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

December 8, 2018

- 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
- Discuss about the ProteinOrtho strategy to create innane immune system groups, based on protein comparison, but grouped by common domain's architectures.
- If yes, what is the best option to detect protein orthologs. Based on complete protein? or splitting by protein domains?
- Plot: Distribution of architectures? (by specie)
- Biological meaning of the detected innate immune proteins on tunicates.
- Best way to publish the paper and the new pipeline?

1 Introduction

The challenge of having a feasible and fast tool to characterize the sequences of many new genomes of non-model organisms has been increasing the need to develop strategies that extract the maximum information of the gen architecture. This trend has become the last purpose of the gene annotation where given a candidate coding sequence its biological function is assigned to further functional or evolutionary studies [1] [5] [2] [26] [27].

Nevertheless, to annotate, researchers tackle two main computational problems, one to detect first on genomic subregions the corresponding coding regions and then after or simultaneously assign functionality. Due to the broad complexity of the regulation of the eukaryotic gene expression which relies on the recognition of alternative motifs of subsequences that ends in alternatives transcripts is today more harder the definition of a gene boundary making the annotation a more complex task than CDS predictions[31].

At the begining, gene annotation was based on ab initio methods that screened for specific signals in eukaryotic genes, such as TATA box signals, exon/intronic definitions and polyadenylation signals [11]. Tools like GENSCAN introduced originally a probabilistic model of human gene structure [9]. Then, in 2005 a great advance in the annotation processes was got by the introduction of the tool AUGUSTUS [25] which allows quite accurately gene prediction and today there are versions that can be trained by integrating RNAseq data, short cDNAs etc. AUGUSTUS corresponds to the Generalized Hidden Markov Model. Other methods as GeneID uses a rules-based heuristic method to assemble the typical signals of genes in a product or model more likely. In the post-genomic era, gene annotation pipelines have been handled the heterogeneity of the available expression information of ESTs, RNA-seq or proteomic data in two main ways: by manual curation as such as the VEGA-HAVANA project of the Wellcome Trust Sanger Institute and by using automated process or hybrids methods which integrate homologies with coherent patterns of known genes of other species to characterize all the functional elements that make up the genes, ending with the assignment of a biological function to a genomic

sequence [1] [5]. For the case of vertebrates as well as some tunicate species which are the closest relatives to vertebrates [14] their genomes have undergone a rigorous annotation process widely accepted in comparative genomics of vertebrates such as the annotation pipeline of the European molecular biology laboratory (EMBL) that integrates the experimental information of 70 vertebrate and other metazoan species on the genome browser Ensembl [1].

In this work we focused on some species of the sub-phylum Tunicata since are key organisms to study the evolution of the immune system due to its phylogenetic position on the Tree of Life just before the biological phenomenon known as the immunology big-bang that gave rise to the origin of the Adaptive Immune System [4]. Additionaly, this group is characterized by the great diversity of style of life and the world widely distribution in ecological niches that might force them to design different immune response to survive in their 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, 10].

Nevertheless, the global importance of this group, the genomic studies are scarce as the comparative as well. So far, the genomes of three solitary ascidians have been annotated: the sessiles *Ciona savignyi* and *Ciona intestinalis* mapped on its 14 chromosomes [13, 24] and the pelagic *Oikopleura dioica* [15], [23]. However, for colonial ascidians, only the genome of *Botrillus schlosseri* mapped to 13 chromosomes [30] and recently the detection of ncRNAs in the draft genome of the carpet sea squirt *Didemnum vexillum* [29] are available. From preliminary comparative genomics between some tunicates and genomes of some chordates have been identified expansion of gene families by events of local duplications, but also gene loss. Some genes conserved between the amphioxus and the human have also been identified based on the tunicate genome of *C. intestinalis*. The complexity of the genomic organization of those tunicates has led to different authors to formulate the idea of the existence in their evolution of processes of genomic re-structuring in all or some tunicates genomes [22]. In addition, since the rate of evolution of other chordates has been constant, it is considered that in the tunicates evolution rates are high and therefore specific patterns of organization of all their genes are expected [3, 22]. Based on complex events that might shape the variety of tunicate genome structure we posit that they might act not only in the evolution of the genome architecture but also in the evolution of complex systems like the immune system. To study that idea we based first on Paalsson, et al. 2007 who claimed that the immune system comes from a few ancestral proteins, selected and fixed throughout the evolution and proposed that the entire immune system has been based fundamentally from nine ancestral domains which are conserved throughout the evolution of multiple organisms [21]. We propose in this survey to use protein domains as the most elementary evolutionary module of the immune system to built a picture of domain architectures turnover that might resemble the effect of evolutionary driving force that led to the complexity of the immune system in tunicates. Since domain variability implies new protein organizations (without considering that there is also re-ordering of exons), then it is extremely interesting to describe if the domain variability would have favored the creation of new protein architectures that could be positively selected in tunicates[21]. Therefore data used in this survey are built on the domains of receptors associated with the innate immune system of known canonical architectures. Our alternative homology search is proposed to complement the classical homologies searches like BLAST which are not very sensitive to traceback homologues relationships among gene with complex structures[8]. In some genes of the inmune system is not rare to find copies and reshuffling of domains. In this survey we searched for canonical and like-canonical protein domain architecture of the innate immune system genes since similar to other invertebrates, tunicates rely only on innate immunology[18]. We designed different comparison strategies for a flexible screening of architectures since tandem copies or rearrangement of domains have been reported in the evolution of domains in proteins [17].

