A new strategy to characterize the domain architecture structure of proteins of the innate inmune system in tunicate species

Cristian A. Velandia-Huerto*, Ernesto Parra, Federico D. Brown, Adriaan Gittenberger,
Peter F. Stadler and Clara I. Bermúdez-Santana

January 1, 2019

- Include information about D.vexillum sequencing and assembly process: Clara and Ernesto.
- Change everything related with another databases different to Pfam.
- Discuss about the biological meaning of the different architecture strategies
- If yes, what is the best option to detect protein orthologs. Based on complete protein? or splitting by protein domains?
- Adapt draft to journal template...Which one?.

1 Introduction

In the last years genomes of non-model organisms have become available at a rapidly accelerating rate. As a consequence, their annotation and comparative analysis has become a task limited by time and resource consumption. Gene architectures serve as a convenient and relatively easily accessible source of information about an organism's metabolic and regulatory capabilities, and allow the efficient extraction of candidates for subsequent, more detailed functional or evolutionary studies [1, 2, 5, 27, 28].

Conceptually, gene annotation comprises two tasks: first the identification of genomic subregions that code for proteins, and second the assignment of functionality. Often, both issues are addressed simultaneously, using sequence similarity to identify homologs of a query with known function and the same time using homology as argument to transfer functional annotation. The complexity of eukaryotic genes, with its extensive use of alternative splicing and alternative transcription starts, however, makes the identification of homologs, and in particular the distinction of orthologs and paralogs, a non-trivial, and often surprisingly difficult task [32]. In addition, homology-based annotation is by definition limited to know query sets of sufficiently well-characterized genes, typically from model organisms.

Modern gene annotation pipelines therefore are built around probabilistic models that are trained on known gene structures. The first generation of such tools, such as `GENSCAN` [10] primarily focused on promoter signals, intron/exon boundaries, and polyadenylation signals [12]. State-of-the-art tools, such as `AUGUSTUS` [26] or `GeneID` [6] accept diverse types of training data, in particular RNA-seq based transcript information. The underlying model can be implemented in very different ways: while `AUGUSTUS` is a generalized Generalized Hidden Markov, a rules-based heuristic is used in `GeneID`. In the case of a select few model organisms, gene annotation has evolved into major, long-term data curation projects such as VEGA-HAVANA and GENCODE for the human genome. *** Include a couple of more sentences and a more inclusive list of the major curation projects *** Well curated gene models are in important resource also for training gene models for application to related species.

The annotation of genes and open reading frames is complemented by systematic efforts to establish homology – and in particular orthology – information ref. This in turn is forms the basis for defining a function-based gene nomenclatures cite HGNC, and efforts to achieve a systematic, orthogy-aware nomenclature at least across some important clades ** mention VGNC for vertebrated ** [1] [5]. As a group, vertebrate genomes certainly feature the most thoroughly curated and and most complete functional annotation.

In this work we focus on the sub-phylum Tunicata. As sister group of the vertebrates they occupy a key position in the Tree of Life to understand the prerequisites for key innovations in the vertebrate lineage. Here, we are in particular concerned with the evolution of the immune system just before the "immunology big-bang" [4] that gave rise to the origin of the Adaptive Immune System. Like other invertebrates, Tunicata rely on innate immunity only [19]. However, they feature a great diversity of life-styles and the world widely distribution in ecological niches that may have forced them to evolve different immune responses to ensure survival in their respective habitats. Since tunicates can live as solitary sessile or pelagic or to live in colonies they have complex relationships between the environment, so diversity in the composition of gene of the immune system is expected [3, 11].

Despite the global importance of this group, genomic studies and comparative analyses have remained scarce so far. So far only the genomes of three solitary ascidians have been annotated in substantial depth so far: the sessiles *Ciona savignyi* and *Ciona intestinalis* mapped on its 14 chromosomes [14, 25] and the pelagic *Oikopleura dioica* [16, 24]. More recently, the genome of a single colonial ascidians, *Botrillus schlosseri* (assembled to 13 chromosomes) [31] has become available. The carpet sea squirt *Didemnum vexillum* has been sequenced and analyzed for its ncRNAs [30]. To-date only a very fragmented draft assembly is available, however.

Comparisons between tunicate and other chordate genomes have identified both expansions of gene families but also substantial losses References needed. The genomic organization of tunicates, as exemplified by *Ciona* and *Oikopleura* shows substantial differences compared to both vertebrates and amphioxus, the common outgroup to the Olfactores [15], and has led different authors to formulate the idea of the existence in their evolution of processes of genomic re-structuring in all or some tunicates genomes [23]. since you say "different authors", we need several reference here! While the other chordate lineages have maintained a fairly constant rate of evolution, tunicates feature a systematically accelerated rate of evolution which likely is linked to specific patterns of organization of their entire gene complement [3, 23].

We suspect, therefore, that the chordate immune system also has undergone substantial changes, restructuring, and diversification. As a first step towards understanding the evolution of the chordate immune system we generate her a global overview based on the hypothesis of Paalsson *et al.* [22] that the immune system derives from a small number of ancestral proteins comprising nine ancestral domains. We therefore focus in this survey on protein domains as the most elementary evolutionary building blocks of the immune system and investigate the turnover of domain architectures as a means to capture at a global level the evolutionary driving force that led to the complexity of the immune system in tunicates. In particular, we are interested in the emergence of novel protein architectures throughout the Tunicata. In order to focus on gene families that are likely associated with immune system functions, we consider genes constructed from domains of receptors that are known to be associated with the innate immune system. It is not uncommon in inmune system is not rare to find copies and reshuffling of domains [18] more references. We therefore employ here a domain-based approach to homology search to overcome the limitations of classical homology search schemes in the face of domain-level changes.

Theory

We represent each protein a as an ordered sequence $P(a)$ of domains. In practice the domains depend on one of several annotation systems, a point to which we will return below. Two proteins can then be compared a different levels of stringency:

O What exactly is the match criterion?: Do you need that every domain of the query $Q = (Q_1, Q_2, \dots, Q_n) \in \mathfrak{G}$ matches the domain list $(P_1(a), P_2(a), \dots, P_m(a))$ of a target protein AND that the order is preserved?

D For the writing below you require that the target shares two distinct domains with the query list Q? Is this correct? Or do you mean that you need at least one match for all distinct domains and consider only hits with at least two distinct domains. Again I could not parse your writing below. By construction every match w.r.t. **O** is also a match w.r.t **D**.

B The domain-based comparisons are complemented by a simple `blast`-based sequence comparison between the query protein sequences used to construct Q and the target protein sequence. Parameter settings are described in Materials & Methods.

A I cannot understand how this is different from **O**.

The different annotation systems and comparison methods have been integrated into a single workflow summarized in Fig. **??**. The query $\mathfrak{G}$ comprises for each protein a a domain lists $P_m(a)$ that is specific for each domain annotation system m, that is, one of *Pfam*, *TIGRFAM*, *Superfamily*, *Gene3D* and *Panther*. For details on the construction of the domain lists we refer to Materials & Methods. In the initial step, comparisons are performed separately for the different annotation systems m.

