Okinawa Institute of Science and Technology

Graduate University

Thesis submitted for the degree

Doctor of Philosophy

A Study of Horizontally Transferred Glycosyl Hydrolase Family 6 Genes in Tunicate Genomes

by

Kun-Lung Li

Supervisor:
Hiroshi Watanabe

Co-Supervisor:
Noriyuki Satoh

September 2021

Declaration of Original Authorship

I, Kun-Lung Li, declare that this thesis entitled “A Study of Horizontally Transferred Glycosyl Hydrolase Family 6 Genes in Tunicate Genomes” and the data presented in it are original and my own work.

I confirm that:

- No part of this work has previously been submitted for a degree at this or any other university.
- References to the work of others have been clearly acknowledged. Quotations from the work of others have been clearly indicated and attributed to them.
- In cases where others have contributed to part of this work, such contribution has been clearly acknowledged and distinguished from my own work.
- Chapter 1 (page 5-30) of this work is based on a previously published paper stated below. The co-authors had checked the originality of this thesis. No other part of contents has been previously published elsewhere.
  - Li, K.-L.; Nakashima, K.; Inoue, J.; Satoh, N. Phylogenetic analyses of Glycosyl Hydrolase Family 6 genes in tunicates: possible horizontal transfer. *Genes* **2020**, *11*, 937, doi:10.3390/genes11080937.

Date: 2021-09-22

Signature: Kun-Lung LI

Abstract

Tunicates are the closest extant relatives of vertebrates. Tunicates produce cellulose-containing tunic and exhibit very characteristic lifestyles among animals. Their unique ability to synthesize cellulose results from a horizontally transferred cellulose synthase gene (*CesA*). Interestingly, a Glycosyl Hydrolase Family 6 (GH6) hydrolase-like domain exists at the C-terminus of tunicate CesA but not in cellulose synthases of other organisms. This led to the identification of another independent GH6-encoding gene, *GH6-1*, in tunicate genomes. These GH6-encoding genes exist exclusively in tunicates within the animal kingdom. The existence of GH6-encoding genes and the combination of GH6 and cellulose synthase domains raised the question of the evolutionary origin and function of GH6s in tunicates. To answer these questions, I first examined the phylogenetic relationship of GH6-encoding genes by comparing their sequence signatures. Tunicate *CesA* and *GH6-1* genes represent two independent orthologous groups, but the origin of these genes before a horizontal transfer event could not be ascertained. Secondly, I examined the expression of tunicate *CesA* and *GH6-1* genes in *Ciona intestinalis* type A, a model ascidian tunicate. The gene expression in embryos at early developmental stages was examined by quantitative reverse transcription PCR and *in situ* hybridization. Obvious expression of both *CesA* and *GH6-1* were found at embryonic stages of *Ciona* embryo at epidermis. The observed expression profiles were also compared with a set of single-cell transcriptome data provided by our collaborators. Embryonic cells of late tailbud stage I showed that both *GH6-1*- and *CesA*-expressing cells are mostly in cell clusters of epidermal identity. Localized signal in the reporter assay also suggest the existence of specific enhancers upstream to *Ciona GH6-1* gene. Finally, I used genome-editing technique to generate *GH6-1* knockout larvae of *C. intestinalis* type A and observed that affected embryos show perturbed papillae formation and metamorphosis. My study showed that *GH6-1*, a gene very likely originated from horizontal gene transfer, is recruited to function in ascidian early development. This observation would help to address how tunicates evolved by obtaining their unique anatomy and ecology.

Acknowledgements

I greatly thank my supervisors, Hiroshi Watanabe and Noriyuki Satoh, for their valuable guidance and suggestion in every step of my research.

I am grateful to trainings provided by and discussion with the members in Marine Genomics Unit, especially Keisuke Nakashima, Koki Nishitsuji, Sakura Kikuchi, Kanako Hisata, Konstantin Khalturin, Hitoshi Tominaga, Yuuri Yasuoka, Jun Inoue, Tsai-Ming Lu, and Seiya Kitanobo on molecular cloning, animal aquaculture, animal experiments, and bioinformatics analysis. I also got practical suggestion on aquaculture from Jeffrey Jolly, Koichi Toda, and OIST Marine Science Station staff. I would also acknowledge the training on *Ciona* transgenic experiments from prof. Hiroki Takahashi, National Institute for Basic Biology.

I thank prof. Yasunori Sasakura (National BioResource Project, NBRP, Japan Agency for Medical Research and Development / University of Tsukuba) for providing TALEN protocol and vectors, and Reiko Yoshida, Satoe Aratake, Akihiro Yoshikawa, Chikako Imaizumi, Megumi Koutsuka, and other members in the NBRP for providing *Ciona intestinalis* type A adults used in this research.

I am indebted to Chihiro Arasaki, Shoko Yamakawa, and Tomomi Teruya, who took care of all aspects of research administrations.

I appreciate the assistance on designing the research, trainings on methodology, and reviewing manuscript by Keisuke Nakashima, Jun Inoue, and prof. Noriyuki Satoh and the English editing by Steven Aird when I was preparing the paper manuscript, “Phylogenetic Analyses of Glycosyl Hydrolase Family 6 Genes in Tunicates: Possible Horizontal Transfer”, that had been published in *Genes*, 2020.

List of Abbreviations

BLAST:	Basic Local Alignment Search Tool
cDNA:	complementary DNA
DIC:	differential interference contrast (microscopy)
EDTA:	ethylenediaminetetraacetic acid
EST:	expressed sequence tag
FAM–TAMRA:	fluorescein amidite–tetramethylrhodamine
GFF:	General Feature Format (for nucleotide sequence)
GFP:	green fluorescent protein
GH6:	glycosyl hydrolase family 6, also known as: glycoside hydrolase family 6
GT2:	glycosyltransferase like family 2
HGT:	horizontal gene transfer
hpf:	hours post-fertilization
kbp:	thousand base pairs
MOPS:	3-morpholinopropane-1-sulfonic acid
PBS:	phosphate buffered saline
PCR:	polymerase chain reaction
RT-qPCR:	reverse transcription quantitative polymerase chain reaction
sj:	the *swimming juvenile* mutant of *Ciona intestinalis*
TALEN:	transcription activator-like effector nuclease
TE buffer:	Tris-EDTA buffer
UMAP:	Uniform Manifold Approximation and Projection

Table of Contents

List of Figures and List of Tables

List of Figures

List of Tables

Introduction

(1) Tunicates, the sister group to vertebrates

Tunicates are a group of marine animals, consisting of approximate 3000 extant species (Appeltans et al., 2012). The name of tunicates stemmed from the tunic, also called the test, a cellulose-containing fibrous layer that covers their adult body (Satoh, 2014). Tunicates are classified into three classes: Ascidiacea, Thaliacea, and Appendicularia. Among the classes, ascidians are the most diverse and comprise of about 2,900 described species (Appeltans et al., 2012). Generally, ascidians have a benthic, sessile, filter-feeding adult stage and a larval stage swimming freely with a muscular tail (Lemaire, 2011). Anural larvae of a few ascidian species were described to be a derived character (Jeffery et al., 1999; Huber et al., 2000). Thaliaceans (the salps, doliolids, and pyrosomes) are phylogenetically nested within ascidians (Tsagkogeorga et al., 2009; Govindarajan et al., 2010; Delsuc et al., 2018); they swim by jet propulsion in open ocean. Different from ascidians and thaliaceans, appendicularians (larvaceans) do not maintain a rigid tunic but repeatedly secrete cellulose-containing ‘house’, which directs water flow and facilitates filter-feeding.

