HHV7	biogrid_3.5.173,Jones, et al. 2008	GCF_000848125.1_ViralProj14625
11676	HIV1	*Human Immunodeficiency Virus 1*	na	HIV1	biogrid_3.5.173,Jones, et al. 2008	GCF_000864765.1_ViralProj15476
11709	HIV2	*Human Immunodeficiency Virus 2*	na	HIV2	biogrid_3.5.173,Jones, et al. 2008	GCF_000856385.1_ViralProj14991
333760	HPV16	*Human papillomavirus 16*	na	HPV16	biogrid_3.5.173,GEO_hic,Jones, et al. 2008	GCF_000863945.3_ViralProj15505
333759	HPV10	*Human papillomavirus type 10*	na	HPV10	biogrid_3.5.173,Jones, et al. 2008	GCF_000864905.1_ViralProj15504
10600	HPV6b	*Human papillomavirus type 6b*	na	HPV6b	biogrid_3.5.173,Jones, et al. 2008	GCF_000861945.1_ViralProj15454
43179	ICT	*Ictidomys tridecemlineatus*	F	thirteen-lined ground squirrel	SFA	GCF_000236235.1 SpeTri2.0
8187	lcf	*Lates calcarifer*	na	barramundi perch	fish	GCF_001640805.1_ASM164080v1

7897	lcm	*Latimeria chalumnae*	na	coelacanth	fish	GCF_000225785.1_LatCha1
7918	LEO	*Lepisosteus oculatus*	na	spotted gar	fish	GCF_000242695.1 LepOcu1
9785	lav	*Loxodonta Africana*	A	African savanna elephant	SFA	GCF_000001905.1_Loxafr3.0
9544	MAM	*Macaca mulatta*	na	Rhesus monkey	biogrid_3.5.173,GEO_hic	GCF_003339765.3 Mmul_10
9993	MARM	*Marmota marmota*	F	European marmot	SFA	GCF_001458135.1 marMar2.1
9103	mgp	*Meleagris gallopavo*	na	turkey	biogrid_3.5.173	GCF_000146605.2_Turkey_5.0
79684	MIO	*Microtus ochrogaster*	F	prairie vole	SFA	GCF_000317375.1_MicOch1.0
10090	mmu	*Mus musculus*	A	house mouse	biogrid_3.5.173,SFA,GEO_hic	GCF_000001635.26_GRCm38.p6
83332	mtv	*Mycobacterium tuberculosis H37Rv*	na	na	biogrid_3.5.173,Jones, et al. 2008	GCF_000195955.2_ASM19595v2
243273	mge	*Mycoplasma genitalium*	na	na	Jones, et al. 2008	GCF_000027325.1_ASM2732v1
1026970	ngi	*Nannospalax galili*	S	Upper Galilee mountains blind mole rat	SFA	GCF_000622305.1_S.galili_v1.0
56216	NEL	*Neotoma lepida*	A	desert woodrat	SFA	GCF_001675575.1 ASM167557v1
367110	ncr	*Neurospora crassa OR74A*	na	na	biogrid_3.5.173	GCF_000182925.2_NC12
4098	nto	*Nicotiana tomentosiformis*	na	tobacco	biogrid_3.5.173	GCF_000390325.2_Ntom_v01
61853	nle	*Nomascus leucogenys*	na	northern white-cheeked gibbon	GEO_hic	GCF_000146795.2_Nleu_3.0
105023	nfu	*Nothobranchius furzeri*	na	turquoise killifish	fish	GCF_001465895.1_Nfu_20140520
8208	ncc	*Notothenia coriiceps*	na	black rockcod	fish	GCF_000735185.1_NC01
10160	OCD	*Octodon degus*	F	degu	SFA	GCF_000260255.1_OctDeg1.0
9258	oaa	*Ornithorhynchus anatinus*	A	platypus	SFA	GCF_000002275.2_Ornithorhynchus_anatinus_5.0.1
1230840	ORA	*Orycteropus afer afer*	*A*	aardvark	SFA	GCF_000298275.1_OryAfe1.0
9986	ocu	*Oryctolagus cuniculus*	na	rabbit	biogrid_3.5.173,GEO_hic	GCF_000003625.3_OryCun2.0
39947	osa	*Oryza sativa Japonica*	na	rice	biogrid_3.5.173	GCF_001433935.1_IRGSP-1.0
8090	ola	*Oryzias latipes*	na	Japanese medaka	fish	GCF_002234675.1_ASM223467v1
30732	ORM	*Oryzias melastigma*	na	Indian medaka	fish	GCF_002922805.1 Om_v0.7.RACA
9940	oas	*Ovis aries*	na	sheep	biogrid_3.5.173	GCF_000298735.2_Oar_v4.0
9598	ptr	*Pan troglodytes*	A	chimpanze e	biogrid_3.5.173,SFA	GCF_002880755.1_Clint_PTRv2
9691	PAP	*Panthera pardus*	na	leopard	cats	GCF_001857705.1_PanPar1.0
74533	ptg	*Panthera tigris altaica*	na	tiger	cats	GCF_000464555.1_PanTig1.0
121224	phu	*Pediculus humanus corporis*	na	human body louse	biogrid_3.5.173	GCF_000006295.1_JCVI_LOUSE_1.0
8167	PEF	*Perca flavescens*	na	yellow perch	fish	GCF_004354835.1 PFLA_1.0

