Supplemental information #1: A new strategy to characterize the domain architecture structure of proteins of the innate inmune system in tunicate species

all

December 12, 2018

Golden Standard properties

Golden Standard Group ($\mathfrak{G}$) is composed with immune system sequences annotated in immune-specific databases (Insect Innate Immunity Database (IIID) and ImmuneDB). From the first one have been retrieved sequences from 5 species of insects: (*Nasonia vitripennis*, *Apis mellifera*, *Drosophila melanogaster*, *Anopheles gambiae* and *Acyrthosiphon pisum*). For the second one, additional sequences from mouse and human. All the detailed steps to retrieve the immune system sequences are described on the main text. Table 1 describes the number of sequences that have been annotated in those immune system databases. At the same time, complete annotation from the latter species have been retrieved from `Ensembl` in order to compare the percentage of proteins that are annotated as Immune System proteins by the referenced databases. First, all the reported proteins along the immune-system databases represented less than 3% in insect, while in mouse $\sim$ 9.32% and particulary in human a $\sim$ 36.90%. Also, the number of proteins that belong from human are the most frequent in $\mathfrak{G}$, where represent a 84.74% followed by mouse sequences (12.51%), while insects in total reported 2.75%.

Once all the immune system proteins have been retrieved from `Ensembl`, the annotation of domains was organized later to define domain's architectures in to define $\mathfrak{G}$.

The arrangement of those protein domain architecures was performed independently for each annotation database (`Gene3d`, `PANTHER`, `Pfam`, `PIRSF`, `PRINTS`, `Prosite`, `SMART`,`SUPERFAMILY` and `TIGRFAM`) generating the input architectures to be reduced by the *reduction system* (described in the main text). The direct effect performing this reduction of redundancy in the protein's architecture is the reduction of the number of domains inside the proteins in $\mathfrak{G}$ (Figure 1). The distribution of domains along $\mathfrak{G}$ reported in all the protein databases that most of the domains comes from proteins that have been annotated on human. Additionaly, the distribution of domains along all the proteins shown that 75% of all proteins reported architectures with ≤ 2 domains. At the same time, $\mathfrak{G}$ also reported different distributions along the specific protein databases, in this case in the database `SMART` were reported the greatest number of domains inside a protein (152) and also, the database `Prosite` reported similar distributions. In overall, the greatest number of domains could be accessed. As an example in Table 2 is described the proteins that have the biggest number of domains, annotated with the `Pfam` database for each of the Gold Standard species. As expected, the annotation for the protein domains retrieved from `Ensembl` is related (when are avaliable) with the immune system. The annotation in general are related with detection

Table 1: Number of retrieved annotated proteins related to the Immune System. The complete annotation of those proteins have been retrieved from `Ensembl`, including the relationships with their correspondend gene. The column *Ann. Proteins IS* are those proteins that have been reported on Immune system databases (`ImmuneDB` or `IIID`).

Specie	**Complete Ann. Genes**	**Complete Ann. Proteins**	**Ann. Genes IS**	**Ann. Proteins IS**	**%. Prot.**	**Database**
A. pisum	26195	26195	65	81 (63)	0.3092	IIID
A. gambiae	11840	13011	326	333 (326)	2.5594	IIID
A. mellifera	10830	10830	105	106 (104)	0.9788	IIID
D. melanogaster	12315	24557	181	242 (201)	0.9855	IIID
N. vitripennis	13195	13253	350	368 (344)	2.7767	IIID
M. musculus	22090	55419	7043	5160	9.3109	ImmuneDB
H. sapiens	22413	94703	1803	34944	36.8985	ImmuneDB

proteins, like transmembrane receptors, or recognition of specific patterns, as lectins or directly Immunoglobulin domains.

When the protein's architectures in $\mathfrak{G}$ were analyzed along all the species, the distribution and the number of domains were calculated. The complete result is shown in Figure 2, where all the information were analysed indenpendently by protein database and the distribution of protein architectures was calculated and its relation with the domain number. In this analysis were taken into account the number of architectures from 1 because this set represents not only the proteins that have been only one domain but also, these ones that reported repetitions on the same domain and were collapsed after the reduction to one. In this set of 1 domain, is posible to identify architectures that have been detected in a wide range of proteins along the annotation of the Golden Standard species; this pattern was detected for all the databases and also reported architectures that have been present in more than 1832 proteins, as seen for the Immunoglobulins domain (Accession number: `2.60.40.10`). As shown in Figure 3, this domain has been annotated in all of the species in `Ensembl` as: Immunoglobulin-like or Immunoglobulin-like-fold for the Gene3d database. Also, this region has another annotations for the same region by others protein domain databases, even including another protein domains and consequently generating another protein architectures.