Methods and Materials

Protein domain architectures of reference

We started with annotated and curated genes from `InnateDB` [6] and `Insect Innate Immunity Database` (IIID) [7] 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 start-

ing point interfaces records from `InnateDB` to `Ensembl` (v.86) by using `Perl` scripts and `biomaRt` R library [16]. 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 five species of chordates *C. robusta*, *C. savignyi*, *P. marinus*, *D. rerio* and *L. chalumnae*, annotated in `Ensembl` (v.81). The method to compare architectures is explained later in the section *Architecture comparison strategies* 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: *P. marinus*, *D. rerio* and *L. chalumnae*, 10 species of Tunicata: *Oikopleura dioica*, *Botryllus schlosseri*, *Botrylloides leachii*, *Ciona robusta*, *Ciona savignyi*, *Didemnum vexillum*, *Molgula occidentalis*, *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 genomic annotation were retrieved are described in Table 1.

Reduction system

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 $G_m^a = (P_i, P_{i+1}, \ldots, P_{i+k})$ 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).

Comparison Strategies

In order to detect in tunicate species architectures from the *gold standard set* $\mathfrak{G}$, we designed four scanning strategies. For practical reasons pair-wise comparisons were performed between the same annotation system m, i.e., architectures from *Pfam* in one tunicate specie were compared to the subset of architecture m of type *Pfam* on the protein a in $\mathfrak{G}$. List of gene architecture from tunicates and the other chordates were greedily 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.

[1] `https://www.ncbi.nlm.nih.gov/sites/batchentrez`
[2] http://tunicatapviridis.bioinf.uni-leipzig.de/Download.html
[3] http://tunicatacoblonga.bioinf.uni-leipzig.de/Download_clob.html
[4] http://tunicata.bioinf.uni-leipzig.de

Sub-phylum	Specie	Source	Version
V	*L. chalumnae*	`Ensembl FTP`	Release 81
V	*D. rerio*	`Ensembl FTP`	Release 81
V	*P. marinus*	`Ensembl FTP`	Release 81
T	*M. occidentalis*	`ANISEED`	v.1
T	*M. oculata*	`ANISEED`	v.1
T	*C. robusta*	`Ensembl FTP`	Release 81
T	*P. viridis*	BogotaUNAL[2]	Draft version
T	*C. savignyi*	`Ensembl FTP`	Release 81
T	*C. oblonga*	BogotaUNAL[3]	Draft version
T	*D. vexillum*	BogotaUNAL[4]	This version
T	*B. schlosseri*	`ANISEED`	`botznik-chr.fa`
T	*B. leachii*	`ANISEED`	v.1
T	*O. dioica*	`Genoscope FTP`	Version 3
C	*B. floridae*	`JGI genome portal`	v.1 and v.2
H	*S. kowalevshii*	`NCBI FTP`	Skow_1.1
E	*P. miniata*	`Echinobase`	v2.0
E	*S. purpuratus*	`Echinobase`	v4.2

Table 1: Genomic data source. Labels **V, T**, **C**, **C** and **E** represent: vertebrates, tunicates, cephalochordates, hemichordates and echinoderms respectively. **NA=** not available. Please add links of the databases as a comment.