230844	PEM	*Peromyscus maniculatus bairdii*	A	prairie deer mouse	SFA	GCF_000500345.1_Pman_1.0
36329	pfa	*Plasmodium falciparum 3D7*	na	malaria parasite P. falciparum	biogrid_3.5.173,Jones, et al. 2008	GCF_000002765.4_ASM276v2
8081	pret	*Poecilia reticulata*	na	guppy	fish	GCF_000633615.1_Guppy_female_1.0_MT
10116	rno	*Rattus norvegicus*	A	Norway rat	biogrid_3.5.173,SFA	GCF_000001895.5_Rnor_6.0
3988	rcu	*Ricinus communis*	na	castor bean	biogrid_3.5.173	GCF_000151685.1_JCVI_RCG_1.1
4932	sce	*Saccharomyces cerevisiae S288c*	na	baker's yeast	biogrid_3.5.173,GEO_hic	GCF_000146045.2_R64
8030	sasa	*Salmo salar*	na	Atlantic salmon	fish	GCF_000233375.1_ICSASG_v2
4896	spo	*Schizosaccharomyces pombe*	na	fission yeast	GEO_hic,biogrid_3.5.173	GCF_000002945.1_ASM294v2
88036	smo	*Selaginella moellendorffii*	na	na	biogrid_3.5.173	GCF_000143415.4_v1.0
11723	SIV	*Simian Immunodeficiency Virus*	na	SIV	biogrid_3.5.173	GCF_000863925.1_ViralProj15501
1891767	SV40	*Simian Virus 40*	na	Macaca mulatta polyomavirus 1	biogrid_3.5.173	GCF_000837645.1_ViralProj14024
4081	sly	*Solanum lycopersicum*	na	tomato	biogrid_3.5.173	GCF_000188115.3_SL2.50
4113	sot	*Solanum tuberosum*	na	potato	biogrid_3.5.173	GCF_000226075.1_SolTub_3.0
42254	SOA	*Sorex araneus*	A	European shrew	SFA	GCF_000181275.2 SorAra2.0
7668	spu	*Strongylocentrotus purpuratus*	na	purple sea urchin	biogrid_3.5.173	GCF_000002235.4_Spur_4.2
9823	ssc	*Sus scrofa*	A	pig	biogrid_3.5.173,SFA	GCF_000003025.6_Sscrofa11.1
12242	TMV	*Tobacco Mosaic Virus*	na	TMV	biogrid_3.5.173	GCF_000854365.1_ViralProj15071
9999	URP	*Urocitellus parryii*	F	Arctic ground squirrel	SFA	GCF_003426925.1 ASM342692v1
237631	USM	*Ustilago maydis 521*	na	na	biogrid_3.5.173	GCF_000328475.2 Umaydis521_2.0
10245	VAV	*Vaccinia Virus*	na	na	biogrid_3.5.173	GCF_000860085.1_ViralProj15241
29760	vvi	*Vitis vinifera*	na	wine grape	biogrid_3.5.173	GCF_000003745.3_12X
8355	xla	*Xenopus laevis*	na	African clawed frog	biogrid_3.5.173	GCF_001663975.1_Xenopus_laevis_v2
4577	zma	*Zea mays*	na	Maize	biogrid_3.5.173	GCF_000005005.2_B73_RefGen_v4

Table 2 Translocation events per super family (Counts). Super family annotation is provided.