Tunicates, cephalochordates (lancelets), and vertebrates constitute the chordate phylum. However, historically tunicates had been recognized as mollusks by Aristotle or as echinoderms by Lamarck based on their adult appearance (Satoh, 2003; Satoh, 2014; Holland, 2016). In the 19th century, embryologist Alexander Kowalevsky carefully observed the anatomy of tunicate larvae and started a grouping of tunicates with other chordate animals (Kowalewski, 1866).

Albeit the adult tunicates show somewhat peculiar morphology and structures, developing tunicate embryos share many important characters with vertebrates. While the notochord and dorsal tubular central nervous system are shared by all three major clades of chordates, the existence of ectodermal placodes and neural-crest like cells in tunicates support the close relation of tunicates and vertebrates (Manni et al., 2004; Mazet et al., 2005; Jeffery, 2007; Stolfi et al., 2015). Based on the embryonic characters and recent phylogenomics studies, tunicates are now recognized as the closest invertebrate relatives to vertebrates (Delsuc et al., 2006; Delsuc et al., 2008; Delsuc et al., 2018).

Due to the important evolutionary position of tunicates, genomes of several tunicate species have been sequenced. Among those, the draft genome of *Ciona intestinalis* (Dehal et al., 2002) was released before the completion of human genome project in the year 2003. Available genomic resources and the ease of embryological manipulations have made tunicates good model animals for evolutionary and developmental biological studies in recent decades (Lemaire, 2011; Satoh, 2014). Comparative studies on genomes and embryonic development have shed light on the evolutionary history of chordates and vertebrates: genome evolution (Dehal et al., 2002), the origin of neural crest cells (Jeffery, 2007), central nervous system development (Imai et al., 2009) are just a few of the important fields of study.

Please note that the name of the species ‘*Ciona intestinalis*’ or ‘*C. intestinalis* type A’ used in this manuscript would match the species nomenclature of the first released “*Ciona*” genome and archived sequence data in many public databases, including the National Center for Biotechnology Information (NCBI, U.S. National Library of Medicine). On the other hand, a few recent studies examined the divergence of *Ciona intestinalis* type A and type B lineages (Ohta et al., 2020; Satou et al., 2021), and *C. intestinalis* type A is sometimes called *C. robusta* (Brunetti et al., 2015).

Aside from chordate common characters, tunicates have various lineage-specific diversification that made them very different from other chordate relatives. Among those diversification, the ability to utilize cellulose is unique in the animal kingdom (Satoh, 2016).

(2) Tunicate cellulose synthase

Cellulose is the largest biomass on Earth (Coughlan, 1985). Cellulose is produced by a taxonomically diverse group of organisms, including plants, algae, bacteria, protists, and fungi (Lin and Aronson, 1970; Matthysse, 1983; Brown and Saxena, 2007). Cellulose is a polymer composed of repeating β 1-4-linked D-glucose. Cellulose molecules form hydrogen-bonded microfibril that has a high tensile strength, and therefore cellulose is used as the physical barrier by various organisms. In plants, cellulose is synthesized by rosette-like protein complexes [(Brown, 2006), and reviewed in (Endler and Persson, 2011)]. Within the complex, the most-studied component is cellulose synthase, encoded by *CesA* genes. The *Arabidopsis thaliana* genome contains 10 *CesA* genes (*CesA1-10*), which encode proteins with homology to bacterial cellulose synthases. It had been shown that CesA1-CesA9 proteins had associations with CesA complexes during either primary or secondary cell wall formation, while the role of CesA10 is not yet clear (Endler and Persson, 2011; Griffiths et al., 2015).

Animals do not synthesize cellulose, but the only exception is tunicates (Satoh, 2016). Scientist had noticed a fibrous component showing polysaccharide characters in the tunic and called it as tunicin, which was later confirmed to be a form of cellulose (Endean, 1961). In another definition, the tunicin was defined as "the alkali-insoluble fibrous fraction of the tunic" and contains cellulose and other associated amino acids and proteoglycans (Van Daele et al., 1992). Before the identification of a tunicate cellulose synthase, cellulose synthesis complexes had also been observed with the freeze fracture technique in a colonial ascidian, *Metandrocarpa uedai* (Kimura and Itoh, 1996).

With the aid of a draft genome, transcript of the first tunicate cellulose synthase gene, *Ci-CesA*, was identified in *Ciona intestinalis* (Dehal et al., 2002; Nakashima et al., 2004). Later, more tunicate cellulose synthase genes were also identified (Matthysse et al., 2004; Sagane et al., 2010; Nakashima et al., 2011).

The predicted tunicate cellulose synthase (CesA) proteins are large. For example, *Ciona intestinalis* CesA protein (Ci-CesA) has the size of 1277 amino acids. The N-terminal part of Ci-CesA contains transmembrane helices and conserved motifs that are seen in members of the Glycosyltransferase-like Family 2 [GT2, PF13641 of the Pfam database (El-Gebali et al., 2019; Mistry et al., 2021)]. The molecular phylogenetics analysis also grouped the aforementioned 'cellulose synthase domain' of tunicate CesA with cellulose synthases of other organisms among GT2 members (Nakashima et al., 2004).

The cellulose synthesis ability of tunicate cellulose synthases were also examined (Matthysse et al., 2004; Sasakura et al., 2005): expression constructs of the *C. savignyi* cellulose synthase gene could restore the cellulose biosynthesis in a cellulose synthase mutant of *Agrobacterium tumefaciens* (Matthysse et al., 2004), and a transposon-mediated mutation that disrupted *CesA* expression made a mutant called "*swimming juvenile* (*sj*)", in which *C. intestinalis* larvae lose cellulose microfibrils (Sasakura et al., 2005).

Intriguingly, the C-terminal part of the tunicate CesA protein contains a hydrolase-like domain, which shows sequence similarity to cellulases of the Glycosyl Hydrolase Family 6 (GH6, Pfam: PF01341). However, the GH6 domain of tunicate CesA proteins (tunicate CesA-GH6) have amino acid substitutions at the probable active site (Koivula et al., 2002; Matthysse et al., 2004; Nakashima et al., 2004; Sagane et al., 2010; Nakashima et al., 2011) and therefore it may lack hydrolytic activity. The physiological role of tunicate CesA-GH6 has remained unclear.

The origin of the tunicate *CesA* gene had also been questioned. Phylogenetic analyses clustered the tunicate cellulose synthase domain (GT2) closer to the corresponding bacterial synthases than plant cellulose synthases, while the analyses of the GH6 domain was not conclusive (Sagane et al., 2010; Nakashima et al., 2011). Based on molecular phylogeny and the unique structure of tunicate *CesA*, Nakashima et al. (2004) hypothesized that a bacterial genomic region that contained both a *GT2/CesA* gene and a *GH6* gene, was transferred horizontally to ancestral tunicates, and that the two genes/domains later merged to form the tunicate *CesA* gene.