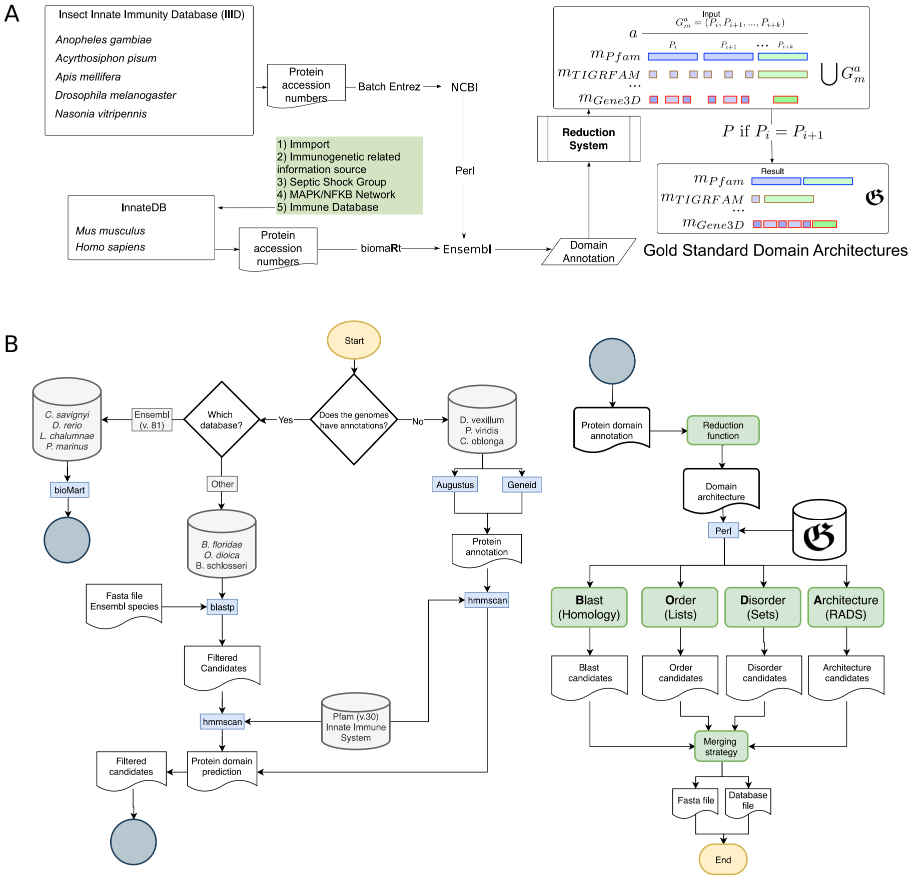

Figure 1: **A.** Workflow to generate $\mathfrak{G}$. Innate immune system databases were used to obtain the accession numbers of the related proteins. Next, domain annotation was accesed throught `biomaRt` from `Ensembl` (v.81). Post processing include reduction of the consecutive repetitions of protein domains and finally, the definition of $\mathfrak{G}$. **B.** Methodological steps to obtain innate immune system candidates based on $\mathfrak{G}$ definition. Used programs from software packages (HHMer and blast) and in-house `Perl` scripts have been highlighted in blue. In green are indicated the `Perl` scripts that perform the reduction function and each step of comparison of architectures (A, B, D, O).

Methods and Materials

Comparison Strategies

List of gene architecture from tunicates and the other chordates were greedily what does greedily mean here? compared with elements from $\mathfrak{G}$ by each m using the followed strategies **O**rder, **D**isorder, **B**last homology and **A**rquitecture or (**O, D, B, A**) respectively.

Reduction system

This subsection need complete rewriting. First the Theory section needs to be cleaned up.

To set out our work, we have defined a reference *gold standard set*. Our survey started building a raw set of domains as follow: let be $G_m^a = (P_i, P_{i+1}, \dots, P_{i+k})$ a sub-sequence of ordered domains P in each protein a of the innate immune system of organisms taking from `InnateDB` and `Insect Innate Immunity Database` (IIID) which have been annotated by m annotation system given by m where m belongs to one of the following sources *Pfam*, *TIGRFAM*, *Superfamily*, *Gene3D* and *Panther*. Since each domain P has a starting s_k and ending e_k point in a, we defined an order $P_i \prec P_j$ if and only if $s_i \leq s_j$. Next, we join all the domains in each protein a by each annotation m as

$$\bigcup G_m^a$$

Since is very commonly found copies of domains in proteins of the immune system, consecutive domains in G_m^a were reduced to a list of unique representative domains P if $P_i = P_{i+1}$. From now on we will refer to this new set as *gold standard set* $\mathfrak{G}$ (Figure 1A).

Protein domain architectures of reference

We started with annotated and curated genes from `InnateDB` [7] and `Insect Innate Immunity Database` (IIID) [8] in order to define a *gold standard* set of domain architectures of proteins of the innate immune system. At `InnateDB` many other immune-specific databases are linked as `Immport`, `Immunogenetic related information source (IRIS)`, `Septic Shock Group`, `MAPK/NFKB Network`, and `Immunome Database`. Our starting point interfaces records from `InnateDB` to `Ensembl` (v.86) by using `Perl` scripts and `biomaRt` R library [17]. In this step, were mostly retrieved accession numbers and sequences belonging to human (GRCh38) and mouse (GRCm38) genomes. Then, to increase the set of gene associated with the innate immune system, the information from the `IIID` was used to obtain data of insects like *Nasonia vitripennis*, *Apis mellifera*, *Drosophila melanogaster*, *Anopheles gambiae* and *Acyrthosiphon pisum*. The latter genomes were chosen because both have annotations on *NCBI* and *Ensembl*. For those cases, genes annotated on `IIID` were retrieved using `Batch Entrez`[1]. Accession numbers from `NCBI` were translated into the accession number of `Ensembl`. Then, we proceed to retrieve the data of insects in a similar way like in human and mouse. A reference set of domains was obtained independently by each domain annotation database after using a *reduction system* described in *Reduction function subsection*1. We used the *gold standard* set for further comparisons of domain architectures of 17 studied species described on Additional File 1.

Re-assembly of *D. vexillum* genome

Include information about re-assembly strategy from D. vexillum, methodological steps...

Genomic data sources

Genomic information source comes from 3 Vertebrata species: *Petromizon marinus*, *Danio rerio* and *Latimeria chalumnae*, 10 species of Tunicata: *Oikopleura dioica*, *Botryllus schlosseri*, *Botrylloides leachii*, *Ciona robusta*, *Ciona savignyi*, *Didemnum vexillum*, *Perohora viridis*, *Clavelina oblonga*, *Molgula occidentalis* and *Molgula oculata*, 1 specie from Cephalochordata: *Branchiostoma floridae*, represents the final set of chosen Chordates. As an outgroup, a set of 2 species from Echinorderms: *Strongylocentrotus purpuratus* and *Patiria miniata* and additionally 1 Hemichordate specie: *Saccoglossus kowalevskii* were studied. The protein database sources are described in Additional File 1.

the following is rather incomprehensible. The concepts need to be described in the Theory section. see there.

[1] `https://www.ncbi.nlm.nih.gov/sites/batchentrez`

Architecture Comparison Strategies

Order comparison

To trace back similar architecture organizations between annotated genes in tunicates with the architectures in the *gold standard set*, tunicate domain architectures were represented as query sets $Q_m^a = (\mathcal{P}_k, \mathcal{P}_{k+1}, \ldots, \mathcal{P}_n)$ and are defined as a sub-sequence of ordered domains $\mathcal{P}$ in each protein a by m like was previously defined. Comparing the order between P_is and $\mathcal{P}_i$s we defined the number $Q_m^a success(o)$

$$Q_m^a success(o) = \begin{cases} 1, & \textit{if } P_i \leq \mathcal{P}_k \\ 0, & \textit{otherwise} \end{cases} \tag{1}$$

If $Q_m^a success(o) = 1$ we say that exist in $\boldsymbol{\mathfrak{G}}$ an architecture organization preserving order equal to an architecture organization Q_m^a. If $Q_m^a success(o) = 0$ then we say those architectures are not related.