Super family Id	**Super family Name**	**Counts**	**Super family Description**
0010.21	SH3	630	Src homology-3 domain

0011.26	Ig	616	Immunoglobulin superfamily
0465.3	Ank	529	Ankyrin repeat superfamily
0001.27	EGF	414	EGF superfamily
0361.4	C2H2-zf	390	Classical C2H2 and C2HC zinc fingers
0022.32	LRR	282	Leucine Rich Repeat
0020.25	TPR	246	Tetratrico peptide repeat superfamily
0229.11	RING	242	Ring-finger/U-box superfamily
0186.14	Beta_propeller	222	Beta propeller clan
0221.11	RRM	210	RRM-like clan
9999.0	Unknown	210	null
0159.16	E-set	187	Ig-like fold superfamily (E-set)
0466.3	PDZ-like	165	PDZ domain-like peptide-binding superfamily
0016.22	PKinase	164	Protein kinase superfamily
0266.9	PH	141	PH domain-like superfamily
0023.34	P-loop_NTPase	121	P-loop containing nucleoside triphosphate hydrolase superfamily
0220.12	EF_hand	115	EF-hand like superfamily
0511.3	Retroviral_zf	95	Retrovirus zinc finger-like domains
0271.7	F-box	79	F-box-like domain
0003.21	SAM	74	Sterile Alpha Motif (SAM) domain
0390.4	zf-FYVE-PHD	47	FYVE/PHD zinc finger superfamily
0063.25	NADP_Rossmann	37	FAD/NAD(P)-binding Rossmann fold Superfamily
0357.4	SMAD-FHA	37	SMAD/FHA domain superfamily
0123.18	HTH	34	Helix-turn-helix clan
0680.1	WW	34	WW domain
0167.15	Zn_Beta_Ribbon	33	Zinc beta-ribbon
0006.20	C1	25	Protein kinase C, C1 domain
0214.13	UBA	24	UBA superfamily
0306.4	HeH	24	LEM/SAP HeH motif
0188.10	CH	23	Calponin homology domain
0459.3	BRCT-like	23	BRCT like
0537.2	CCCH_zf	22	CCCH-zinc finger
0004.20	Concanavalin	20	Concanavalin-like lectin/glucanase superfamily
0072.20	Ubiquitin	19	Ubiquitin superfamily
0033.14	POZ	17	POZ domain superfamily
0154.11	C2	11	C2 superfamily
0007.18	KH	9	K-Homology (KH) domain Superfamily
0021.18	OB	8	OB fold
0029.20	Cupin	8	Cupin fold
0049.15	Tudor	8	Tudor domain 'Royal family'
0164.13	CUB	8	CUB clan

0172.17	Thioredoxin	8	Thioredoxin-like
0212.9	SNARE	8	SNARE-like superfamily
0392.4	Chaperone-J	8	Chaperone J-domain superfamily
0124.15	Peptidase_PA	7	Peptidase clan PA
0137.15	HAD	7	HAD superfamily
0364.4	Leu-IlvD	7	LeuD/IlvD-like
0575.2	EFTPs	7	Translation proteins of Elongation Factors superfamily
0541.2	SH2-like	6	SH2, phosphotyrosine-recognition domain superfamily
0173.11	STIR	5	STIR superfamily
0192.13	GPCR_A	5	Family A G protein-coupled receptor-like superfamily
0244.9	PGBD	5	PGBD superfamily
0602.2	Kringle	5	Kringle/FnII superfamily
0671.1	AAA_lid	5	AAA+ ATPase lid domain superfamily
0041.13	Death	4	Death Domain Superfamily
0178.16	PUA	4	PUA/ASCH superfamily
0183.14	PAS_Fold	4	PAS domain clan
0642.1	SOCS_box	4	SOCS-box like superfamily
0015.20	MFS	3	Major Facilitator Superfamily
0030.16	Ion_channel	3	Ion channel (VIC) superfamily
0084.13	ADP-ribosyl	3	ADP-ribosylation Superfamily
0198.16	HHH	3	Helix-hairpin-helix superfamily
0497.3	GST_C	3	Glutathione S-transferase, C-terminal domain
0661.1	Gain	3	GPCR autoproteolysis inducing
0027.15	RdRP	2	RNA dependent RNA polymerase
0028.22	AB_hydrolase	2	Alpha/Beta hydrolase fold
0055.13	Nucleoplasmin	2	Nucleoplasmin-like/VP (viral coat and capsid proteins) superfamily
0107.12	KOW	2	KOW domain
0202.11	GBD	2	Galactose-binding domain-like superfamily
0492.3	S4	2	S4 domain superfamily
0531.2	AMP-binding_C	2	AMP-binding enzyme C-terminal domain superfamily
0005.27	Kazal	1	Kazal like domain
0025.14	His_Kinase_A	1	His Kinase A (phospho-acceptor) domain
0026.20	CU_oxidase	1	Multicopper oxidase-like domain
0070.13	ACT	1	ACT-like domain
0088.16	Alk_phosphatase	1	Alkaline phosphatase-like
0109.12	CDA	1	Cytidine deaminase-like (CDA) superfamily
0110.12	GT-A	1	Glycosyl transferase clan GT-A
0113.13	GT-B	1	Glycosyl transferase clan GT-B

