ANISEED 2017: Extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets.

Matija Brozovic[1], Christelle Dantec[1], Justine Dardaillon[1], Delphine Dauga[2], Emmanuel Faure[3,4], Mathieu Gineste[1], Alexandra Louis[5], Magali Naville[6], Kazuhiro R. Nitta[7], Jacques Piette[1], Wendy Reeves[8], Céline Scornavacca[9], Paul Simion[9], Renaud Vincentelli[10], Maelle Bellec[11], Sameh Ben Aicha[12], Marie Fagotto[11], Marion Guéroult-Bellone[2], Maximilian Haeussler[13], Edwin Jacox[1], Elijah K Lowe[14, 15], Mickael Mendez[7], Alexis Roberge[11], Alberto Stolfi[16], Rui Yokomori[17], C. Titus Brown[18], Christian Cambillau[10], Lionel Christiaen[19], Frédéric Delsuc[9], Emmanuel Douzery[9], Rémi Dumollard[12], Takehiro Kusakabe[20], Kenta Nakai[17], Hiroki Nishida[21], Yutaka Satou[22], Billie Swalla[23], Michael Veeman[8], Jean-Nicolas Volff[6], and Patrick Lemaire[1,3]*

Supplementary methods: Inference of gene homology relationships between ascidians, vertebrates and three deuterostome outgroups.

To infer homology and orthology relationships between ascidians, vertebrates and other invertebrate deuterostome, we first used SiLiX (1) to build clusters of homologous genes from 18 species:

6 vertebrates:

- Rodent: *Mus musculus*;
- Primate: *Homo sapiens*;
- Bird: *Gallus gallus*,
- Turtle: *Pelodiscus sinensis*;
- Coelacanth: *Latimeria chalumnae*
- Shark: *Callorhinchus milii.*
- Source of data: ENSEMBL 84 (http://mar2016.archive.ensembl.org/index.html) except *C. milii* (http://esharkgenome.imcb.a-star.edu.sg/download/).

9 ascidians:

- Phlebobranchs: *Ciona robusta, Ciona savignyi, Phallusia mammillata, Phallusia fumigata,*
- Stolidobranchs: *Halocynthia roretzi, Halocynthia aurantium, Botryllus schlosseri, Molgula oculata, Molgula occidentalis*
- Source of data: ANISEED (https://www.aniseed.cnrs.fr/aniseed/download/download_data)

1 lancelet:

- *Branchiostoma belcheri* (genome.bucm.edu.cn/lancelet/download_data.php)

2 non-chordate deuterostome outgroups:

- *Strongylocentrotus purpuratus*: (http://www.echinobase.org/Echinobase/SpDownloads)
- *Saccoglossus Kowalevskii* (https://www.hgsc.bcm.edu/other-invertebrates/acorn-worm-genome-project).

Starting from the longest annotated protein for each gene model, we ran Silix on the whole dataset (first relaxed iteration on the whole set: silix -n -i 0.20 -r 0.40 -l 100 -m 0.50. Second more stringent iteration to break a large cluster containing ~30% of sequences after round 1: silix -n -i 0.35 -r 0.80 -l 100 -m 0.50). 12885 clusters were selected as they contained: i) one tunicate and at least one vertebrate or one outgroup species, or ii) two tunicate species which did not both belong to the *Halocynthia* genus. The sequences within each of these clusters were aligned using the MAFFT Multiple alignment software (2), and the alignments trimmed using trimAl (3) (-gt 0.3 -resoverlap 0.75 -seqoverlap 80). Low quality sequences in the resulting alignments were removed if they were both smaller than 100 amino-acid long and presented more than 70 % of missing data. Phylogenetic trees were then infered in RaxML (4) under the LG+Γ4+F evolution model with 100 bootstrap replicates to estimate node support. All branches with bootstrap support inferior to 35% were collapsed. These collapsed trees were then analyzed using a custom C++ program in order to map onto each trees the apparent duplication events and thus to reconstruct all pairwise orthology relationships between all genes. Table S1 presents some statistics on orthology relationships.

1. Miele,V., Penel,S. and Duret,L. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, **12**, 116.
2. Katoh,K. and Standley,D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.*, **30**, 772–780.
3. Capella-Gutiérrez,S., Silla-Martínez,J.M. and Gabaldón,T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
4. Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma. Oxf. Engl.*, **30**, 1312–1313.

Table S1: Orthology relationships between tunicate and human genes and between tunicate orders (stolidobranch vs Phlebobranch).

	Total gene number	Genes with Hsap ortholog		Mean number of Hsap orthologs	Genes with orthologs in a different tunicate order		Genes with ortholog in a different tunicate order w/o Hsap ortholog	
		number	% of total		number	% of total	Number	% of total
Cirobu	15284	4405	29%	2,81	8538	56%	4270	28%
Cisavi	12172	3663	30%	2,84	6838	56%	3311	27%
Phmamm	19508	4257	22%	2,89	9275	48%	5057	26%
Phfumi	10090	2987	30%	2,80	6119	61%	3166	31%
Boschl	46519	4387	9%	2,92	9006	19%	4905	11%
Moocul	15313	3976	26%	2,88	7559	49%	3812	25%
Moocci	30639	4312	14%	2,91	8260	27%	4249	14%
Harore	16083	4244	26%	2,82	8027	50%	3973	25%
Haaura	11436	3958	35%	2,84	7548	66%	3767	33%