(3) Horizontal gene transfer may have greatly affected tunicate biology

Horizontal gene transfer (HGT, or lateral gene transfer) is the movement of genetic material between genomes that have no parent-offspring relationship. It could happen between different species or among organelle and nuclear genomes. Bacterial genomes are greatly shaped by HGT and some of them may contain more than 10% transferred genes (Garcia-Vallve et al., 2000; Koonin et al., 2001; Soucy et al., 2015). Although animals usually inherit genetic information from parents (Martin, 2017), many horizontally transferred genes are maintained in animal genomes and expressed (Dunning Hotopp, 2011; Boto, 2014; Husnik and Mccutcheon, 2018). HGT may well be one of the most important forces shaping animal evolution (Boto, 2014).

Tunicate cellulose synthase, which is generally assumed to be originated by horizontal transfer (Sagane et al., 2010; Sasakura et al., 2016; Bhattachan and Dong, 2017), also shaped tunicate evolution. In addition to providing an important material for structural support of the tunic/house, the tunicate *CesA* gene also affects larva-to-juvenile metamorphosis in *Ciona* (Sasakura et al., 2005). In a mutant line of *C. intestinalis*, *swimming juveniles* (*sj*), a tandem array of *Minos* transposon was inserted 327–328 base pairs upstream of the *Ci-CesA* transcriptional start site. The *sj* mutant larvae showed an abnormal appearance of the tunic and a disrupted metamorphosis. Normally, swimming *Ciona* larvae first adhere to a substrate and retract adhesive papillae before the subsequent metamorphic events including tail resorption and body axis rotation (Cloney, 1982). However, the *sj* mutants started metamorphic events of adhesive papillae retraction and body axis rotation, but they retained the larval tail and kept swimming (Sasakura et al., 2005). The trunk of *sj* metamorphosed to juvenile-like structures but the larval tail remained, therefore they were named as *swimming juveniles*. Also, the *sj* larvae had a lower efficiency of adhering to the substrate. Functional suppression of *Ci-CesA* with morpholino oligonucleotide injection led to a phenotype similar to that of *sj*: abnormal tunic and metamorphosis without tail resorption (Sasakura et al., 2005). These results indicate that the *Ci-CesA* gene, the cellulose synthase protein, and the product cellulose not only are responsible for formation of the covering structures but also contribute to the control of metamorphic events. Although the metamorphic events were altered, the *sj* mutant could keep growing until sperm maturation, showing that the *CesA* gene and a cellulose-containing tunic were not necessary for the survival and maturation (Sasakura et al., 2005). In addition, the affected adult tunic was very soft, indicating that cellulose contributes to the physical strength of the tunic (Sasakura et al., 2005).

The ascidians show a sessile adult stage, which is not seen in extant vertebrates or cephalochordates. The *Ciona CesA* gene affects the secretion of tunic and regulation of metamorphosis; these are both important in the evolution of a sessile lifestyle of ascidians (Sasakura, 2018a).

In addition to the well-reported tunicate *CesA* gene, in another ascidian species, leathery sea squirt *Styela clava*, three cold-shock protein genes had been described to have originated from horizontal gene transfer (Wei et al., 2020). The expression of *S. clava* cold-shock protein genes responds to cold temperature and may provide low-temperature stress response (Wei et al., 2020). Furthermore, the *rusticalin* gene, which can be discovered in invertebrate chordates, coral, and placozoan animals, was also proposed to have entered tunicate genomes by horizontal gene transfer event (Daugavet et al., 2019).

(4) Another GH6-encoding gene in tunicate genomes

To confirm whether a cellulose synthase homolog exist in animals other than tunicates, a previous study analyzed the genomic resources that have been greatly accumulated in the recent decades (Inoue et al., 2019). In that study, no cellulose synthase gene was found in the genome of non-tunicate animal taxa. It also revealed another independent gene (a group of possibly orthologous genes) in tunicate genomes, named *GH6-1*, which encodes a GH6 domain protein (Inoue et al., 2019). The evolutionary origin and physiological role of this newly identified *GH6-1* gene was unknown.

To investigate the evolutionary origin of tunicate GH6-encoding genes, I first analyzed the sequence signature and phylogenetic relationship of tunicate *GH6-1* genes. Then, in order to understand the physiological function of tunicate *GH6-1* genes, I investigated the expression of *GH6-1* in early developmental stages of *Ciona intestinalis* type A. Further, I prepared *GH6-1* knockout animals to observe the physiological effects.

Chapter 1 Phylogenetic analyses of tunicate Glycosyl Hydrolase Family 6 genes

This chapter was based on and modified from a peer-reviewed publication (Li et al., 2020).

1.1 Introduction

Glycosyl hydrolases (also called glycoside hydrolases) are a widespread group of enzymes which hydrolyze glycosidic bonds. Glycosyl hydrolases are essential to animal physiology: examples include amylase (Buisson et al., 1987; Boehlke et al., 2015) for digestion, chitinase (Ohno et al., 2016; Patel and Goyal, 2017) for immunity and digestion, lysozymes (Blake et al., 1965; Uversky et al., 2010) for anti-bacterial defense, and hyaluronidase (Gmachl and Kreil, 1993; Modelski et al., 2014) for fertilization and bee venom. Based on amino acid sequence similarities, glycosyl hydrolases are grouped into families: as of September, 2021, there are 171 different families of glycosyl hydrolases on the Carbohydrate Active Enzymes database (Henrissat, 1991; Henrissat and Davies, 1997; Lombard et al., 2014). Meanwhile, enzymes of the same sequence-similar family may evolve to acquire new specificities of catalyzing substrates (Henrissat, 1991; Lombard et al., 2014).

The *Ciona intestinalis* cellulose synthase, Ci-CesA, is the first reported animal protein that contain a Glycosyl Hydrolase Family 6 domain (GH6, Pfam: PF01341) (Nakashima et al., 2004). Tunicate CesAs show a unique combination of a Glycosyltransferase-like family 2 (GT2) domain (Pfam PF13641, or CESA_CelA_like, Conserved Domain Database cd06421) and a GH6 domain (Matthysse et al., 2004; Nakashima et al., 2004; Sagane et al., 2010; Nakashima et al., 2011). A recent report (Inoue et al., 2019) revealed the existence of *GH6-1*, another gene group in tunicate genomes that also encodes for GH6 domain. In *Ciona intestinalis* genome, other glycosyl hydrolases (families: 9, 18, 20, 23, 38, 47, 116) had also been predicted (Lo et al., 2003; Davison and Blaxter, 2005; Intra et al., 2008).

Well-studied GH6 enzymes are cellulases, catalyzing cellulose hydrolysis (Henrissat et al., 1998; Koivula et al., 2002; Nakashima et al., 2004). Many herbivorous or xylophagous animals rely on symbiotic microorganisms, which have endogenous cellulases, to utilize the cellulose in the diet (Watanabe and Tokuda, 2001). Some animals were described to have acquired cellulase from horizontal gene transfer (Watanabe and Tokuda, 2001; Boto, 2014). Another gene family, Glycosyl Hydrolase Family 9, encoding for possible cellulases in a different family of glycosyl hydrolases, was found in five animal phyla (Lo et al., 2003; Davison and Blaxter, 2005). The sequence similarity and intron positions of these Glycosyl Hydrolase Family 9 genes led to the suggestion that a cellulase existed in a Metazoa ancestor (Lo et al., 2003; Davison and Blaxter, 2005). However, none of these cellulases of other animals belongs to the GH6 family. Until now, GH6-domain-containing genes have been found only in bacteria, fungi, tunicates, and a few other non-animal eukaryotes.