0114.12	HMG-box	1	HMG-box like superfamily
0117.11	uPAR_Ly6_toxin	1	uPAR/Ly6/CD59/snake toxin-receptor superfamily
0125.15	Peptidase_CA	1	Peptidase clan CA
0144.13	Periplas_BP	1	Periplasmic binding protein like
0196.12	DSRM	1	DSRM-like clan
0236.17	PDDEXK	1	PD-(D/E)XK nuclease superfamily
0381.4	Metallo-HOrase	1	Metallo-hydrolase/oxidoreductase superfamily
0426.4	HRDC-like	1	HRDC-like superfamily
0449.3	G-PATCH	1	DExH-box splicing factor binding site
0505.3	Pentapeptide	1	Pentapeptide repeat
0547.2	GF_recep_C-rich	1	Growth factor receptor Cys-rich
0552.2	Hect	1	Hect, E3 ligase catalytic domain
0607.2	TNF_receptor	1	TNF receptor-like superfamily
0630.1	PSI	1	Plexin fold superfamily
0672.1	p35	1	Baculovirus p35 protein superfamily
0677.1	GHMP_C	1	GHMP C-terminal domain superfamily

Table 3 Indel events per super family (Counts). Super family annotation is provided

Super family Id	Super family Name	Counts	Super family Description
9999.0	Unknown	8536	null
0023.34	P-loop_NTPase	688	P-loop containing nucleoside triphosphate hydrolase superfamily
0020.25	TPR	520	Tetratrico peptide repeat superfamily
0123.18	HTH	453	Helix-turn-helix clan
0219.14	RNase_H	347	Ribonuclease H-like superfamily
0361.4	C2H2-zf	228	Classical C2H2 and C2HC zinc fingers
0665.1	BET	216	BET superfamily
0021.18	OB	196	OB fold
0072.20	Ubiquitin	187	Ubiquitin superfamily
0063.25	NADP_Rossmann	173	FAD/NAD(P)-binding Rossmann fold Superfamily
0148.11	Viral_Gag	157	Viral Gag protein
0186.14	Beta_propeller	152	Beta propeller clan
0016.22	PKinase	148	Protein kinase superfamily
0266.9	PH	143	PH domain-like superfamily
0031.13	Phosphatase	142	Phosphatase superfamily
0523.3	GAG-polyprotein	140	LTR-copia-type polyprotein segment
0671.1	AAA_lid	140	AAA+ ATPase lid domain superfamily
0550.2	SRCR	123	SRCR-like
0114.12	HMG-box	109	HMG-box like superfamily
0010.21	SH3	106	Src homology-3 domain
0137.15	HAD	106	HAD superfamily
0214.13	UBA	100	UBA superfamily
0172.17	Thioredoxin	98	Thioredoxin-like
0128.12	vWA-like	95	von Willebrand factor type A