In general larger numbers in protein domains are not frequent in proteins in comparison to the lower ones, as seen when the range of the data is compared and also, supporting the pattern observed on Figure 1.

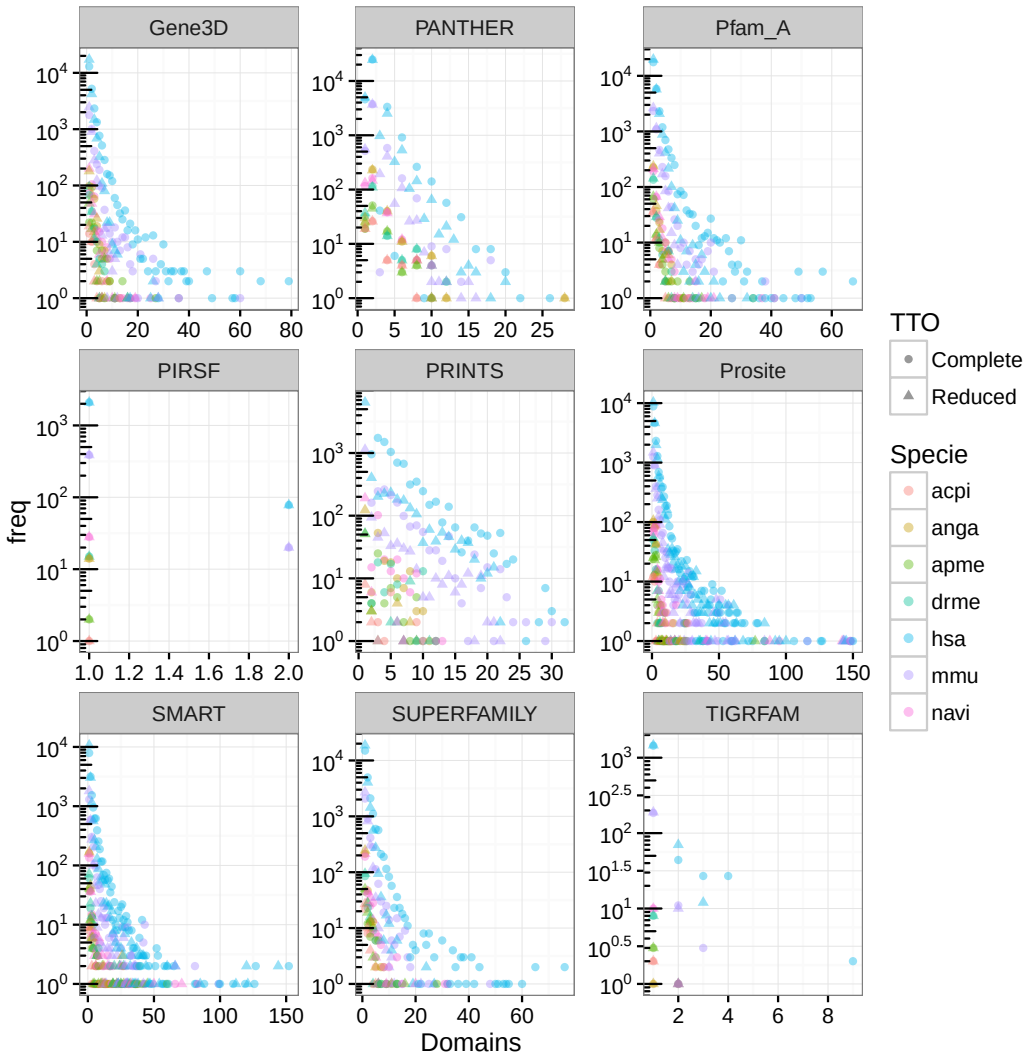

Figure 1: Distribution of the number of domains in $\mathfrak{G}$. **Acpi**: *A. pisum*, **anga**: *A. gambiae*, **apme**: *A. mellifera*, **drme**: *D. melanogaster*, **hsa**: *H. sapiens*, **mmu**: *M. musculus* and **Navi**: *N. vitripennis*.