Disorder comparison

Since rearrangements of domains are also expected we used a second more flexible comparison between elements of Q_m^a and $\boldsymbol{\mathfrak{G}}$ without considering order in $\mathcal{P}$ domains. Now the rules are defined as follows:

$$Q_m^a success(d) = \begin{cases} 1, & Q_m^a \subseteq \boldsymbol{\mathfrak{G}} \quad \textit{and} \quad |Q_m^a| \geq 2 \\ 0, & \textit{otherwise} \end{cases} \tag{2}$$

If $Q_m^a success(d) = 1$ we say that exist in $\boldsymbol{\mathfrak{G}}$ an architecture composition similar to an architecture organization Q_m^a. If $Q_m^a success(d) = 0$ then we say those architectures are not related. Note that here the order of domains is not a constrain to classify a query set Q_m^a as success.

Blast homology comparison

A classical homology strategy with `blastp` was used [20]. For these homology searches, pairwise comparisons were done between the proteins used to built both query and *gold standard* sets. After running BLAST following the combination of parameters:

```
blastall -p blastp -d <DB> -i <QUERY> -f 9 -F 'm S' -M BLOSUM45 -e 100 -b
10000 -v 10000 -m 8
```

were filtered candidate homologous if they satisfied:

- E-value ≤ 0.001.
- Coverage to query length ≥ 60 %.
- Identity $\geq 30\%$.

Arquitecture comparison

Before the application of reduction system, there are different architectures composed by only one domain that had not been taken into account with the O,D,B strategies. In order to complement the search strategies, a comparison between *gold standard* architectures and query architectures was performed applying the methodology reported by `RADS`[2] [29]. First, a domain architecture database was created with the *gold standard* domains using the program:

```
makeRadsDB -i <DOMAIN_DISTRIBUTION1> <DOMAIN_DISTRIBUTION2> -s
<Seq_fasta_1> <Seq_fasta_2> seqFile2.fa -o <OUT_DB>}.
```

And the comparison was applied against all the query architectures with:

```
rads -c -d <DB> -m <Matrix> -q <Input> -o <OUT_FILE>
```

[2] `http://domainworld.uni-muenster.de/programs/rads/`

Where DB, corresponds to the output file OUT_DB from makeRadsDB. Matrix file (pfam-30.dsm) was obtained directly from RADS site; Input file correspond to the domain's distribution organised as pfam_scan output file. Final candidates were retrieved if reported a similarity normalized score ≥ 0.75.

Merging output of comparison strategies

In order to identify the best candidates to be related with the immune system, all the previously results from Order, Disorder, Blast and Architecture strategies were merged and combined with a Perl script. Candidates that have been detected only by Blast (B) strategy were not taken into account. Considering all other possible combinations of strategies, it is important to note that $O \subset D$, it means that combinations as O, OB, OA, OBA are not possible. In this way the remaining combinations 10 were considered to detect the candidates for the innate immune system.

Cleaning specific Hidden Markov Models (HMMs) for each domain

Specific Hidden Markov Models (HMMs) for each domain on the different annotation sources were obtained using the program hmmfetch by screening on Interpro (Version 60). Then HMMs related with innate inmune system in $\mathfrak{G}$ were retrieved. The final list was used in further steps.

Screening of architectures domains in Ensembl-non-annotated tunicate and cephalochordate species

Ensembl-non-annotated genomes

The protein annotation for the cephalochordate *B. floridae* and the tunicates *O. dioica* and *B. schlosseri* are based on the scheme reported at JGI Genome Portal (http://genome.jgi.doe.gov/Brafl1/Brafl1.download.ftp.html) for *B. floridae*, Oikoarrays (http://oikoarrays.biology.uiowa.edu/Oiko/Downloads.html) for *O. dioica* and ANISEED database (http://www.aniseed.cnrs.fr/aniseed/download/download_data) for *B. schlosseri*. In first place, candidates related to the immune system in the species *C. robusta*, *C. savignyi*, *L. chalumnae*, *P. marinus* and *D. rerio* were used as query sequences to perform pairwise homology searches with blastp.

```
blastall −p blastp −d <DB> −i <QUERY> −F 'm S' −m 8 −o <OUT_FILE>
```

After filtering, hits candidates with high level of similarity were considered as a set of putative candidates, as described below:

- E-value ≤ 0.001
- Coverage ≥ 60 %
- Identity against query $\geq 30\%$.

After that, an exhaustive search of HMM domains was conducted by the suite HMMer to detect domains using the mapped HMMs in $\mathfrak{G}$. Best candidates to annotate protein domains derived from PFAM was obtained filtering all of the candidates that reported a bitscore $\geq$ Gathering cut-off from Pfam (v.30) and reported an internal i-E-value and c-E-value ≤ 0.01. For the inference of domain architecture in these proteins, previously described approach had been applied, including the comparison against the *gold standard* set. The O, D, B and A strategies were merged generating the candidates that were overlapping between all the strategies, as described on Figure 1.

Draft genomes without annotation

For the recently reported draft genomes of *C. oblonga* and *P. viridis* a *de novo* gene prediction was performed directly on the assembled contigs using GeneID[6] with the following parameters:

```
geneid −3 −P <Parameter file> <FASTA FILE> −A >> <GFF3 file>
```

Here the *Parameter file* was fetched via `FTP` for the tunicates: *C. intestinalis*[3] and *O.dioica*[4]. The final result was a `GFF3` file describing the coordinates on the candidate genes, and additionally the set of possible protein candidates in a `fasta` format. Over those candidates `HMMer` was ran to detect domains that intersect with the mapped HMMs in $\mathfrak{G}$. Again, filters by each m was used, the *Reduction system*1 step and comparison strategies **O, D, B, A** were done (Figure 1).

For the draft genome of the carpet sea squirt *D. vexillum* [30] a *de novo* gene prediction was performed directly on the assembled contigs using `AUGUSTUS` [26] with the following parameters:

Put AUGUSTUS parameters

and the complete prediction of homologous architectures was obtained applying the pipeline Name of my pipeline!.

Orthology detection between candidate innate immune system proteins

In order to detect orthologous groups among innate immune system candidates, proteins that reported the same architecture relationships respect to the gold standard proteins, were compared with `ProteinOrtho` (v.5.16) [21], as follows:

```
proteinortho5.pl -force -graph -clean -keep -project=<name-project>
-step=1 <fasta files>
proteinortho5.pl -force -graph -clean -keep -step=2
-project={name-project} <fasta files>
proteinortho5.pl -force -graph -clean -keep -step=3
-project=<name-project> <fasta files>
```

Detected 1:1 or co-orthologous relationships between pre-defined groups of orthology were obtained with `Perl` script. At the same time, available annotation from $\mathfrak{G}$ were obtained from `Ensembl` using `biomaRt`. For protein candidates that belongs from studied species and shared orthologous relations with a $\mathfrak{G}$ protein, the retrieved annotation from from `Ensembl` and `Interpro` accession numbers were associated and reported.

Gain and losses of domains

Derived from our orthologous relationships we obtained raw and family-specific presence/absence matrices of domains. The combined presence/absence matrices were subjected to analysis with Count [13], reconstructing the family history by Dollo parsimony. The phylogenetic distribution of this species were obtained from [71] for tunicates, and for the other organisms from Ensembl compara [72].

Results

Global distribution of domains

A total of 8846 genes associated with the innate immune system were recovered from `InnateDB`, of which 7043 and 1803 belong to human and mouse respectively. After interfaced these records with Ensembl Genome Browser a total of 35136 and 5179 proteins were identified. Next, to integrate domains from the source `Insect Innate Immunity Database` (IIID) a total of 1312 proteins were recovered distributed as follows: *N. vitripennis* 393 (368), *A. mellifera* 170(106), *D. melanogaster* 298(242), *A. gambiae* 366 (333) and *A. pisum* 85(81), the number in parenthesis corresponds to the *bone fide* annotation in `Ensembl`. Finally the domain structure was traced back with `Biomart` in `Ensembl`. With the final set of domains we built the *gold standard* set $\mathfrak{G}$ as described in Additional File 2.