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 $Q_m^a success(o)$

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

If $Q_m^a success(o) = 1$ we say that exist in $\mathfrak{G}$ 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 $\mathfrak{G}$ without considering order in $\mathcal{P}$ domains. Now the rules are defined as follows:

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

If $Q_m^a success(d) = 1$ we say that exist in $\mathfrak{G}$ 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 [19]. 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`[5] [28]. 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>
```

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.`

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

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 [] 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*[6] and *O.dioica*[7]. 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* [29] a *de novo* gene prediction was performed directly on the assembled contigs using AUGUSTUS [25] 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) [20], as follows:

Change protein ortho parameters according to the discussion about groups and orthology relationships.

```
#Step 1:
proteinortho5.pl -synteny -dups=3 -verbose -cpus=8 -clean
-keep -project=<name-project> -step=1 -temp=${temp-directory} -p=blastn
<fasta file>
```

[6] ftp://genome.crg.es/pub/software/geneid/cintestinalis.param_Apr_26_2006
[7] ftp://genome.crg.es/pub/software/geneid/odioica.param_Nov_10_2006

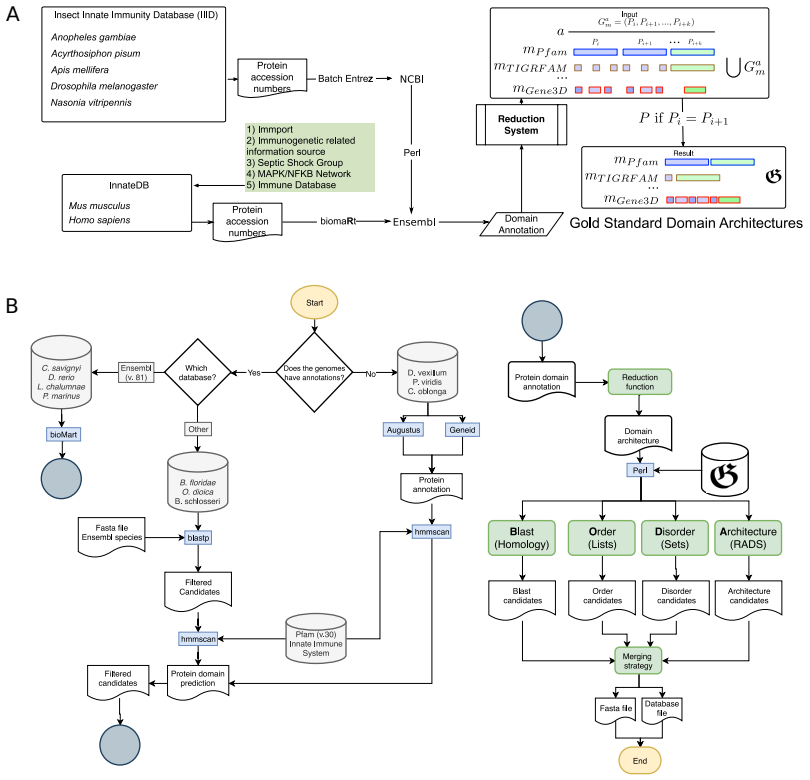

Figure 1: **A.** Workflow to generate $\mathfrak{G}$. Innate immune system databases were used to obtain the accession numbers of the related proteins. Next, domain annotation was accesed throught `biomaRt` from `Ensembl` (v.81). Post processing include reduction of the consecutive repetitions of protein 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).