Table 2: Highest number of protein domains reported in 𝔊 for the Pfam database

Specie	**Numb.**	Pfam **Gene| Protein**	**Most common Domain ACC., Name, (%), No.**	**Annotation**
A. pisum	7	ACYPI008584|ACYPI008584-PA	PF01344, Kelch motif, 71.43,5	NA
A. gambiae	13	AGAP000929|AGAP000929-PA	PF00084, Selectin, 84.61, 11	C-type lectin (CTL) - selectin like
A. mellifera	34	GB47938|GB47938-PA	PF00008, EGF-like domain, 26.47, 9	Gene CTL4
D. melanogaster	27	FBgn0243514|FBpp0301780	PF02363, Cysteine rich repeat, 100.00, 27	Eater is a transmembrane receptor of the Nimrod family specifically expressed in hemocytes and required for the phagocytosis of Gram positive bacteria and the attachment of hemocytes to sessile niches.(http://flybase.org/reports/FBgn0243514.html)
N. vitripennis	41	NV14569|NV14569-PA	PF00008, EGF-like domain, 56.10, 23	NA
M. musculus	50	ENSMUSG00000040249|ENSMUSP00000044004	PF00057, Low-density lipoprotein receptor domain class A, 60.00, 30	LDL receptor related protein 1
H. sapiens	67	ENSG00000154358|ENSP00000455507	PF07679, Immunoglobulin I-set domain, 91.04,61	obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF

Table 3: Highest number of protein domains reported in $\mathfrak{G}$ for the `Pfam` database after the application of the reduction system, as explained in the main text, to generate protein architectures.

Specie	**Number**	`Pfam` **Gene| Protein**	**Most common Domain ACC., Name, (%), No.**	**Annotation**
A. pisum	5	ACYPI007708|ACYPI007708-PA	All domains represented equally (1) in the arquitecture	NA
A. gambiae	7	AGAP007237|AGAP007237-PA	PF07679, Immunoglobulin I-set domain, 28.57, 2	heme peroxidase 4
A. mellifera	17	GB47938|GB47938-PA	PF12661; PF07699; PF07645; PF02494; PF00754; PF00084 and PF00008, 11.76, 2	Gene CTL4
D. melanogaster	14	FBgn0029167|FBpp0075495	PF08742, 35.71, 5	Hemolectin (large multidomain protein produced by hemocytes and involved in the clotting reaction.) `http://flybase.org/reports/FBgn0029167.html`
N. vitripennis	21	NV14569|NV14569-PA	PF00008, EGF-like domain, 28.57, 6	NA
M. musculus	16	ENSMUSG00000027878|ENSMUSP00000078741	PF00008, EGF-like domain, 37.50, 6	Notch 2
H. sapiens	32	ENSG00000197558|ENSP00000485256	PF01826, Trypsin Inhibitor like cysteine rich domain, 40.62, 13	SCO-spondin