A hypothesis regarding the origin of tunicate *CesA* genes had been proposed (Nakashima et al., 2004): a bacterial genomic region that contained both a *GT2/CesA* gene and a GH6 gene, was transferred horizontally to ancestral tunicates, and that the two genes/domains later merged to form the tunicate *CesA* gene. This hypothesis was further strengthened, when it was observed that actinobacterial genomes contain sequences of high guanine-cytosine content (GC-rich) that can be transformed into enhancers in the tunicate cellular environment (Sasakura et al., 2016).

Because of sequence divergence between the GH6 domain of tunicate CesA proteins (tunicate CesA-GH6) and GH6 proteins of other organisms, previous studies could not determine their evolutionary relationships (Nakashima et al., 2004; Sagane et al., 2010). Together with the newly found tunicate *GH6-1*, the evolutionary relationship of the tunicate GH6-encoding genes with other GH6 genes remains uncertain.

In eukaryotes, conservation of splice sites (location of boundaries between exons and introns) is often found among orthologous genes (Rogozin et al., 2003; Putnam et al., 2007). Assuming that tunicate GH6-containing genes were transferred horizontally from bacteria, acquisition of spliceosomal introns in tunicate *CesA-GH6* or *GH6-1* genes could be interpreted as a eukaryote-specific character (Nixon et al., 2002; Patel and Steitz, 2003). A previous survey (Bhattachan and Dong, 2017) found that no splice sites were shared between the tunicate *CesA* genes and plant cellulose synthase genes; therefore, it was concluded that ancient *CesA* genes without introns transferred into tunicate genomes and plant genomes independently.

The foregoing findings raised the question of how the tunicate ancestor acquired the precursor of the *CesA-GH6* and *GH6-1* genes. Three possible evolutionary scenarios have been proposed (Figure 1.1) (Nakashima et al., 2004; Inoue et al., 2019). Scenario 1: Two *GH6* genes were transferred, one of which merged with a GT2-containing gene from the same prokaryote genomic region transferred to an ancestral tunicate and formed the tunicate *CesA* gene. The second *GH6* gene gave rise to the current *GH6-1*. Scenario 2: A *GH6* gene was transferred and duplicated. After a single transfer of prokaryote *GT2-GH6* region into a tunicate genome, a duplication occurred. One copy did not include or retain the *GT2* part and became *GH6-1*, while the other copy (an ancient '*GH6* gene') merged with the GT2 domain and became part of tunicate *CesA* (joined GT2-GH6 domains). Scenario 3: A *GT2* gene and a *GH6* gene were transferred independently into an ancestral tunicate. The *GH6* gene duplicated thereafter. One copy of the *GH6* genes fused with the *GT2* gene to form the tunicate *CesA* gene. The other copy remained an independent *GH6-1* gene.

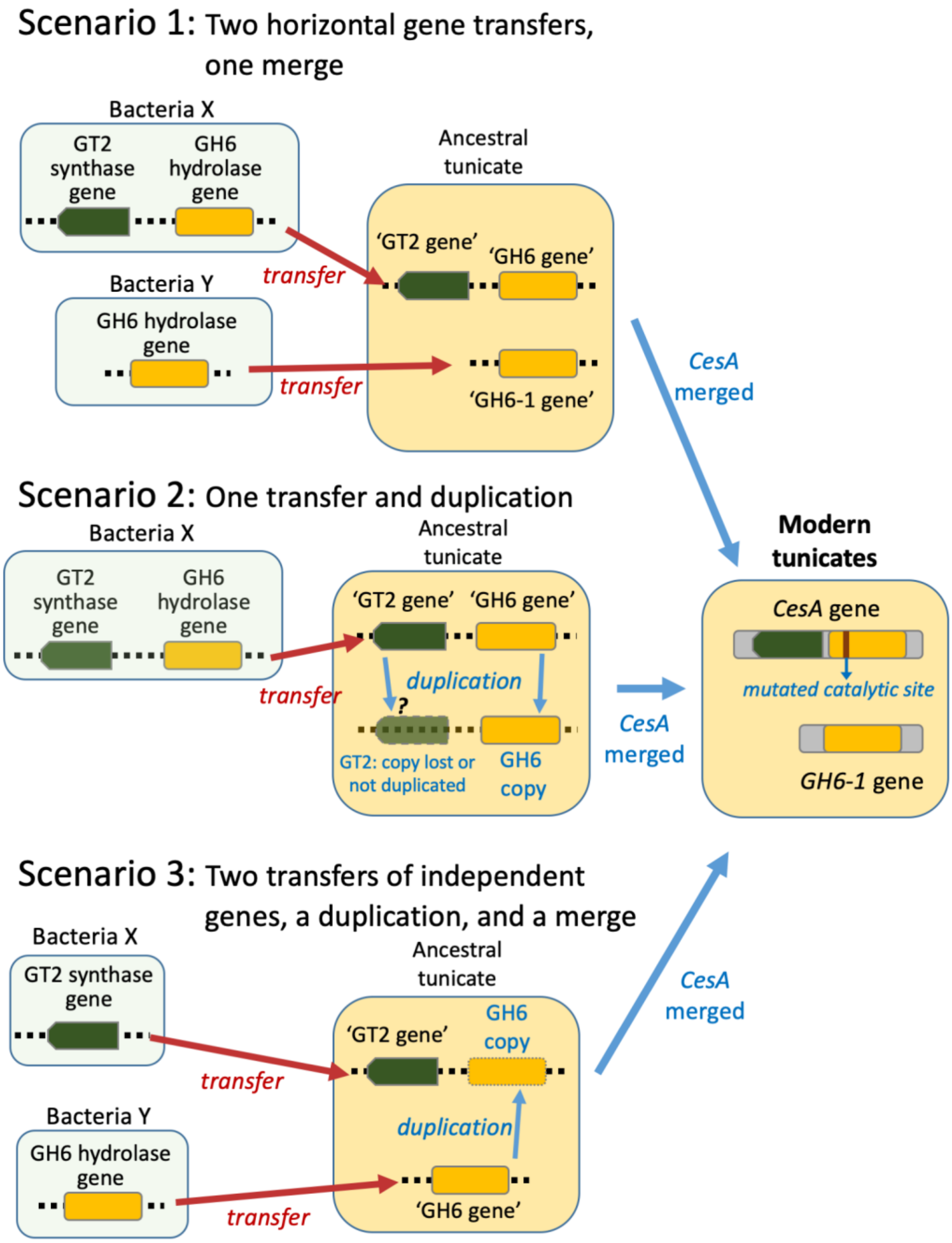

Figure 1.1. Possible scenarios on the origin of tunicate GH6 domain-containing genes

Three scenarios have been proposed to explain the existence of two GH6 domain-containing genes in extant tunicate genomes.

In this chapter, I assessed possible origins of tunicate GH6s by examining phylogenetic relationships of GH6-containing genes in diverse organisms. I also compared sequence characters and exon boundaries among tunicate GH6 domains to understand their evolutionary changes in tunicate genomes.