```
#Step 2:
proteinortho5.pl -synteny -dups=3 -graph -cpus=8 -verbose -clean
-step=2 -project={name-project} -keep -startat=${step} -stopat=4 -p=blastn
<fasta file>
#Step 3:
proteinortho5.pl -synteny -dups=3 -graph -cpus=16 -verbose -clean -step=3
-keep -project=<name-project> -p=blastn <fasta file>
```

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 [12], 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 1.

ABDO strategy comparisons of domains

The distribution of genes and proteins that belong from *gold standard set* $\mathfrak{G}$ is shown on Additional File 1: 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 2, exists 4 different groups of species: those ones that have been annotated in `Ensembl` as: *Cŕobusta*, *C. savignyi*, *P. marinus*, *D. rerio* and *L. chalumnae*. Next, 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, like: *D. vexillum*, *C. oblonga* and *P. viridis*, where does not exists 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*[8] and with `AUGUSTUS` using transcriptome data. 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 with the automated pipeline NAME_PIPELINE generated in this study.

Table 2 summarizes the final innate immune system candidates, based on homology architecture strategies (**ABDO**), which have been applied independently. The final number of immune system set results of the merging of all the candidates obtained from the applied strategies, except for the candidates that have been detected only by the **B** strategy. The main reason is that comparison does not represent an pure architecture relation, because the complete sequence is used in the pairwise alignment, this method has been frequently used to detect homologs in distant or related species, according to the design of the homology strategy based on the `blastp` parameters CITE. When it was applied, always the B strategy reported higher frequencies of protein candidates in comparison to the O or D strategies; and those results are more similar to the A strategy, which does not have a previous reduction step.

Then taking all into account, Table 2 shows the 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. This distribution of relationships against $\mathfrak{G}$ set

[8]For those genomes, in Table 2 are referenced as *Ciro* and *Oidi* in parenthesis, respectively

Specie	**Annotated Genes**	**Annotated Proteins**	**Annotated Prot. with Pfam domains**	**Orderded Prot**	**Disorder Prot**	**Blast Prot**	**Architecture**	**Total Prot IS**
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 2: Final distribution of annotated genes and found candidate Immune system proteins. The percentage in relation of the total of reported proteins is in parenthesis.

shows a highest number of proteins that have been detected as immune system candidates by only 1 architecture strategy. Additionaly, 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 $\mathfrak{G}$. 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 respecto to $\mathfrak{G}$ (Figure 2B). Mainly those relationships have been detected with human and mouse proteins, as described earlier, those protein candidates are the most frequent proteins in $\mathfrak{G}$.