0208.11	UBC	90	Ubiquitin conjugating enzyme like superfamily
0154.11	C2	88	C2 superfamily
0220.12	EF_hand	87	EF-hand like superfamily
0001.27	EGF	84	EGF superfamily
0081.13	MBD-like	83	MBD-like DNA-binding domain
0390.4	zf-FYVE-PHD	82	FYVE/PHD zinc finger superfamily
0168.15	PAN	77	PAN-like
0049.15	Tudor	75	Tudor domain 'Royal family'
0183.14	PAS_Fold	70	PAS domain clan
0459.3	BRCT-like	67	BRCT like
0159.16	E-set	66	Ig-like fold superfamily (E-set)
0575.2	EFTPs	65	Translation proteins of Elongation Factors superfamily
0126.18	Peptidase_MA	58	Peptidase clan MA
0029.20	Cupin	53	Cupin fold
0129.14	Peptidase_AA	52	Peptidase clan AA
0221.11	RRM	51	RRM-like clan
0451.3	FnI-like	51	von Willebrand Factor like superfamily
0041.13	Death	48	Death Domain Superfamily
0280.7	PIN	42	PIN domain superfamily
0202.11	GBD	37	Galactose-binding domain-like superfamily
0188.10	CH	36	Calponin homology domain
0464.3	5_3_exonuc_C	36	5'-3'-exonuclease C-terminal sub-region
0537.2	CCCH_zf	35	CCCH-zinc finger
0004.20	Concanavalin	34	Concanavalin-like lectin/glucanase superfamily
0556.2	PapD-like	32	PapD-like superfamily,immunoglobulin-like beta sandwich
0678.1	CLIP	31	CLIP domain superfamily
0167.15	Zn_Beta_Ribbon	30	Zinc beta-ribbon
0074.13	Matrix	28	Retroviral matrix superfamily
0240.8	PFK	27	PFK-like superfamily
0589.2	KIX_like	27	Kix domain of CBP (creb binding protein) and MED13/15
0113.13	GT-B	26	Glycosyl transferase clan GT-B
0257.9	Acetyltrans	25	N-acetyltransferase like
0087.13	Acyl-CoA_dh	23	Acyl-CoA dehydrogenase, C-terminal domain-like
0497.3	GST_C	21	Glutathione S-transferase, C-terminal domain
0011.26	Ig	20	Immunoglobulin superfamily
0179.14	ATP-grasp	20	ATP-grasp superfamily
0145.15	Golgi-transport	19	Golgi-transport
0229.11	RING	19	Ring-finger/U-box superfamily
0511.3	Retroviral_zf	19	Retrovirus zinc finger-like domains
0544.2	AcylCoA_ox_dh_N	17	Acyl-coenzyme A oxidase/dehydrogenase N-terminal
0573.2	KA1-like	17	Kinase associated domain 1-like
0036.24	TIM_barrel	16	Common phosphate binding-site TIM barrel superfamily
0103.12	Gal_mutarotase	16	Galactose Mutarotase-like superfamily
0630.1	PSI	16	Plexin fold superfamily
0039.12	HUP	15	HUP - HIGH-signature proteins, UspA, and PP-ATPase.
0125.15	Peptidase_CA	15	Peptidase clan CA
0306.4	HeH	15	LEM/SAP HeH motif
0028.22	AB_hydrolase	14	Alpha/Beta hydrolase fold
0109.12	CDA	14	Cytidine deaminase-like (CDA) superfamily

0175.11	TRASH	13	TRASH superfamily
0181.10	ABC-2	13	ABC-2-transporter-like clan
0600.2	S15_NS1	13	S15/NS1 RNA-binding domain superfamily
0616.1	SPOC	13	SPOC domain-like superfamily
0012.18	Histone	12	Histone superfamily
0192.13	GPCR_A	12	Family A G protein-coupled receptor-like superfamily
0265.7	HIT	12	HIT superfamily
0357.4	SMAD-FHA	11	SMAD/FHA domain superfamily
0369.4	GHD	11	Glycosyl hydrolase domain superfamily
0381.4	Metallo-HOrase	11	Metallo-hydrolase/oxidoreductase superfamily
0614.1	L27	11	L27 domain superfamily
0632.1	FERM_M	11	FERM middle domain superfamily
0005.27	Kazal	10	Kazal like domain
0056.12	C_Lectin	10	C-type lectin-like superfamily
0058.16	Glyco_hydro_tim	10	Tim barrel glycosyl hydrolase superfamily
0178.16	PUA	10	PUA/ASCH superfamily
0223.8	MACRO	10	MACRO domain superfamily
0389.4	TRAF	10	TRAF domain-like superfamily
0014.22	Glutaminase_I	9	Class-I Glutamine amidotransferase superfamily
0092.12	ADF	9	Actin depolymerizing Factor
0298.5	tRNA_bind_arm	9	tRNA-binding arm superfamily
0512.3	CRAL_TRIO	9	CRAL-TRIO domain superfamily
0561.2	LisH	9	LisH-like
0644.1	Fz	9	Frizzled cysteine-rich domain-related superfamily
0015.20	MFS	8	Major Facilitator Superfamily
0062.13	APC	8	APC superfamily
0064.12	CPA_AT	8	CPA/AT transporter superfamily
0070.13	ACT	8	ACT-like domain
0105.13	Hybrid	8	Barrel sandwich hybrid superfamily
0258.7	DALR	8	DALR superfamily
0314.4	PP-binding	8	ACP-like superfamily
0329.4	S5	8	Ribosomal protein S5 domain 2-like superfamily
0332.4	AcetylDC-like	8	Acetyl-decarboxylase like superfamily
0479.3	PLD	8	Phospholipase D superfamily
0483.3	PreATP-grasp	8	Probable substrate-binding preceding ATP-grasp domain
0673.1	GYF	8	GYF domain
0263.8	His-Me_finger	7	His-Me finger endonuclease superfamily
0343.4	MHC	7	MHC antigen-recognition domain
0344.4	4Fe-4S	7	4Fe-4S ferredoxins
0375.4	Transporter	7	Transporter superfamily, four TM region
0396.4	Marvel-like	7	MARVEL domain containing superfamily
0533.2	PRTase-like	7	PRPP synthetase-associated protein 1
0647.1	FG_rpt	7	Nucleoporin FG repeat
0661.1	Gain	7	GPCR autoproteolysis inducing
0034.15	Amidohydrolase	6	Amidohydrolase superfamily
0050.12	HotDog	6	HotDog superfamily
0110.12	GT-A	6	Glycosyl transferase clan GT-A
0190.12	HSP20	6	HSP20-like chaperone superfamily
0198.16	HHH	6	Helix-hairpin-helix superfamily
0236.17	PDDEXK	6	PD-(D/E)XK nuclease superfamily