1.2 Methods

1.2.1 Genetic information acquisition

To reanalyze the tunicate cellulose synthase gene (*CesA*) and *GH6-1* genes and gene models [many of which were mentioned in previous studies (Table 1.1) (Matthysse et al., 2004; Nakashima et al., 2004; Sagane et al., 2010; Inoue et al., 2019)], I retrieved the corresponding gene models and genomic information from these databases: NCBI GenBank [reported genes in literatures, *Salpa thompsoni* genomic sequence assemblies GCA_001749815.1 (Jue et al., 2016) and transcriptome GFCC00000000.1 (Batta-Lona et al., 2017), and the *Ciona savignyi* transcriptome GGEI00000000.1 (Wei and Dong, 2018)], the Ghost database (Kyoto University) for *Ciona intestinalis* type A (Satou et al., 2005; Satou et al., 2008), the *Botryllus schlosseri* Genome Project (transcripts only, Stanford University) (Voskoboynik et al., 2013), the OikoBase for *Oikopleura dioica* (Denoeud et al., 2010; Danks et al., 2013), and the Aniseed database (transcripts and genomes of all other species, as well as the genomes of *C. savignyi* and *B. schlosseri*) (Tassy et al., 2010).

Table 1.1. Tunicate GH6-containing genes or gene models and related genes analyzed in this chapter

Species	**Domain Content**	**Short Name of the Gene Used in this Chapter ***	**Source Database**	**Accession/ID of Gene, Transcript, or Protein**	**Note**
Ciona intestinalis type A *(C. robusta)*	GH6	*CinGH6-1*	GenBank	XM_002119543.4 / XP_002119579.1	
	CesA+GH6	*CinCesA*	GenBank	NM_001047983.1 / BAD10864.1	As reported in (Nakashima et al., 2004)
Ciona savignyi	GH6	*CsaGH6-1*	GenBank (Transcriptome)	GGEI01013363.1	
	CesA+GH6	*CsaCesA*	GenBank	AY504665.1 / AAR89623.1	As reported in (Matthysse et al., 2004)
Salpa thompsoni	GH6	*SthGH6-1a*	GenBank (Transcriptome)	GFCC01117283.1	Possible lineage-specific duplication
	GH6	*SthGH6-1b*	GenBank (Transcriptome)	GFCC01119318.1	No possible catalytic Asp; possible lineage-specific duplication.
	CesA+GH6	*SthCesA*	GenBank (Transcriptome)	GFCC01072613.1	
Molgula occidentalis	GH6	*MoxGH6-1*	Aniseed database	Moocci.CG.ELv1_2.S285391.g07021.01.t	
	CesA+GH6	*MoxCesAa*	Aniseed database	Moocci.CG.ELv1_2.S469068.g15915.01.t	Short GH6 part
	GH6	*(MoxCesAbGH6)*	Aniseed database	Moocci.CG.ELv1_2.S469068.g15914.01.t	Very short
	GH6	*MoxCesAcGH6*	Aniseed database	Moocci.CG.ELv1_2.S469068.g15913.01.t	
Molgula oculata	GH6	*MocGH6-1*	Aniseed database	Moocul.CG.ELv1_2.S112948.g12660.01.t	
	CesA+GH6	*MocCesAa*	Aniseed database	Moocul.CG.ELv1_2.S71617.g04842.01.t	Rhodopsin-like GPCR domain at upstream part
	GH6	*MocCesAbGH6*	Aniseed database	Moocul.CG.ELv1_2.S69739.g04625.01.t	

* Gene names were assigned after considering phylogenetic information examined in this study and in that by Inoue et al. (2019).

Table 1.1. Tunicate GH6-containing genes or gene models and related genes analyzed in this chapter (continued)

Species	**Domain Content**	**Short Name of the Gene Used in this Chapter ***	**Source Database**	**Accession/ID of Gene, Transcript, or Protein**	**Note**
Botryllus schlosseri	GH6	*BscGH6-1*	*Botryllus schlosseri* Genome Project	g9326	
	GH6	*(BscGH6-1b)*	*Botryllus schlosseri* Genome Project	g61144	Short, similar to BscGH6-1
	GH6	*BscCesAaGH6*	*Botryllus schlosseri* Genome Project	g44331	Similar to BscCesAbGH6 (89.6% identity in the matching 222 AA region)
	GH6	*BscCesAbGH6*	*Botryllus schlosseri* Genome Project	g45080	Similar to BscCesAaGH6
Botrylloides leachii	GH6	*BleGH6-1*	Aniseed database	Boleac.CG.SB_v3.S133.g02304.01.t	
	CesA+GH6	*BleCesA*	Aniseed database	Boleac.CG.SB_v3.S157.g03251.01.t	
Oikopleura dioica	GH6	*OdiGH6-1*	OikoBase/GenBank	GSOIDT00010490001 / CBY09680.1	
	GH6	*(OdiGH6-1b)*	OikoBase/GenBank	GSOIDT00021901001 / CBY33927.1	98% identical to OdiGH6-1
	CesA+GH6	*OdiCesA2*	GenBank	AB543593.1 / BAJ65326.1	As reported in (Sagane et al., 2010; Nakashima et al., 2011)
	CesA+GH6	*OdiCesA1*	GenBank	AB543594.1 / BAJ65327.1	As reported in (Sagane et al., 2010; Nakashima et al., 2011)

* Gene names were assigned after considering phylogenetic information examined in this study and in that by Inoue et al. (2019).

Although the recorded transcripts or annotated gene models were retrieved, I wished to confirm whether there is any hidden GH6-encoding genetic information that failed to be annotated as a gene model in each tunicate genome. Therefore, I recorded the genomic location (the coordinates on chromosomes, scaffolds, or contigs) of each predicted *GH6-1* and *CesA* gene. When the genomic locations of transcript/models were unknown, as in the cases of *C. savignyi*, *S. thompsoni*, and *O. dioica*, the GH6-containing transcripts were used to search (blastn in BLAST, Basic Local Alignment Search Tool, using default parameters) against its corresponding genome/genomic assembly: the databases used were listed as above. The genomic locations of tunicate GH6-containing genes were listed in Table 1.2. Next, I used the GH6 domains in *C. intestinalis* type A predicted proteins of CesA (GenBank: BAD10864.1) and GH6-1 (NCBI: XP_002119579.1) as queries to search (tblastn in BLAST, with default parameters, e-value threshold = 1×10^{-10}) against the other seven tunicates' genomic databases or assemblies and used *O. dioica* predicted proteins (GH6-1, CBY09680.1 and CesA2, BAJ65326.1) to search (tblastn, with default parameters, e-value threshold = 1×10^{-10}) the *C. intestinalis* type A genome and recorded the genomic locations of results. When I used *C. intestinalis* CesA-GH6 or GH6-1 sequences to search *C. intestinalis* genome, the only results passed the threshold were the same genomic areas encoding these two genes. Therefore, I used *O. dioica* predicted proteins to search against *C. intestinalis* genome to confirm that there is no other GH6-like sequence in *C. intestinalis* genome. I found that the location of retrieved transcripts/gene models mostly matched with the BLAST search (tblastn) results, with minor exceptions: a few additional open reading frames (ORF) or short gene models were newly discovered. For example, an ORF of *M. oculata* coding for a 39-amino-acid (AA) peptide and a gene model of *B. schlosseri*, Boschl.CG.Botznik2013.chr9.g44329, coding for a 166-AA peptide, were found in BLAST searches. These short peptides/gene models have similar sequences to a GH6 domain, but those are either far shorter (less than 140 AA) than a typical GH6 domain (Pfam PF01341, with sizes of around 300 AA) or were evaluated as 'no significance' in protein profile searches [hmmscan, HmmerWeb version 2.41.1 (Potter et al., 2018), searched against the Pfam database]. Therefore, I interpreted that there is no better hidden representative of GH6 genes in these genomes.

