

Supplemental information #4: A new strategy to characterize the domain architecture structure of proteins of the innate immune system in tunicate species

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

March 27, 2019

Didemnum vexillum re-annotation

Annotation of coding regions with Augustus

Mapping previous ncRNA annotation on new assembly

Previous ncRNA annotation was retrieved from Velandia-Huerto, *et al* [] in fasta format. All the contigs which have been reported an ncRNA have been obtained from the old *D. vexillum* genome version¹. This multifasta file was mapped onto the new genome with **blast**: `¡PARAMETERSi`. According to the set of blast parameters the number of contigs were increasing into the new genome. At the same time, if one contig reported more than one candidates into the new genome, was chosen this/those one (s) that reported the highest bitscore. Having this previous information as an additional source of information in order to clean the true position of the annotated ncRNAs in the new genomes. After mapping all the candidates with **blast**, the true locations were obtained after applying those filters:

- Identity have to be $\geq 85\%$.
- E-value $\leq 10^{-10}$.
- Relation of sizes between the homology region of the query (r_h) and their calculated size (r_s) have to be $\frac{r_h}{r_s} \geq 0.9$

From the 264 reported loci of ncRNAs, XY were retrieved in the new genome, GFF file is reported along with all the coding and non-coding regions on Supplemental File X.

¹<http://tunicata.bioinf.uni-leipzig.de/Download.html>