[3] ftp://genome.crg.es/pub/software/geneid/cintestinalis.param_Apr_26_2006

[4] ftp://genome.crg.es/pub/software/geneid/odioica.param_Nov_10_2006

ABDO strategy comparisons of domains

The distribution of genes and proteins that belong from *gold standard set* $\mathfrak{G}$ is shown on Additional File 2: Table 1, where most of the current innate inmune system proteins belongs from human (84.74%) and mouse (12.51%). In order to compare $\mathfrak{G}$ to the query species, the previously explained **ABDO** strategies were applied. As shown in Table 1, exists 4 different groups of species: those ones that have been annotated in `Ensembl` as: *C. robusta*, *C. savignyi*, *P. marinus*, *D. rerio* and *L. chalumnae.* Also, those ones that have gene and protein annotations but, in an independent databases and without the prediction of domains, as: *B. floridae*, *B. schlosseri* and *O. dioica.* The other third one is composed by those genomes that have a *de novo* assembly, as: *D. vexillum*, *C. oblonga* and *P. viridis*, which did not reported predictions of genes or proteins, then this prediction were performed as described in Methods and Materials with the `GeneID` program using the previously constructed gene models from *C. robusta* and *O. dioica*[5] and with `AUGUSTUS` for *D. vexillum.* Finally, the last group is composed by the outgroup species from hemichordates: *S. kowalevskii* and from echinoderms: *P. miniata* and *S. purpuratus*, where the **ABDO** predictions have been calculated applying the automated pipeline NAME_PIPELINE generated in this study.

At the same time, Table 1 summarizes the final innate immune system candidates, based on homology architecture strategies (**ABDO**), which have been reported independently, too. The final number of innate immune system proteins set is described on column **Total Prot. ISS**, which was obtained after merging steps on the ABDO results.

When all the strategies were applied, always the B strategy reported higher frequencies of protein candidates in comparison to the O or D; and those results are more similar to the A strategy, which does not have a previous reduction step. Then, those results show a highest distribution from annotated immune system proteins in vertebrates (median = 63.49% ±4.87, n=3), in comparison to tunicates (median = 31.88% ±22.27, n=12), cephalochordates (43.14%, n=1), and the outgroup composed by hemichordates (42.89%, n=1) and echinoderms (median = 45.15% ±11.56, n=2). High standard deviation in tunicates are a consequence of the inclusion of the new draft genomes (from *C. oblonga* and *P. viridis*), where the prediction of genes was *de novo* by `GeneID`. In this context, only considering the tunicates genomes that had a previous annotation, the estimated values changed: (median =40.21% ±13.37, n=8).

Due the application of ABDO strategies was independently, it is possible to identify the relationships between the query species and the $\mathfrak{G}$ species in the final set of immune system candidates through an architecture relationships. As shown in Figure 2A different combinations of possible architecture comparison strategies have been merged to four different sets, according to the number of (ABDO) strategies that reported a successfully architecture comparisons. Here, $1 = (A, D)$; $2 = (AB, AD, BD, DO)$; $3 = (ABD, ADO, BDO)$ and $4 = (ABDO)$; that is 10 from 15 possible combinations, because (AO, BO) always map to (ADO, BDO), due $D \subseteq O$ and additionally, B has not been considered at all because this comparison does not represent a pure architecture relation, due complete sequence is used in the pairwise alignment against $\mathfrak{G}$ protein sequences. At the same time, thist described distribution of relationships against $\mathfrak{G}$ proteins shows a highest number of proteins that have been detected as innate immune system candidates by only 1 architecture strategy. Additionally, the number of candidates that have been detected with all the strategies (set 4) reported frequencies ≤ 5000 proteins on the subject species, reporting the highest distribution on the specie *D. rerio.* This group represents, along all the comparisons, the most conserved set of immune proteins respect to the defined $\mathfrak{G}$. In this approach, it is important to point out that query proteins could report different ABDO strategies against $\mathfrak{G}$, showing that relationships between query proteins set (defined as Q with m elements) and $\mathfrak{G}$ proteins set (defined as G, with n elements) could not defined as a function $f(x)$ with comparison strategies $S = (A, B, D, O)$ because it does not follows: $m_1 \neq m_2 \rightarrow f(m_1) \neq f(m_2)$.

Considering in overall the immune system proteins along all the studied species, it is possible to identify the proportion of shared innate immune system proteins respect to $\mathfrak{G}$ species (Figure 2B). Mainly those relationships have been detected with human and mouse proteins, as described earlier those $\mathfrak{G}$ species reported the most frequent proteins in $\mathfrak{G}$ set. Delete Ciin of the plot!.

[5] For those genomes, in Table 1 are referenced as *Ciro* and *Oidi* in parenthesis, respectively

Specie	**Annotated Genes**	**Annotated Proteins**	**Annotated Prot. with Pfam domains**	**Ordered Prot**	**Disorder Prot**	**Blast Prot**	**Architecture**	**Total Prot IIS**
P. miniata	30399	30399	20192(66.42)	1936(6.37)	2527(8.31)	11577(38.08)	15707(51.67)	16210(53.32)
S. purpuratus	33663	35786	23640(66.06)	3248(9.08)	4218(11.79)	15420(43.09)	10706(29.92)	13230(36.97)
S. kowalevskii	32367	22111	14888(67.33)	1973(8.92)	2571(11.63)	9737(44.04)	8280(37.45)	9483(42.89)
B. floridae	50817	50817	25430(50.04)	5499(10.82)	7352(14.47)	21767(42.83)	5496(10.82)	21920(43.14)
O. dioica	17212	17212	5709(33.17)	1342(7.80)	1633(9.49)	4577(26.59)	4760(27.66)	5065(29.43)
M. occidentalis	30639	33023	13050(39.52)	1195(3.62)	1486(4.50)	7170(21.71)	11152(33.77)	11341(34.34)
M. oculata	15313	16616	9985(60.09)	1336(8.04)	1689(10.16)	6615(39.81)	8355(50.28)	8620(51.88)
B. schlosseri	46519	46519	8709(18.72)	1790(3.85)	2520(5.42)	6148(13.22)	6760(14.53)	7316(15.73)
B. leachii	15839	15839	9833(62.08)	1271(8.02)	1698(10.72)	6243(39.42)	8032(50.71)	8369(52.84)
C. robusta	17153	17302	8994(51.98)	1371(7.92)	1584(9.15)	6005(34.71)	4188(24.21)	4958(28.66)
C. savignyi	12172	20157	14016(84.84)	2087(10.35)	3273(16.24)	10049(49.85)	10074(49.98)	10923(54.19)
P. viridis (Ciro)	6077	2221773	2806(0.13)	56(0.00)	65(0.00)	12724(0.57)	1896 (0.09)	1865(0.08)
P. viridis (Oidi)	3025	1811030	2110(0.12)	60(0.00)	72(0.00)	10329(0.57)	1352 (0.07)	1319(0.07)
D. vexillum	26546	72326	36075(49.88)	2920(4.04)	4136(5.72)	16889(23.35)	26400(36.50)	27349(37.81)
C. oblonga (Ciro)	19507	1174882	4032(0.34)	120(0.01)	161(0.01)	4070(0.35)	2828(0.24)	2810(0.24)
C. oblonga (Oidi)	4832	950470	2856(0.30)	125(0.01)	164(0.02)	8746(0.92)	1957 (0.21)	1939(0.20)
P. marinus	13114	11444	9214(80.51)	1650(14.42)	2143(18.73)	6227(54.41)	6023(52.63)	7008(61.24)
D. rerio	31953	44489	38629(86.83)	11762(26.44)	13654(30.69)	28031 (63.01)	20992(47.18)	31395(70.57)
L. chalumnae	22628	23603	19509(82.65)	4461(18.90)	5824(24.67)	9127(38.67)	10765(45.61)	14986(63.49)

Table 1: Final distribution of annotated genes and found candidate Innate Immune system proteins. The percentage in relation of the total of annotated proteins (column *Annotated Proteins*) is reported in parenthesis. In last column **IIS** refers to: Innate Immune system.