Table 1.2. Selected tunicate GH6-containing genes (models) and genomic locations

Species, **Genome version**	**Domain content**	**Short name of the gene used in this manuscript**	**Accession/ID of gene or transcript**	**Chromosome or genomic scaffold location**
Ciona intestinalis type A *(C. robusta),* HT Reference Genome (2019 version, Ghost database)	GH6	*CinGH6-1*	XM_002119543.4	Chr3: 2879389-2883393
	CesA+GH6	*CinCesA*	NM_001047983.1	Chr7: 3393711-3406917
Ciona savignyi, Ciona_savignyi_ENS81_Genome (Aniseed database)	GH6	*CsaGH6-1*	GGEI01013363.1	Reftig R35: 1310928-1318413
	CesA+GH6	*CsaCesA*	AY504665.1	Reftig R2: 2893501-2913534
Salpa thompsoni, GenBank assembly: GCA_001749815.1 (NCBI)	GH6	*SthGH6-1a*	GFCC01117283.1	Scaffolds: 3051, 10336, and 13318
	GH6	*SthGH6-1b*	GFCC01119318.1	Scaffolds: 26886, contig 211572 and 272169
	CesA+GH6	*SthCesA*	GFCC01072613.1	Scaffolds: 5822, 14051, 39468, 41682, 48268
Molgula occidentalis, Molgula_occidentalis_ELv12_Genome (Aniseed database)	GH6	*MoxGH6-1*	Moocci.CG.ELv1_2.S285391.g07021.01.t	Scaffold 285391: 1115-4105
	CesA+GH6	*MoxCesAa*	Moocci.CG.ELv1_2.S469068.g15915.01.t	Scaffold 469068: 7725-12872
	GH6	*(MoxCesAbGH6)*	Moocci.CG.ELv1_2.S469068.g15914.01.t	Scaffold 469068: 6382-7088
	GH6	*MoxCesAcGH6*	Moocci.CG.ELv1_2.S469068.g15913.01.t	Scaffold 469068: 2541-4008
Molgula oculata, Molgula_oculata_ELv12_Genome (Aniseed database)	GH6	*MocGH6-1*	Moocul.CG.ELv1_2.S112948.g12660.01.t	Scaffold 112948: 22155-24431
	CesA+GH6	*MocCesAa*	Moocul.CG.ELv1_2.S71617.g04842.01.t	Scaffold 71617: 7004-16648
	GH6	*MocCesAbGH6*	Moocul.CG.ELv1_2.S69739.g04625.01.t	Scaffold 69739: 3301-4739
Botryllus schlosseri, Botryllus_schlosseri_botznik2013_Genome (Aniseed database); *Botryllus schlosseri* Genome Project (transcript GFF information, Stanford University)	GH6	*BscGH6-1*	g9326/Boschl.CG.Botznik2013.chr9.g09326.01.t	Chr9: 18,466,231-18,474,930
	GH6	*(BscGH6-1b)*	g61144 (only in the genome project transcript fasta/GFF)	Contig botctg111009: 2832-3434
	GH6	*BscCesAaGH6*	g44331/Boschl.CG.Botznik2013.chr9.g44331.01.t	Chr9: 15,780,634-15,782,886
	GH6	*BscCesAbGH6*	g45080/Boschl.CG.Botznik2013.chr13.g45080.01.t	Chr13: 9,500,441-9,502,510
Botrylloides leachii, Botrylloides_leachii_SBv3_Genome (Aniseed database)	GH6	*BleGH6-1*	Boleac.CG.SB_v3.S133.g02304.01.t	Scaffold 133: 25041-28636
	CesA+GH6	*BleCesA*	Boleac.CG.SB_v3.S157.g03251.01.t	Scaffold S157: 172941-185066
Oikopleura dioica, Odioica_Assembly_reference_unmasked_v3.0 (OikoBase)	GH6	*OdiGH6-1*	GSOIDT00010490001	Scaffold 33: 182048-186143
	GH6	*(OdiGH6-1b)*	GSOIDT00021901001	Scaffold 33: 182048-185894

	CesA+GH6	*OdiCesA2*	AB543593.1	Scaffold 314: 1902-7896
	CesA+GH6	*OdiCesA1*	AB543594.1	Scaffold 80: 135240-139503

After retrieving tunicate *CesA* and *GH6-1* gene models, I prepared an expanded sequence alignment including more bacterial/fungal GH6 sequences for the phylogenetic analysis. The same two *C. intestinalis* type A protein models (CesA, BAD10864.1 and GH6-1, XP_002119579.1) were used as queries to perform BLAST searches of the NCBI non-redundant protein sequences database (nr). The blastp (protein-protein BLAST) algorithm was selected, with default parameters (word size = 6; matrix = BLOSUM62; gap cost existence:11, extension:1; conditional compositional score matrix adjustment). A strategy was used to achieve broad sampling of GH6-containing proteins across different taxa. First, the query was used to search all nr sequences excluding tunicates, and the results with the lowest e-values were all sequences from the genus *Streptomyces*. A second search was carried out against "all data excluding tunicates and *Streptomyces.*" Several subsequent searches were performed stepwise, excluding higher taxa (Streptomycetales, Actinobacteria, or Bacteria). Another approach was to search only "Archaea", "Fungi", or "Eukaryotes, excluding tunicates and fungi." A GH6 protein (NCBI: WP_094052291.1) from *Streptomyces* was also used as a query to expand the search results in several eukaryotic taxa (Table 1.3). However, two questionable 'eukaryotic' results, showing higher similarity to bacterial proteins and linkages to other probable bacterial genes, were excluded (Table 1.3). A few selected bacterial and fungal sequences that were used in a previous phylogenetic analysis (Sagane et al., 2010) were also included in later analyses. In search results, some long sequences included conserved domains other than GH6, which were confirmed using InterPro searches (online searches against all available databases). Those extra domains were excluded before downstream analyses. All the selected sequences (listed in Table 1.1 and Table 1.4) contained a GH6 domain (Pfam: PF01341), which was confirmed by a HMMER hmmscan examination [HmmerWeb version 2.41.1 (Potter et al., 2018), searched against the Pfam database]; a GH6 domain in each sequence was identified with an Independent E-value smaller than 1×10^{-5}.