0268.7	Pec_lyase-like	6	Pectate lyase-like beta helix
0296.5	GroES	6	GroES-like superfamily
0364.4	Leu-IlvD	6	LeuD/IlvD-like
0409.4	GAP	6	GTPase activation domain superfamily
0445.4	SNARE-fusion	6	SNARE-fusion membrane complex superfamily
0458.3	IIaaRS-ABD	6	Class II aaRS Anticodon-binding domain-like
0492.3	S4	6	S4 domain superfamily
0502.3	STAS	6	STAS domain superfamily
0506.3	Succ_CoA_synth	6	Succinyl-CoA synthetase flavodoxin domain superfamily
0545.2	APCOP-app_sub	6	subdomain
0547.2	GF_recep_C-rich	6	Growth factor receptor Cys-rich
0003.21	SAM	5	Sterile Alpha Motif (SAM) domain
0026.20	CU_oxidase	5	Multicopper oxidase-like domain
0052.18	NTN	5	NTN hydrolase superfamily
0065.15	Cyclin	5	Cyclin-like superfamily
0116.13	Calycin	5	Calycin superfamily
0212.9	SNARE	5	SNARE-like superfamily
0226.9	M6PR	5	Mannose 6-phosphate receptor
0260.8	NTP_transf	5	Nucleotidyltransferase superfamily
0274.8	WRKY-GCM1	5	WRKY-GCM1 superfamily
0340.4	PTase-anion_tr	5	Phosphotransferase/anion transport protein superfamily
0466.3	PDZ-like	5	PDZ domain-like peptide-binding superfamily
0542.2	RAS_GEF_N	5	Ras guanyl-nucleotide exchange factor activity N-term
0660.1	SHOCT	5	SHOCT superfamily
0022.32	LRR	4	Leucine Rich Repeat
0035.16	Peptidase_MH	4	Peptidase clan MH/MC/MF
0094.12	Peptidase_ME	4	LuxS/MPP-like metallohydrolase
0118.12	Ribokinase	4	Ribokinase-like superfamily
0187.11	LysM	4	LysM-like domain
0196.12	DSRM	4	DSRM-like clan
0261.7	NUDIX	4	NUDIX superfamily
0291.7	KNTase_C	4	Nucleotidyltransferase substrate binding domain
0350.4	PRC-barrel	4	PRC-barrel like superfamily
0399.4	Asp-glut_race	4	Aspartate/glutamate racemase superfamily
0540.2	GCP	4	Gamma-tubulin complex superfamily
0570.2	PPP-I	4	Protease propeptides/inhibitors
0025.14	His_Kinase_A	3	His Kinase A (phospho-acceptor) domain
0033.14	POZ	3	POZ domain superfamily
0040.17	tRNA_synt_II	3	Class II aminoacyl-tRNA and Biotin synthetases
0051.14	NTF2	3	NTF2-like superfamily
0088.16	Alk_phosphatase	3	Alkaline phosphatase-like
0093.14	Peptidase_CD	3	Peptidase clan CD
0209.11	Bet_V_1_like	3	Bet V 1 like
0272.7	RGS	3	RGS-like superfamily
0301.6	PA14	3	PA14 superfamily
0322.4	RND_permease	3	RND permease superfamily
0336.4	FMN-binding	3	FMN-binding split barrel superfamily
0402.4	Cdc48_2-like	3	Cdc48 domain 2-like
0412.4	Frag1-like	3	Frag1 like
0416.4	Anoctamin-like	3	Transmembrane protein families of the Anoctamin type