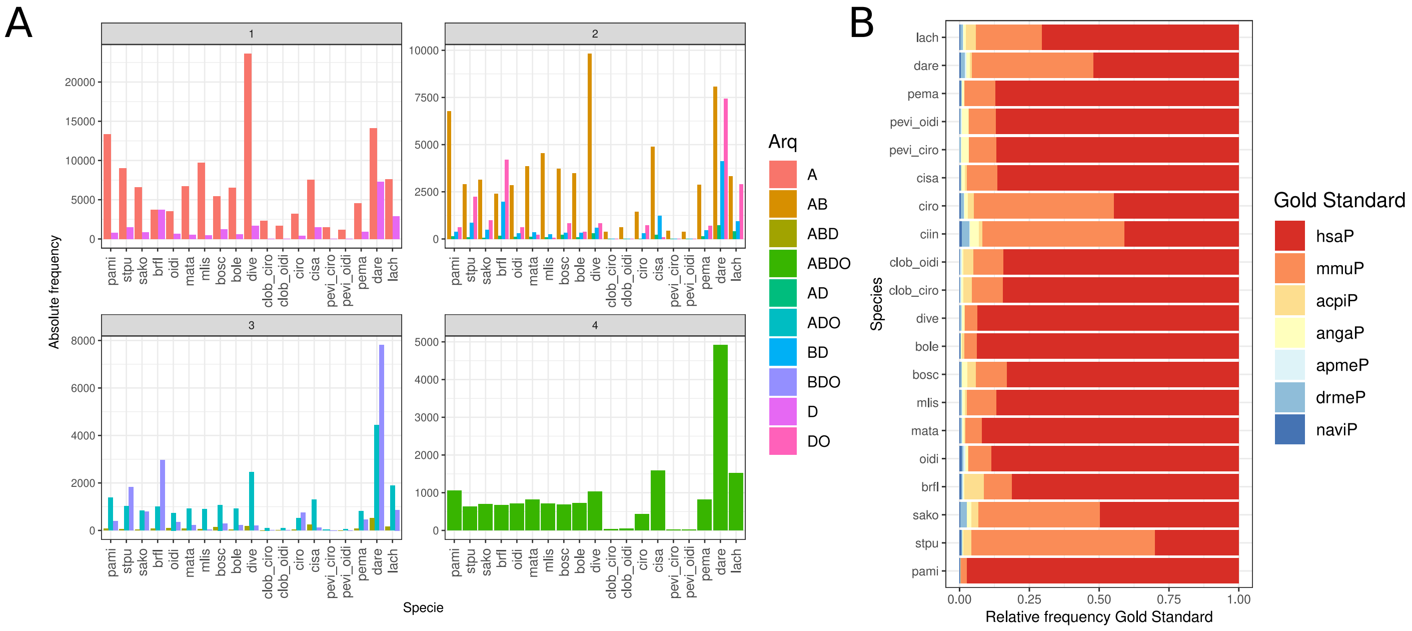

Figure 2: **A)**Frequency of detected proteins with defined architecture comparison strategies classified according to the number of possible combinations of architecture strategies (described in more detail in main text), against $\boldsymbol{\mathfrak{G}}$. **B)** Mean shared proportion homology architecture against gold standard species. naviP=*N. vitripennis*, apmeP=*A. mellifera*, drmeP=*D. melanogaster*, angaP=*A. gambiae* and acpiP=*A. pisum*; and Mammals: mmuP=*M. musculus* and hsaP=*H. sapiens*.

Relationships between Innate Immune system candidates

Obtaining the protein domain architectures after their detection by ABDO strategies, the distribution of the number of domains was studied (Figure 3A). All the studied organisms were grouped in 5 defined taxonomical clades as follows: **Echi** = Echinodermata, **Hemi** = Hemichordata, **Ceph** = Cephalochordata, **Tuni** = Tunicata and **Vert** = Vertebrata, following the sub-phylum assignment on Additional File 1:Table 1. It was possible to identify a high number of proteins that only reported 1 protein domain type (homodomain proteins). Despite the similarity of this results along all clades, distribution from *B. floridae* (Cephalochordata) showed also higher number of proteins with 2 and 3 types of domains with similar numbers as reported on vertebrates; but not in echinoderms, hemichordates and tunicates, that showed similar density distributions. Also, few proteins are composed by more than 5 different types of domains, but the current distributions shows, in all clades, a very long heavy tails with $\geq$ 20 domains types.

Meanwhile, Figures 3B and C, shows the distribution of protein architectures on homodomain and heterodomain proteins along studied species, respectively. For homodomain proteins, 2304 domain architectures were detected along innate immune proteins in echinoderms (54.1%), hemichordates (51.3%), cephalochordates (37.0%), tunicates (43.1%) and vertebrates(52.9%). This distribution also shows a high variation in total number of architectures of one type of domain in echinoderms and tunicates, respect to vertebrates (an excluding Hemichordates and Cephalochordates that have $n = 1$). Please, explain this with a more detailed plot or data for tunicates and echinoderms..

In this homodomain set, the most frequent protein architecture with current annotation from `Pfam` database is: Rhodopsin-like receptor (PF00001) for echinoderms, hemichordates and vertebrates, while for Cephalochordata was Cytochrome P450 (PF00067) and for tunicates the Protein Kinase domain (PF00069).

A number of 1936 heterodomain architectures have been detected along the set of proteins. Specifically, for all clades, was possible to calculate the percentage of found architectures respect to the total number of heterodomain architectures for all clades: echinoderms (42.3%), hemichordates (45.0%), cephalochordates (47.1%), tunicates (22.5%) and vertebrates(50.8%). In more detail, Figure 3C shows the average of found protein architectures along all the clades. Both, the presence percentage and the average number of architectures shows a reduced number of found architectures on tunicates proteins, in comparison to other species of chordates and the outgroup species. Also, a high variability is evident for the complete Tunicata clade. Maybe make a stat test?, please make a plot!.

As a complement for heterodomain proteins, Figure 3D represent the top 5 proteins architectures domains along all defined clades. This distribution is dominated by only 10 domains that are spanned along the protein architectures. By this way, the most frequent domain is an Immunoglobulin domain (Ig 3, PF13927), in vertebrates, tunicates and echinoderms. In relation with this domain, in echinoderms and hemichordates a frequent domain is the Immunoglobulin I-set domain (PF07679). Leucine-rich repeats (PF13855), related with protein-protein interactions are frequent in vertebrates and cephalochordates and additionally related with this funcion, Ankyrin repeats (PF12796) and the Calcium-binding EGF domain (PF07645) have been detected in this list. The most conserved along those clades are: P-kinase (PF00069) and related domains as: Tyrosin kinase (PF07714), and even Pleckstrin homology domain (PF00169).

When the architecture distribution is considered, the most frequent along all species are described on Table 2. Change this table describing by specie, clade or even describe the matrix based on 1:1 orthologous proteins.

Table 2: Most conserved architectures along studied species. Maybe reeplace this table for one more complete table with the most conserved along species? Additional file?

Architecture	**Annotation**	**References**
PF00134,PF16899	Cyclin_N,Cyclin_C_2	?
PF00651,PF07707	BTB,BACK	`https://www.ncbi.nlm.nih.gov/pubmed/15544948`
PF00688,PF00019	TGFb_propeptide,TGF_beta	`https://www.sciencedirect.com/science/article/pii/S0145305X03001812?via%3Dihub`
PF04851,PF00271	ResIII,Helicase_C	?
PF07707,PF01344	BACK,Kelch_1	?
PF15227,PF00643	zf-C3HC4_4,zf-B_box	?