Table 1.3. Existence of GH6 proteins in different taxa

Taxa				**GH6 presence?**
Bacteria				**Present**
Archaea				Not yet observed
Eukaryota	Opisthokonta	Metazoa	Tunicates	**Present**
			Metazoa, except tunicate	No? Contamination? *1
		Fungi		**Present**
		Opisthokonta, except Metazoa and fungi		Not yet observed
	Viridiplantae			No? Contamination? *2
	SAR-Stramenopiles			**Present**
	SAR-Alveolate			**Present**
	SAR-Rhizaria			Not yet observed
	Haptista			**Present**
	Rhodophyta			**Present**
	Other eukaryotes			Not yet observed

*1: A GH6 protein in the *Lucilia cuprina* (a dipteran) genome project, XP_023300643.1, was very similar to bacterial GH6 proteins. It was located at a genomic scaffold that contained other probable bacterial genes. *2: A GH6 protein found in the *Gossypium hirsutum* (upland cotton) genome project, XP_016733546.1, was highly similar to bacterial GH6 proteins and it was located at a genomic scaffold that contained other probable bacterial genes. The above two cases were the only results that contained GH6 domains in each search. I treated these two cases as bacterial contaminants.

Table 1.4. Other GH6-containing proteins used in phylogenetic analyses

Taxa	**(taxa-) Genus**	**Accession Number**
Actinobacteria, except *Streptomyces*	unidentified Actinobacteria	WP_054220290.1
	Actinoplanes	WP_043523455.1
	Actinoplanes	WP_043525061.1
	Brachybacterium	WP_012804931.1
	Cellulomonas	P07984.1
	Geodermatophilus	PYG45136.1
	Geodermatophilus	WP_163476570.1
	Kitasatospora	WP_030394430.1
	Mycobacteriaceae	WP_011562673.1
	Mycobacterium	YP_001848433.1
	Mycobacterium	YP_879613.1
	Mycolicibacterium	WP_005142664.1
	Nocardioides	WP_011758040.1
	Paraoerskovia	SDS23371.1
	Paraoerskovia	WP_043109784.1
	Saccharopolyspora	WP_009943607.1
	Streptacidiphilus	WP_034091765.1
	Thermobispora	P26414.1
	Zhihengliuella	WP_130448960.1
Bacteria, except Actinobacteria	*Cystobacter*	WP_095990604
	Granulicella	WP_089840334.1
	Myxococcus	WP_090493059.1
	Myxococcus	WP_140855929.1
	Plesiocystis	ZP_01907667.1
	Sorangium	YP_001618727.1
	Uncultured bacterium	AHL27895.1
	Vitiosangium	WP_108069950.1
Eukaryotes, except fungi and tunicates	Alveolata- *Stylonychia*	CDW82212.1
	Alveolata- *Symbiodinium*	OLP73243.1
	Haptista- *Chrysochromulina*	KOO25121.1
	Haptista- *Chrysochromulina*	KOO25881.1
	Rhodophyta- *Chondrus*	XP_005713951.1
	Rhodophyta- *Chondrus*	XP_005717841.1
	Rhodophyta- *Gracilariopsis*	PXF45697.1
	Stramenopiles- *Achlya*	OQR88682.1
	Stramenopiles- *Aphanomyces*	KAF0695776.1
	Stramenopiles- *Saprolegnia*	XP_008620974.1
	Stramenopiles- *Thraustotheca*	OQS00291.1
Fungi	*Neocallimastix*	ORY54114.1
	Neocallimastix	ORY77883.1
	Orpinomyces	AAB92678
	Orpinomyces	AAB92679.1
	Piromyces	AAP30749.1
	Piromyces	OUM68810.1
	Talaromyces	APE61639.1
	Talaromyces	ATQ35966.1
	Talaromyces	BAA74458.1
	Termitomyces	KNZ78466.1
	Volvariella	AAT64008.1
***Streptomyces* (genus)**	*Streptomyces*	AAA26776.1
	Streptomyces	WP_033308106.1
	Streptomyces	WP_079662022.1
	Streptomyces	WP_093802936.1
	Streptomyces	WP_094052291.1
	Streptomyces	WP_099499477.1
	Streptomyces	WP_120721336.1
	Streptomyces	WP_150216144.1
	Streptomyces	WP_156692247.1

1.2.2 Bioinformatic analyses

For phylogenetic analyses, the abovementioned GH6-containing proteins (or protein parts) were used to build a multiple sequence alignment with MAFFT v7 online server (strategy: L-INS-I iterative refinement; recommended for <200 sequences with one conserved domain and long gaps) (Katoh and Standley, 2013; Katoh et al., 2019). Poorly aligned regions were removed using trimAl v1.2 (Capella-Gutierrez et al., 2009) when more than 65% of the selected sequences showed gaps in a given position. The appropriate amino acid substitution model was selected using Prottest 3.4.2 (with default parameters) (Darriba et al., 2011) before a maximum likelihood phylogenetic analysis. Phylogenetic reconstructions were performed with MrBayes 3.2.7a (nucmodel = protein, aamodelpr = mixed, ngen = 2500000, nchains = 1) (Ronquist et al., 2012) or RAxML-HPC Blackbox v8.2.12 (substitution model: PROTCATWAGF, rapid bootstrap with automatic bootstopping) (Stamatakis, 2014) via CIPRES Science Gateway (Miller et al., 2010). Consensus trees were visualized with FigTree software (Rambaut).

1.2.3 Sequence comparison

To understand the similarity or difference of tunicate GH6 domains and GH6 proteins of other organisms, signatures of a few tunicate GH6 proteins and a *Hypocrea jecorina* protein (UniProtKB: P07987.1) were compared with signature information on PROSITE (Sigrist et al., 2012). For signal peptide prediction of *Ciona intestinalis* GH6-1 protein, the SignalP-5.0 Server (Almagro Armenteros et al., 2019), selecting the 'eukarya' computation model, was used. Regarding the exon-intron structures, some genes or gene models in the databases had been annotated with exon boundaries. When exon information of genes or gene models was unknown, sequences of transcripts were used to search (blastn of NCBI BLAST, with default parameters) against the corresponding genomic databases: the Ghost database (Satou et al., 2005; Satou et al., 2019) for *C. intestinalis* type A, NCBI genome assembly GCA_001749815.1 for *S. thompsoni* genomic assembly (Jue et al., 2016), OikoBase (Danks et al., 2013) for *O. dioica*, and the Aniseed database BLAST tool for other tunicate species (Tassy et al., 2010). Coding parts of transcripts and genomic sequences were then compared with the Splign utility at NCBI (Kapustin et al., 2008), with default parameters. Tunicate GH6-containing proteins were aligned with MAFFT v7 server (strategy: L-INS-I) (Katoh and Standley, 2013; Katoh et al., 2019) for splice site (exon-boundary) comparisons.

1.3 Results

1.3.1 Tunicate CesA-GH6 domains and tunicate GH6-1 represent two independent monophyletic groups

To determine whether *GH6-1* genes represent a monophyletic group distinct from tunicate *CesA* genes and to understand the relationship of GH6-1 with GH6 proteins in other organisms, I used amino acid sequences of eight predicted tunicate CesA-GH6 domains, sequences of eight predicted GH6-1 proteins, and many predicted GH6 protein sequences from bacteria, fungi, and various eukaryotes (Table 1.4) to reconstruct phylogenetic trees (Figure 1.2 and Figure 1.3). Both Bayesian inference (Figure 1.2) and maximum likelihood (Figure 1.3) approaches provided trees supporting a close relationship of tunicate CesA-GH6 and GH6-1. In addition, CesA-GH6 sequences and GH6-1 sequences formed two separate clusters, although the maximum likelihood bootstrap support values were only 83% for the GH6-1 clade and 61% for the CesA-GH6 clade.