0442.4	Tubulin_C	3	Tubulin, FtsZ and Misato and their C-termini,
0448.3	Cargo_bd_muHD	3	Second domain of Mu2 adaptin subunit (ap50) of ap2 adaptor
0530.2	DNase_I-like	3	DNase I-like
0007.18	KH	2	K-Homology (KH) domain Superfamily
0018.15	bZIP	2	bZIP-like leucine zipper
0046.16	Thiolase	2	Thiolase-like Superfamily
0053.15	4H_Cytokine	2	4-helical cytokine superfamily
0066.14	Trefoil	2	Beta-trefoil superfamily
0068.12	RIIa	2	RIIa-like fold
0069.12	GFP	2	GFP-like superfamily
0108.16	Actin_ATPase	2	Actin-like ATPase Superfamily
0139.11	GADPH_aa-bio_dh	2	Amino acid biosynthesis and glycosomal dehydrogenase
0147.11	Traffic	2	Trafficking protein
0163.11	Calcineurin	2	Calcineurin-like phosphoesterase superfamily
0173.11	STIR	2	STIR superfamily
0206.11	TRB	2	Transcriptional repressor beta-barrel domain
0231.9	MazG	2	all-alpha NTP pyrophosphohydrolase superfamily
0233.8	SufE_NifU	2	SufE/NifU superfamily
0277.7	FAD-oxidase_C	2	FAD-linked oxidase C-terminal domain superfamily
0295.7	Vps51	2	Vps51 domain superfamily
0302.6	Arginase	2	Arginase/deacetylase superfamily
0311.4	SCP2	2	SCP-2 sterol transfer superfamily
0316.4	Acyl_transf_3	2	Membrane acyl transferase superfamily
0321.4	PLAT	2	PLAT domain like superfamily
0334.4	THBO-biosyn	2	Tetrahydrobiopterin biosynthesis-like enzyme superfamily
0353.4	TIMP-like	2	TIMP-like superfamily
0356.4	AMP_N-like	2	Creatinase/prolidase N-terminal domain superfamily
0387.4	DHFred	2	Dihydrofolate reductase-like
0422.4	Fibrinogen_C	2	Fibrinogen C-terminal domain-like
0441.4	AlbA	2	RNA-DNA binding Alba-like superfamily
0462.3	TPA-repeat	2	Transcription elongation factor C-terminal nonapeptide repeat
0487.3	FKBP	2	FKBP-like superfamily
0516.3	ISP-domain	2	Rieske-like iron-sulphur domain
0526.3	SUKH	2	SUKH superfamily
0527.3	Sm-like	2	Sm (Small RNA binding protein domain)
0536.2	HEXAPEP	2	Hexapeptide repeat superfamily
0566.2	Tubulin	2	Tubulin nucleotide-binding domain-like, GTPase
0586.2	CTC1	2	CST, telomere maintenance, complex subunit CTC1
0608.2	Reductase_C	2	FAD/NAD-linked reductase C-terminal domain superfamily
0609.2	sPC4_like	2	PC4-like superfamily
0662.1	Triple_barrel	2	Triple barrel superfamily
0006.20	C1	1	Protein kinase C, C1 domain
0009.20	ENTH_VHS	1	ENTH/ANTH/VHS superfamily
0059.15	6_Hairpin	1	Six-hairpin glycosidase superfamily
0061.13	PLP_aminotran	1	PLP dependent aminotransferase superfamily
0067.13	SIS	1	SIS domain fold
0071.12	His_phosphatase	1	Histidine phosphatase superfamily