Once protein candidates were detected by ABDO strategies on studied species, a further step is the identification of the biological relevance of the new detected protein candidates. Applying a clustering strategy was possible to identify the intrinsic relationships between $\mathfrak{G}$ and the new identified proteins, based on common protein domain architectures. In this way 4240 groups were created and to access for orthology relationships with described methodology with `ProteinOrtho`. The first step to study those relations is the identification of orthologous (1:1) or co-orthologous (1:many or many:many) groups. In this way, a total of 27311 relations were detected, from them: 53.13% reports 1:1 relationships, while co-orthologs are represented by 46.87%. Based on the 1:1 orthology relations and their relationship with protein architectures a matrix was generated with the number of proteins for each specie. In this way, was identified 2740 architectures related with 1:1 relations and was possible to identify 38 architectures that are conserved in all species (Describe all at new Supp. File?). Most of this conserved set of architectures corresponds to homodomain proteins, except for 3 heterodomain architectures (PF00134:PF16899,PF04851:PF00271 and PF07707:PF01344). At the same time, based in the same matrix it is possible to reconstruct the architecture domain history along all studied species by Dollo parsimony [] with `Count`[13].

As represented on Figure 4, a set of 1477 architectures are shared in the base of the Deuterostomia? and in both clades: chordates and ambulacrarians report only gains (g) events. In this way, at the base of Chordata 1619 architectures have been partially loss (l) in Cephalochordata (l:986) in comparison to Olfactores (l:20, g:276) that reported an additional set of gained architectures. In the divergence between *Tunicata* and *Vertebrata*,

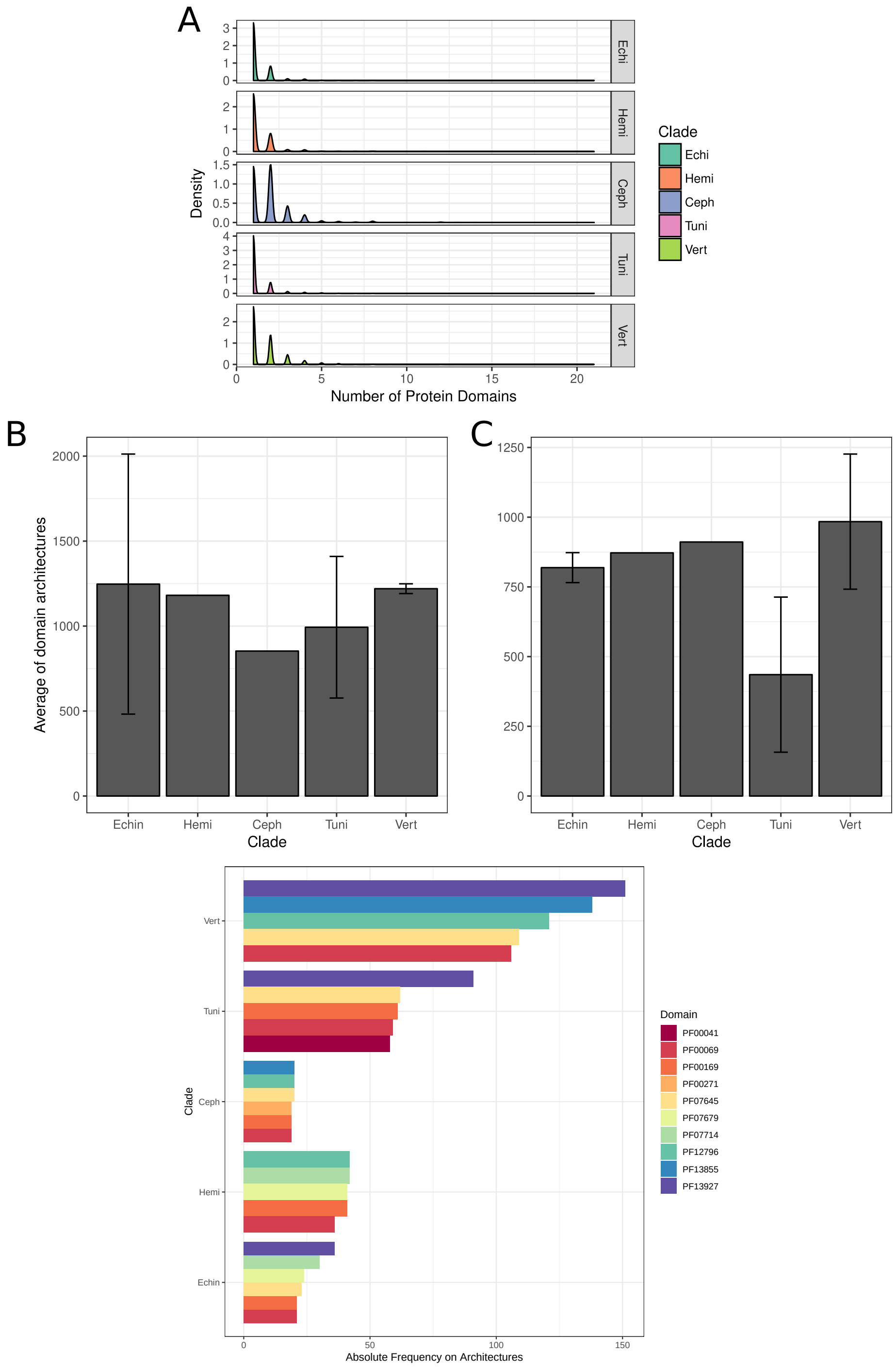

Figure 3: **A)** Domain distribution of innate immune system proteins along studied bilateria species. **B)** Homodomain architecture distribution. **C)** Multidomain architecture distribution. **Echi:** Echinodermata, **Hemi:** Hemichordata, **Ceph:** Cephalochordata, **Tuni:** Tunicata and **Vert:** Vertebrata.

both clades reported specific gains and losses, but with a higher number of losses that could be traced in the base of vertebrates (l:483, g:62). Gain and loss in tunicates could be traced (g:18,l:145), but not at the same magnitude as occurred in vertebrates, with 1748 architectures. More than 70.25% of those architectures have been lost in *O. dioica* (Appendicularians) and also reporting a very few number of gains (g:7). The clade that groups *Stolidobranchia, Phlebobranchia and Aplousobranchia* increased the total number of architectures. In comparison, more loss events have been detected in the clade (Phlebobranchia + Aplousobranchia) (g:17, l:386) than Stolidobranchia (g:34, l:232). In overall through those clades is important to note that Aplousobranchia reported between almost twice loss events (g:6,l:523) than Phlebobranchia (g:2, l:283) and Stolidobranchia (g:34, l:232). At the same time, exists a high number of lost architectures in species where *de novo* gene prediction was predicted (with gene models from *C. robusta* and *O. dioica* using `GeneID`). At the end, final numbers of orthologous architectures are reported for each specie. Specie-specific or clade-specific gains resulted for *P. miniata* the highest number of specific architectures. In terms of numbers, Stolidobranchia reported, in average, biggest values of architectures (972.5) than Aplousobranchia (634) and Phlebobranchia (617), despite the high number of gains registered on *C. savignyi* (g:37, l:284), in general at the base of the Stolidobranchia happened a similar gain event (g:34, l:232). Finally, inside vertebrates the highest number of architectures are present on *D. rerio*, not only the specie-specific gain events (g:86,l:319), but also gains at the base of Osteichthyes (g:102, l:178).