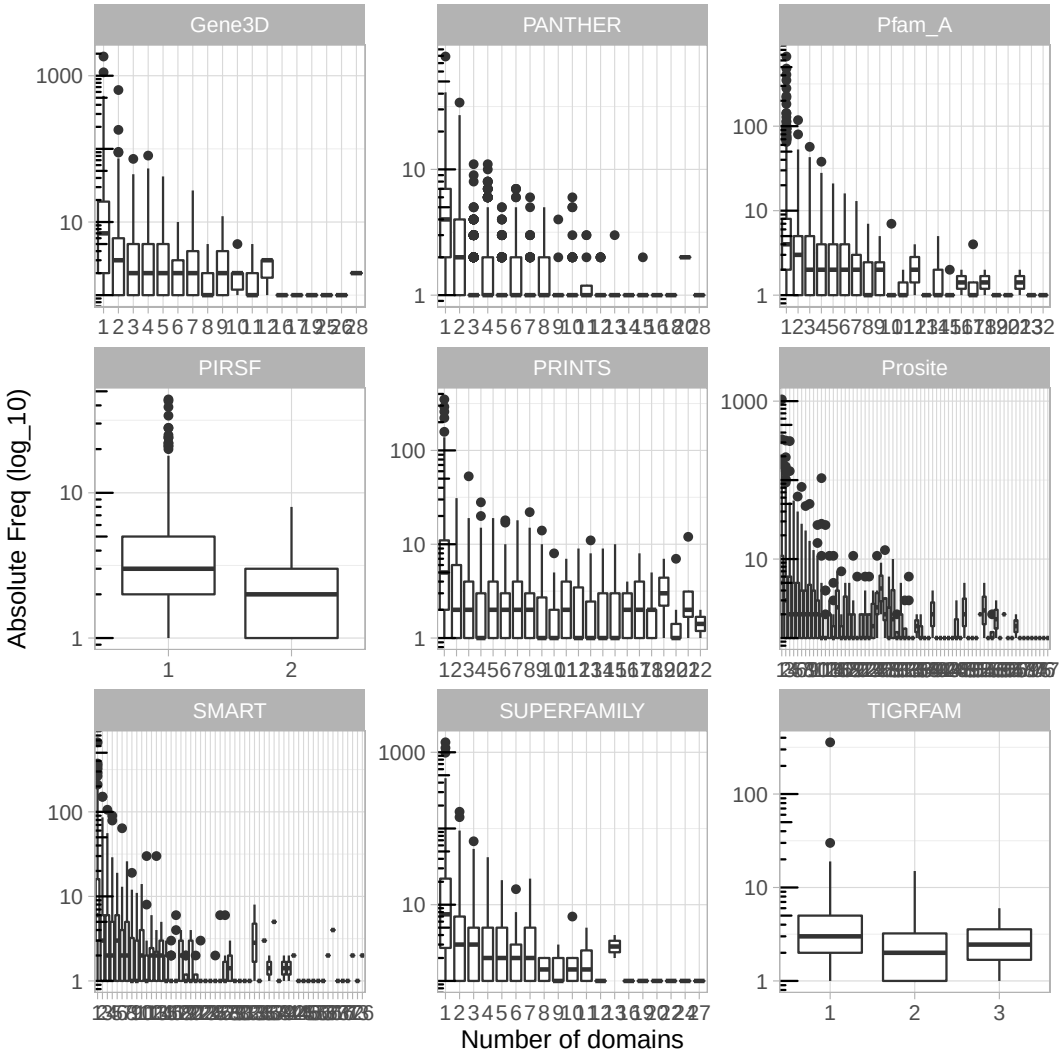

Figure 2: Distribution of protein architectures. The number of domains corresponds to the total number of annotated domains in the protein after the aplication of the reduction process and the Absolute Frequency is the counting of these architectures along all the organisms considered in $\mathfrak{G}$.

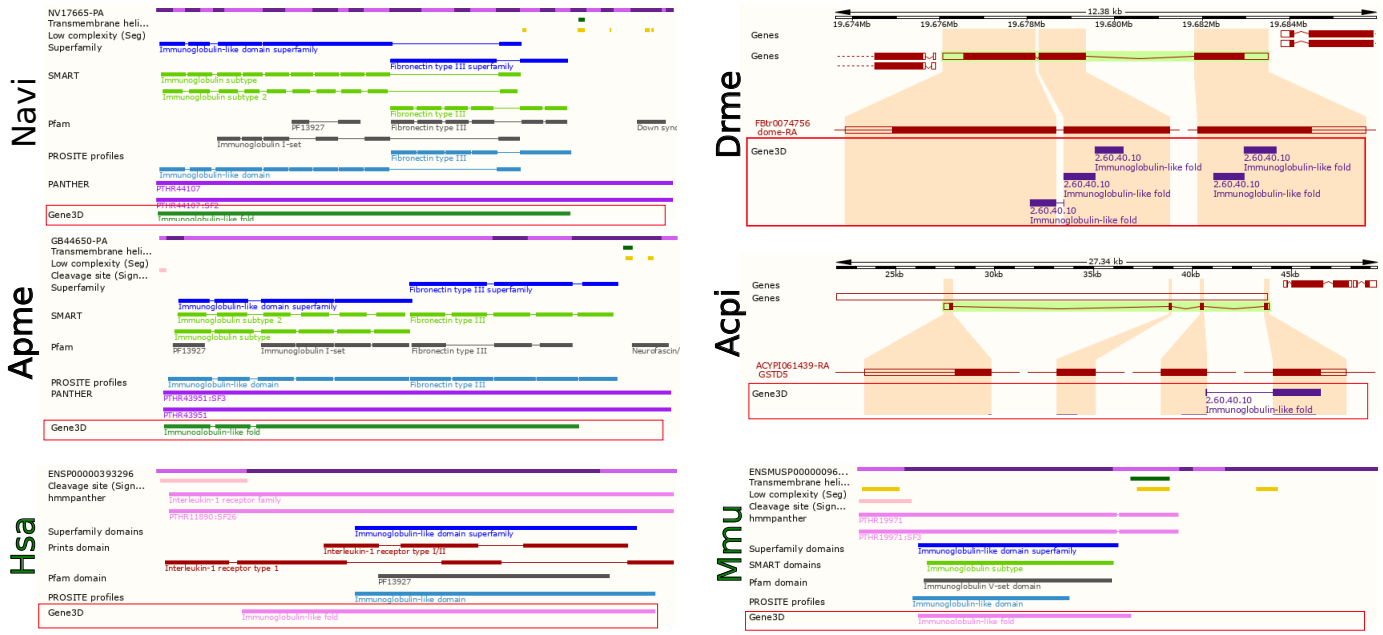

Figure 3: Current annotation of the Immunoglobulin domain (2.60.40.10) along the Gold Standard Species: **Acpi**: *A. pisum*, **apme**: *A. mellifera*, **drme**: *D. melanogaster*, **hsa**: *H. sapiens*, **mmu**: *M. musculus* and **Navi**: *N. vitripennis*. In the current version of `Ensembl Metazoa` the annotation of **anga**: *A. gambiae* for the protein `AGAP010083-PA` have been deprecated and redirected to the gen: `AGAP029053`.