0076.12	FAD_Lum_binding	1	Riboflavin synthase/Ferredoxin reductase FAD binding domain
0079.13	Cystine-knot	1	Cystine-knot cytokine superfamily
0085.14	FAD_DHS	1	DHS-like NAD/FAD-binding domain
0091.12	NAD_Ferredoxin	1	Ferredoxin / Ferric reductase-like NAD binding
0098.14	SPOUT	1	SPOUT Methyltransferase Superfamily
0100.13	C1q_TNF	1	C1q and TNF superfamily
0101.12	PELOTA	1	Pelota - RNA ribose binding superfamily
0104.13	Glyoxalase	1	VOC superfamily
0106.13	6PGD_C	1	6-phosphogluconate dehydrogenase C-terminal-like superfamily
0121.12	Cystatin	1	Cystatin-like superfamily
0151.12	PK_TIM	1	Pyruvate kinase-like TIM barrel superfamily
0155.11	CBM_14_19	1	Carbohydrate binding domain 14/19 clan
0157.11	Kleisin	1	Kleisin superfamily
0161.12	GAF	1	GAF domain-like
0165.11	Cache	1	Cache-like domain
0171.11	Phospoesterase	1	inositol polyphosphate 1 phosphatase like superfamily
0177.16	PBP	1	Periplasmic binding protein clan
0182.13	IT	1	IT (Ion Transporter) superfamily
0191.11	POTRA	1	POTRA domain superfamily
0194.10	DNA_pol_B-like	1	DNA polymerase B like
0213.9	ShK-like	1	Sea anemone toxin k like
0235.9	PspA	1	PspA/ESCRT-III
0243.10	AEP	1	Archaeo-eukaryotic primase
0247.9	2H	1	2H phosphoesterase superfamily
0251.8	MORN	1	MORN repeat
0255.9	ATP_synthase	1	ATP synthase F0 subunit
0271.7	F-box	1	F-box-like domain
0276.8	Nucleot_cyclase	1	Nucleotide cyclase superfamily
0287.7	Transthyretin	1	Transthyretin superfamily
0290.6	EPT_RTPC	1	EPT/RTPC-like superfamily
0299.5	Peptidase_SF	1	Peptidase clan SF
0303.5	H2TH	1	Helix-two-turns-helix superfamily
0318.4	Cytochrome-c	1	Cytochrome c superfamily
0323.4	Patatin	1	Patatin/FabD/lysophospholipase-like superfamily
0325.4	Form_Glyc_dh	1	Formate/glycerate dehydrogenase catalytic domain-like superfamily
0341.4	LDH_C	1	LDH C-terminal domain-like superfamily
0378.4	ANL	1	ANL superfamily
0382.4	DNA-mend	1	DNA breaking-rejoining enzyme superfamily
0383.4	PheT-TilS	1	Phenylalanine- and lysidine-tRNA synthetase domain superfamily
0391.4	CAP_C-like	1	Adenylate cyclase associated (CAP) C terminal like
0392.4	Chaperone-J	1	Chaperone J-domain superfamily
0395.4	Tubby_C	1	Tubby C-terminal domain-like
0407.4	TBP-like	1	TATA-binding protein like
0413.4	Toprim-like	1	Toprim domain
0417.4	BIR-like	1	BIR-like domains
0437.4	EF-G_C	1	Transcription elongation factor G C-terminal
0475.3	Cyclophil-like	1	Cyclophilin-like superfamily

0476.3	tRNA-IECD_N	1	tRNA-intron endonuclease catalytic domain-like N-term
0484.3	Peroxisome	1	Peroxisome-like domain
0489.3	SAF	1	SAF-like beta clip superfamily
0509.3	RBP11-like	1	RBP11-like subunits of RNA polymerase
0513.3	LCCL-domain	1	LCCL-domain like
0521.3	GOLD-like	1	Sec14-like superfamily of golgi trafficking
0534.2	YjgF-like	1	YjgF-like superfamily
0539.2	RNase_III	1	RNase III domain-like superfamily
0546.2	Hexosaminidase	1	beta-N-acetylhexosaminidase-like domain
0585.2	Nucleoporin_A	1	Nucleoporin superfamily
0599.2	GH57_38_middle	1	Families 57/38 glycoside transferase middle domain superfamily
0603.2	AA_dh_N	1	Aminoacid dehydrogenase-like, N-terminal domain superfamily
0629.1	PLA2	1	Phospholipase A2 superfamily
0642.1	SOCS_box	1	SOCS-box like superfamily
0648.1	Aha1_BPI	1	Aha1/BPI domain-like superfamily
0649.1	PseudoU_synth	1	Pseudouridine synthase superfamily
0664.1	UB2H	1	UB2H/UvrB interaction domain superfamily