As mentioned earlier, is possible divide the architecture domains into: homodomain and heterodomain. Figure 3A shows that homodomain proteins are the most frequent in all clades in comparison to heterodomain proteins, except for cephalochordates. In order to analyze the evolutionary history of heterodomain orthologous proteins (1:1), the evolutionary history reconstruction using Dollo parsimony was applied as described earlier (Figure 5). A 437 of 1133 architectures were shared between Deuterostomata, again showing gain events at Ambulacraria (g:29) and Chordata (g:113). Cepahlochordata lost about 42% of the ancestral architectures for chordates, while in Olfactores predominate gain events (g:130, l:9). The ancestral number of heterodomain architectures in Tunicata are 595, specifically in *O. dioica* report the most higher loss of domains 73.9%. At thre same time, Stolidobranchia shows higher heterodomain architectures in comparison to Phlebobranchia and Aplousobranchia. A number of 526 heterodomain architectures are at the base of vertebrates, and about $48,8\%$ are lost in *P. marinus* (g:8, l:257). In contrast, Osteichthyes reported few loss and 10 times gain events (g:80, l:30). At the end, not only in vertebrates, but in all analyzed species, *D. rerio* report the biggest number of heterodomain architectures (525).

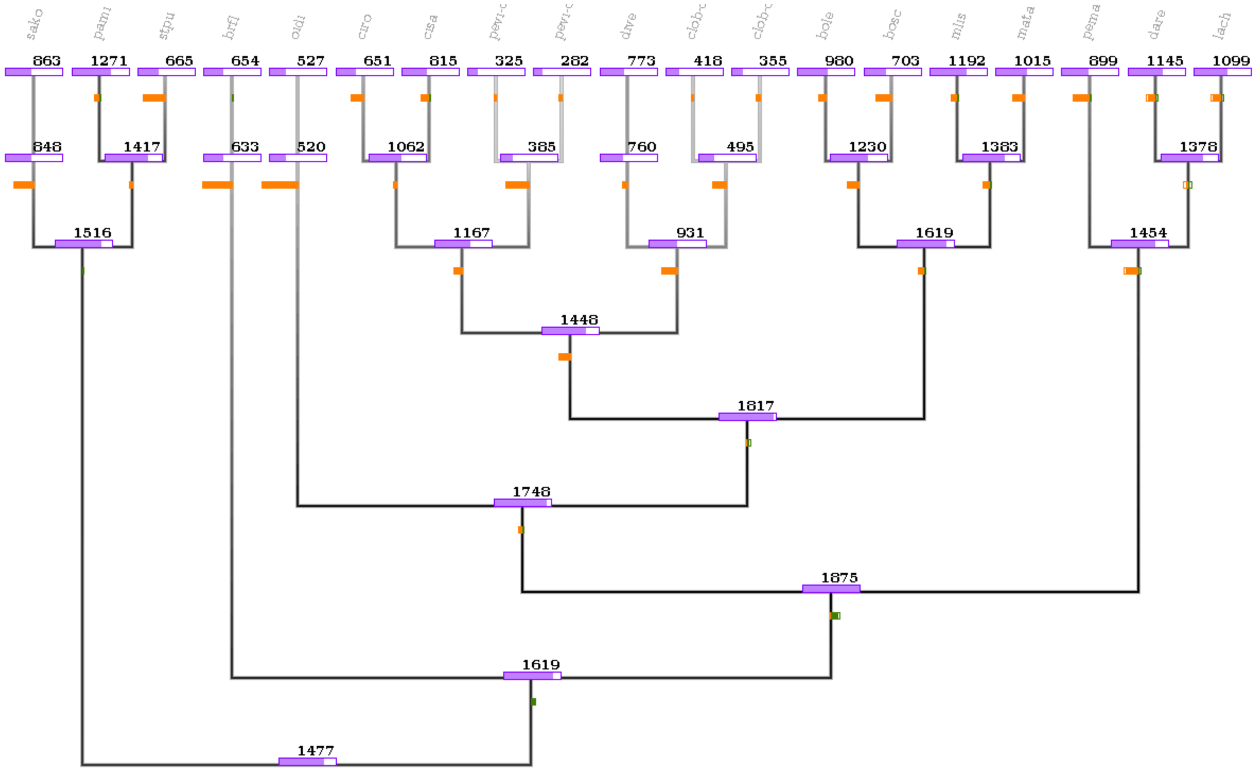

Figure 4: Evolutionary history of 1:1 orthologous protein architectures in Deuterostomata. This is a temporal image, while I'm sure about numbers

In the same way, from the 27311 relations that have been detected, those ones that have been classified as 1:1 orthology (14510) reported a sub-set of 2732 groups have at least one 𝕾 protein. From co-orthologous comparisons (12801) were detected and a subset of 5293 have at least one 𝕾 protein. For those sub-sets with 𝕾 proteins was possible to retrieve the `Interpro` annotation using `biomaRt`. Final tables are reported in Additional

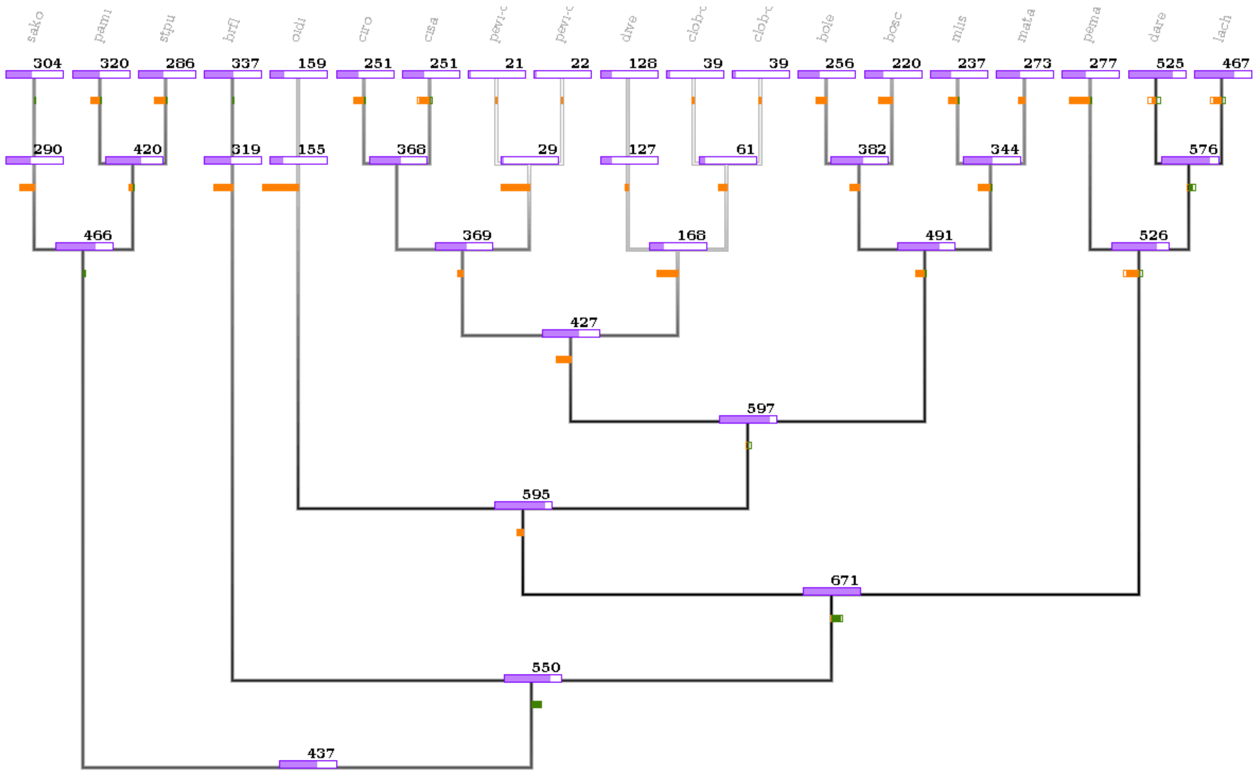

Figure 5: Evolutionary history of 1:1 orthologous **heterodomain** protein architectures in Deuterostomata.