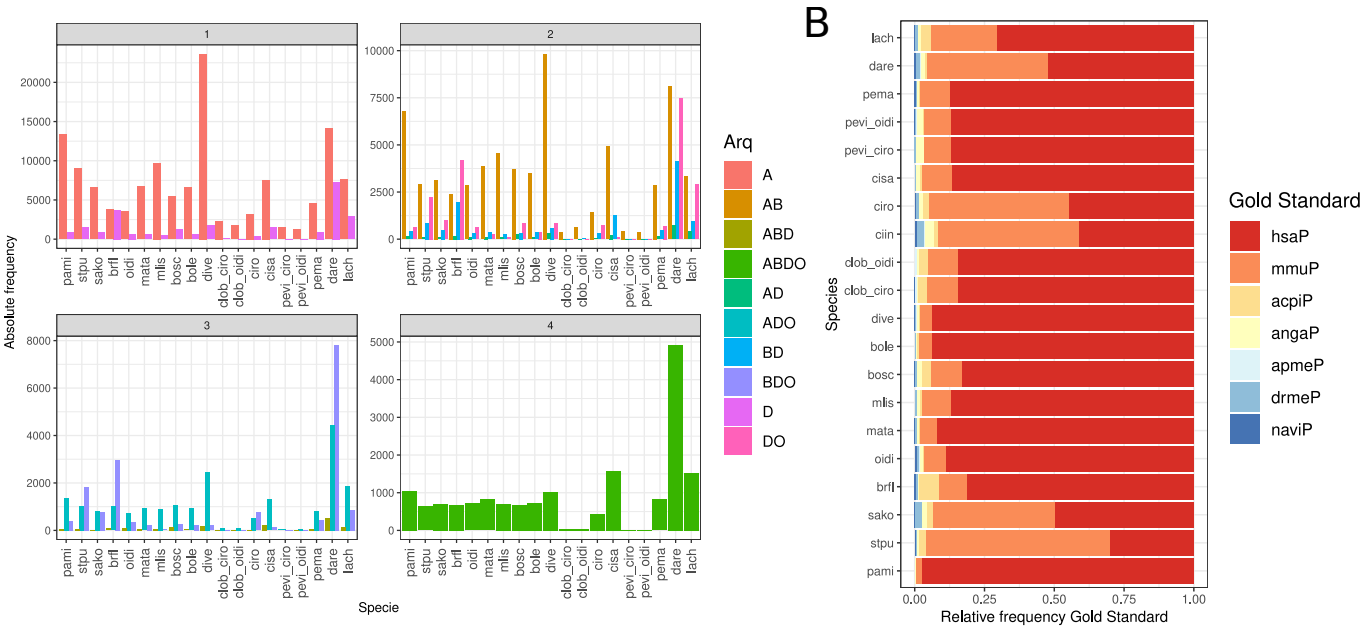

Figure 2: **A.**Frequency of detected proteins with defined architecture comparison strategies classified according to the number of possible combinations of architecture strategies (described in more detail in main text), against $\mathfrak{G}$. **B.** Mean shared proportion homology architecture against 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*.

Orthology relationships between Immune system candidates

The merging step on the resulting immune system candidates generated 11013 groups that shared the same $\mathfrak{G}$ proteins. Applying `ProteinOrtho` was possible to identify.

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.

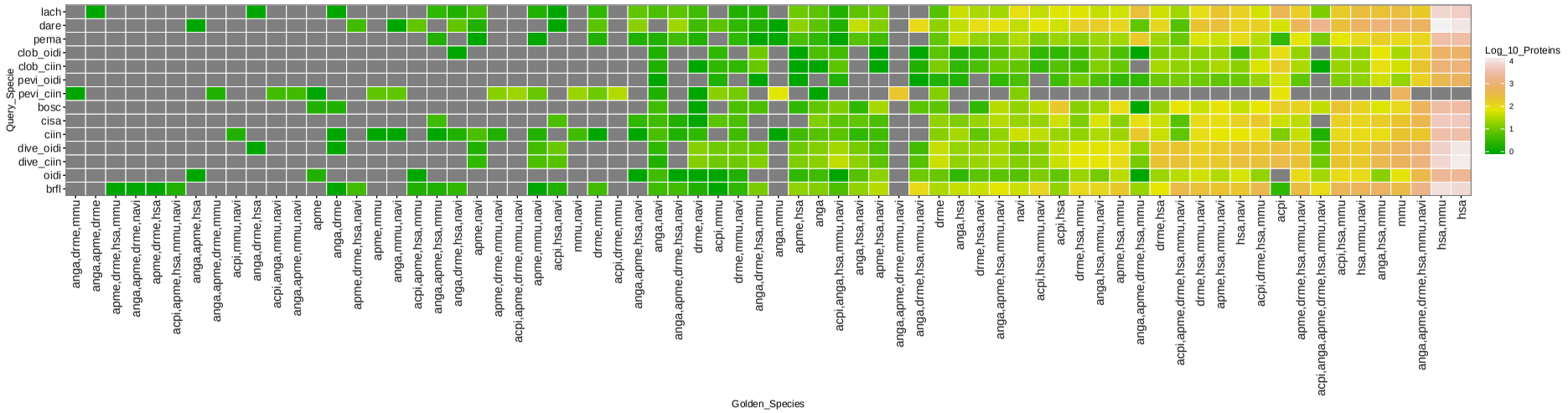

Figure 3: Heatmap of the number of proteins identified in the $\mathfrak{G}$ species by at least one of the architecture strategies. The number of proteins have been transformed by $\log_{10} x$. Combinations of species have been obtained according to the final agrupation given by the relationships between Query species and $\mathfrak{G}$ species. This plot have been organized according to the number of final candidates by set of $\mathfrak{G}$ species, left with lower values to the higher ones on the rigth. navi=*N. vitripennis*, apme=*A. mellifera*, drme=*D. melanogaster*, anga=*A. gambiae* and acpi=*A. pisum*; and Mammals: mmu=*M. musculus* and hsa=*H. sapiens*.

Table 3: Top ten of most frequent domains found on chordate's immune system proteins.

Number Architectures	**Frecuency**	**ACC**	**PFAM Annotation**	**GO term**
96	1143	PF00028	Cadherin domain	GO:calcium ion binding; GO:0005509. GO:homophilic cell adhesion via plasma membrane adhesion molecules; GO:0007156. GO:membrane; GO:0016020
422	1183	PF00096	Zinc finger, C2H2 type	NA
185	1385	PF01391	Collagen triple helix repeat (20 copies)	NA
279	1655	PF00057	Low-density lipoprotein receptor domain class A	Ldl_recept_a; GO:protein binding; GO:0005515
302	1780	PF00084	Sushi repeat (SCR repeat)	NA
430	2293	PF13465	Zinc-finger double domain	NA
680	3566	PF07679	Immunoglobulin I-set domain	NA
683	3966	PF07645	Calcium-binding EGF domain	EGF_CA; GO:calcium ion binding; GO:0005509
543	4150	PF00041	Fibronectin type III domain	fn3; GO:protein binding; GO:0005515
731	4703	PF00008	EGF-like domain	NA

[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] 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`.

[7] 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`.

[8] 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.

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

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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`.

[29] 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.

[30] 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.

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