Supplemental information #4: 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

April 4, 2019

Didemnum vexillum re-annotation

Annotation of coding regions with `Augustus`

Ernesto and Clara are working on that.

Mapping previous ncRNA annnotation 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 first reported assembly of the *D. vexillum* genome[1]. This multifasta file was mapped onto the new genome with `lastz`:

```
lastz_32 <NEW_GENOME>[multiple] <OLD_GENOME> --chain C=0 E=150 H=0 K=4500 L
    =3000 M=254 O=600 Q=human_chimp.v2.q T=2 Y=15000 --format=maf+
```

Aligment files were retrieved in `maf` format and were parsed with `Bio::AlignIO` `Bioperl` library. The criteria to obtain the best genome coordinates was choosen based on the relation between the length of the mapped region into the new genome (m) and the original size of the query contig in the old genome (s). The relation was defined as $R = \frac{m}{s}$, and were defined as the best mapping candidates those ones reported $R = 1$, but in order to retrieve the maximum number of mapping between the two genome versions, $R \geq 0.90$ was also considered.

From 247 contigs, was possible to map 212 in the raw results after the mapping stage with *lastz*. After considering the R relation, those results were parsed, resulting in: 64 ($R = 1$), 35 ($0.95 \leq R < 1$), 39 ($0.90 \leq R < 0.95$) and 32 ($0.85 \leq R < 0.90$), in total 170 contigs that reported high score mapping into the new genome.

Best candidates was choosen based on the final aligment score. For those contigs that reported 1:many relations, those set of positions in the new assembly was also considered for the following analysis.

Sequences from ncRNAs was obtained and mapped against the new *D. vexillum* assembly with `blast`, as follows:

```
blastall -p blastb -d <DB> -i <QUERY> -F F -e 10e-5 -m 8 -o <OUT>
```

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 choosen 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:

[1] `http://tunicata.bioinf.uni-leipzig.de/Download.html`

- 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 $\frac{r_h}{r_s} \geq 0.9$

An additional confirmation step was performed using the Covariance Models from `RFAM`v.11 onto the retrieved fasta sequences, using `infernal` package:

cmsearch −g −Z <NT number (Mb)> −−toponly <FASTA> <CM>

Previously reported ncRNAs from first assembly draft on *D. vexillum* reported 264 loci. Following the explained strategy allowed to identify the new coordinates of ncRNAs on the new genome assembly. A 73 loci was retrieved and had reported additional support from genome alignments, 63 have been identified on the new genome in another location that is different to the correspondent new region of the old contig. Finally, 1441 candidates that belong from 12 ncRNA families have been retrieved applying HMMs directly on the genome sequence. A high number of loci from 5S rRNA (389), U1 (326), U2 (316) and U6 (381) spliceosomal RNAs have been detected, for that reason the complete number of candidates is high as previosuly reported. All of the final candidates were subject of structural evaluation with the correspondent metazoan-specific covariance model. The final list of candidates have been reported in a single GFF3 file (`didemnum_vexillum_ncRNAs_mapped.gff3`). Redundant coordinates are loci from the first assembly that mapped into the new genome in the same position.