Files 3 (oneone_heterodomains_annotation_proteins.txt, oneone_homodomains_annotation_proteins.txt). TODO:

- In this case, I have the interpro annotation for a given architecture based on the golden protein ($\mathfrak{G}$) which have a 1:1 relationship. Just reporting those data? or is more convenient to plot by some way?
- Because the paper is about innate immune system, I was looking for categories which I could classify the final architectures. I have found a nice classification specifically for tunicates on `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5465252/pdf/fimmu-08-00674.pdf` but I have to take some examples for each category and look it into the found candidates...Another option is on this systematic revision of domains: `https://www.sciencedirect.com/science/article/pii/S0378111918311119?via%3Dihub`..
- Dollo parsimony could be applied for analize gain and loss of protein architectures?.
- I could represent the protein architectures as drawings of the proteins, or a comparison between species, but for the most important ones in innate immune system, there is lot of information...
- I would like to report the ABDO method. I have the program, mainly Perl scripts and glue code in bash...or maybe in a server?

Conclusions

References

[1] B. L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, J. F. Banet, K. Billis, C. G. Girón, T. Hourlier, et al. The ensembl gene annotation system. *Database*, 2016:baw093, 2016.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1): 25–29, 2000.

[3] L. Berná and F. Alvarez-Valin. Evolutionary genomics of fast evolving tunicates. *Genome biology and evolution*, 6(7):1724–1738, 2014.

[4] R. M. Bernstein, S. F. Schluter, H. Bernstein, and J. J. Marchalonis. Primordial emergence of the recombination activating gene 1 (RAG1): sequence of the complete shark gene indicates homology to microbial integrases. *Proc. Natl. Acad. Sci. U.S.A.*, 93(18):9454–9459, Sep 1996.

[5] E. Birney, T. D. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, G. Coates, J. Cuff, V. Curwen, T. Cutts, et al. An overview of ensembl. *Genome research*, 14(5):925–928, 2004.

[6] E. Blanco, G. Parra, and R. Guigó. Using geneid to identify genes. *Current Protocols in Bioinformatics*, 18(1):4.3.1–4.3.28, 2007. doi: 10.1002/0471250953.bi0403s18. URL `https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi0403s18`.

[7] K. Breuer, A. K. Foroushani, M. R. Laird, C. Chen, A. Sribnaia, R. Lo, G. L. Winsor, R. E. W. Hancock, F. S. L. Brinkman, and D. J. Lynn. Innatedb: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Research*, 41(D1):D1228–D1233, 2013. doi: 10.1093/nar/gks1147. URL `http://nar.oxfordjournals.org/content/41/D1/D1228.abstract`.

[8] R. M. Brucker, L. J. Funkhouser, S. Setia, R. Pauly, and S. R. Bordenstein. Insect innate immunity database (iiid): An annotation tool for identifying immune genes in insect genomes. *PLOS ONE*, 7(9):1–4, 09 2012. doi: 10.1371/journal.pone.0045125. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0045125`.

[9] K. M. Buckley and J. P. Rast. Diversity of animal immune receptors and the origins of recognition complexity in the deuterostomes. *Developmental & Comparative Immunology*, 49(1):179 – 189, 2015. ISSN 0145-305X. doi: http://dx.doi.org/10.1016/j.dci.2014.10.013. URL `http://www.sciencedirect.com/science/article/pii/S0145305X14002602`.

[10] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268(1):78–94, Apr 1997.

[11] S. B. Carroll. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1):25–36, 2008.

[12] J.-M. Claverie. Computational methods for the identification of genes in vertebrate genomic sequences. *Human Molecular Genetics*, 6(10):1735, 1997. doi: 10.1093/hmg/6.10.1735. URL `+http://dx.doi.org/10.1093/hmg/6.10.1735`.

[13] M. Csuros. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15):1910–1912, Aug 2010.

[14] P. Dehal, Y. Satou, R. K. Campbell, J. Chapman, B. Degnan, A. De Tomaso, B. Davidson, A. Di Gregorio, M. Gelpke, D. M. Goodstein, et al. The draft genome of ciona intestinalis: insights into chordate and vertebrate origins. *Science*, 298(5601):2157–2167, 2002.

[15] F. Delsuc, H. Brinkmann, D. Chourrout, and H. Philippe. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, 439(7079):965–968, Feb 2006.

[16] F. Denoeud, S. Henriet, S. Mungpakdee, J.-M. Aury, C. Da Silva, H. Brinkmann, J. Mikhaleva, L. C. Olsen, C. Jubin, C. Cañestro, et al. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science*, 330(6009):1381–1385, 2010.

[17] S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nat. Protocols*, 4(8):1184–1191, 07 2009. URL `http://dx.doi.org/10.1038/nprot.2009.97`.

[18] K. Forslund and E. L. Sonnhammer. Evolution of protein domain architectures. *Methods Mol. Biol.*, 856:187–216, 2012.

[19] N. Franchi and L. Ballarin. Immunity in Protochordates: The Tunicate Perspective. *Front Immunol*, 8: 674, 2017.

[20] I. Korf, M. Yandell, and J. Bedell. *BLAST*. O'Reilly &amp; Associates, Inc., Sebastopol, CA, USA, 2003. ISBN 0596002998.

[21] M. Lechner, S. Findeiß, L. Steiner, M. Marz, P. F. Stadler, and S. J. Prohaska. Proteinortho: Detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics*, 12(1):124, Apr 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-124. URL `https://doi.org/10.1186/1471-2105-12-124`.

[22] E. M. Palsson-McDermott and L. A. O'Neill. Building an immune system from nine domains. *Biochem. Soc. Trans.*, 35(Pt 6):1437–1444, Dec 2007.

[23] N. H. Putnam, T. Butts, D. E. Ferrier, R. F. Furlong, U. Hellsten, T. Kawashima, M. Robinson-Rechavi, E. Shoguchi, A. Terry, J.-K. Yu, et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198):1064–1071, 2008.

[24] H.-C. Seo, M. Kube, R. B. Edvardsen, M. F. Jensen, A. Beck, E. Spriet, G. Gorsky, E. M. Thompson, H. Lehrach, R. Reinhardt, et al. Miniature genome in the marine chordate oikopleura dioica. *Science*, 294 (5551):2506–2506, 2001.

[25] K. S. Small, M. Brudno, M. M. Hill, and A. Sidow. A haplome alignment and reference sequence of the highly polymorphic ciona savignyi genome. *Genome biology*, 8(3):R41, 2007.

[26] M. Stanke and B. Morgenstern. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.*, 33(Web Server issue):W465–467, Jul 2005.

[27] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research*, 28(1):33–36, 2000.

[28] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, E. P. Nawrocki, L. Zaslavsky, A. Lomsadze, K. D. Pruitt, M. Borodovsky, and J. Ostell. Ncbi prokaryotic genome annotation pipeline. *Nucleic Acids Research*, page gkw569, 2016.

[29] N. Terrapon, J. Weiner, S. Grath, A. D. Moore, and E. Bornberg-Bauer. Rapid similarity search of proteins using alignments of domain arrangements. *Bioinformatics*, 30(2):274–281, 2014. doi: 10.1093/bioinformatics/btt379. URL `http://dx.doi.org/10.1093/bioinformatics/btt379`.

[30] C. A. Velandia-Huerto, A. A. Gittenberger, F. D. Brown, P. F. Stadler, and C. I. Bermudez-Santana. Automated detection of ncRNAs in the draft genome sequence of a colonial tunicate: the carpet sea squirt Didemnum vexillum. *BMC Genomics*, 17:691, Aug 2016.

[31] A. Voskoboynik, N. F. Neff, D. Sahoo, A. M. Newman, D. Pushkarev, W. Koh, B. Passarelli, H. C. Fan, G. L. Mantalas, K. J. Palmeri, et al. The genome sequence of the colonial chordate, botryllus schlosseri. *Elife*, 2:e00569, 2013.

[32] M. Yandell and D. Ence. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.*, 13(5): 329–342, Apr 2012.