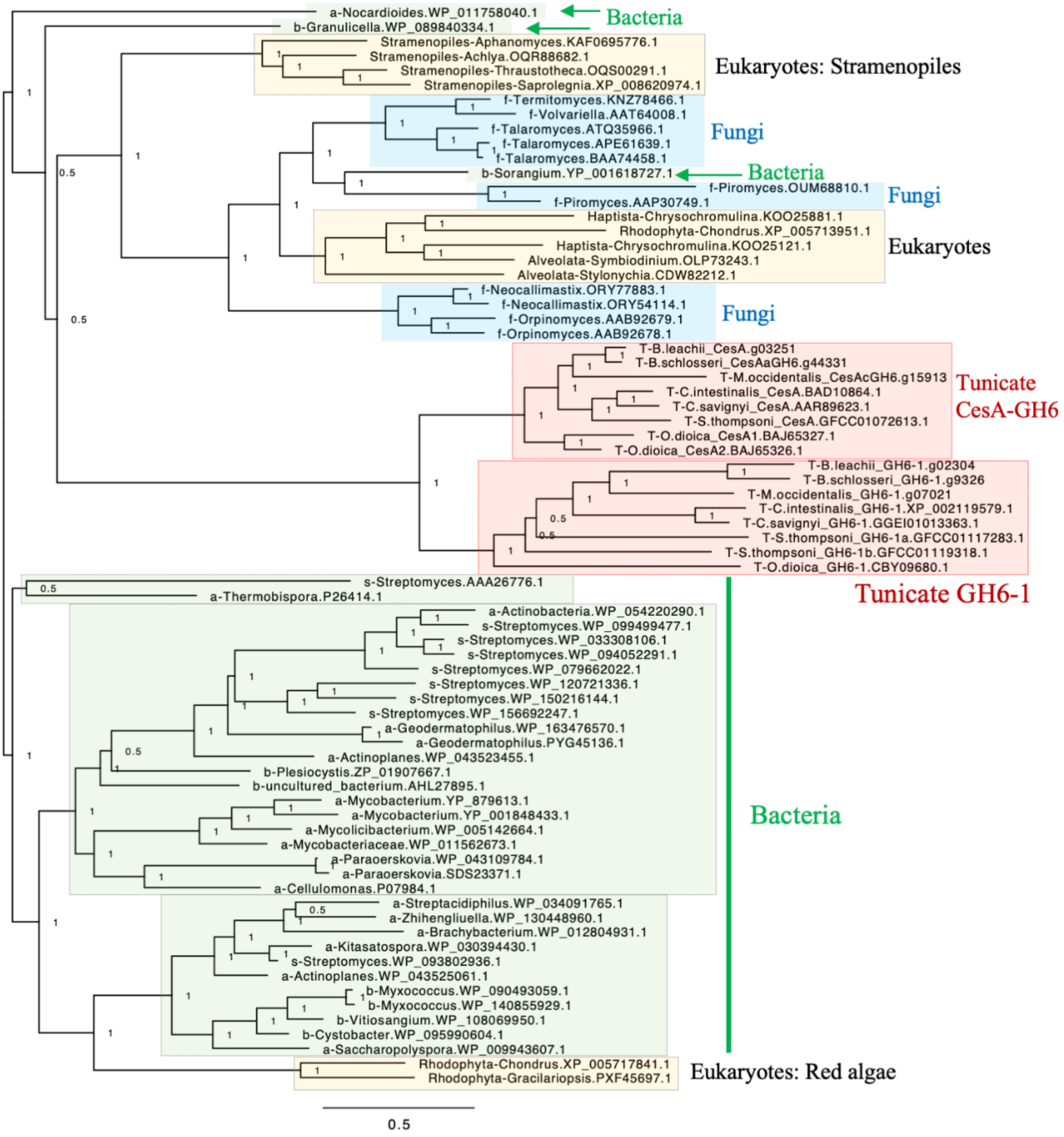
a-Nocardioides.WP_011758040.1
b-Granulicella.WP_089840334.1
Bacteria
Stramenopiles-Aphanomyces.KAF0695776.1
Stramenopiles-Achlya.OQR88682.1
Stramenopiles-Thraustotheca.OQS00291.1
Stramenopiles-Saprolegnia.XP_008620974.1
Eukaryotes: Stramenopiles
f-Termitomyces.KNZ78466.1
f-Volvariella.AAT64008.1
f-Talaromyces.ATQ35966.1
f-Talaromyces.APE61639.1
f-Talaromyces.BAA74458.1
Fungi
b-Sorangium.YP_001618727.1
Bacteria
f-Piromyces.OUM68810.1
f-Piromyces.AAP30749.1
Fungi
Haptista-Chrysochromulina.KOO25881.1
Rhodophyta-Chondrus.XP_005713951.1
Haptista-Chrysochromulina.KOO25121.1
Alveolata-Symbiodinium.OLP73243.1
Alveolata-Stylonychia.CDW82212.1
Eukaryotes
f-Neocallimastix.ORY77883.1
f-Neocallimastix.ORY54114.1
f-Orpinomyces.AAB92679.1
f-Orpinomyces.AAB92678.1
Fungi
T-B.leachii_CesA.g03251
T-B.schlosseri_CesAaGH6.g44331
T-M.occidentalis_CesAcGH6.g15913
T-C.intestinalis_CesA.BAD10864.1
T-C.savignyi_CesA.AAR89623.1
T-S.thompsoni_CesA.GFCC01072613.1
T-O.dioica_CesA1.BAJ65327.1
T-O.dioica_CesA2.BAJ65326.1
Tunicate CesA-GH6
T-B.leachii_GH6-1.g02304
T-B.schlosseri_GH6-1.g9326
T-M.occidentalis_GH6-1.g07021
T-C.intestinalis_GH6-1.XP_002119579.1
T-C.savignyi_GH6-1.GGEI01013363.1
T-S.thompsoni_GH6-1a.GFCC01117283.1
T-S.thompsoni_GH6-1b.GFCC01119318.1
T-O.dioica_GH6-1.CBY09680.1
Tunicate GH6-1
s-Streptomyces.AAA26776.1
a-Thermobispora.P26414.1
a-Actinobacteria.WP_054220290.1
s-Streptomyces.WP_099499477.1
s-Streptomyces.WP_033308106.1
s-Streptomyces.WP_094052291.1
s-Streptomyces.WP_079662022.1
s-Streptomyces.WP_120721336.1
s-Streptomyces.WP_150216144.1
s-Streptomyces.WP_156692247.1
a-Geodermatophilus.WP_163476570.1
a-Geodermatophilus.PYG45136.1
a-Actinoplanes.WP_043523455.1
b-Plesiocystis.ZP_01907667.1
b-uncultured_bacterium.AHL27895.1
a-Mycobacterium.YP_879613.1
a-Mycobacterium.YP_001848433.1
a-Mycolicibacterium.WP_005142664.1
a-Mycobacteriaceae.WP_011562673.1
a-Paraoerskovia.WP_043109784.1
a-Paraoerskovia.SDS23371.1
a-Cellulomonas.P07984.1
Bacteria
a-Streptacidiphilus.WP_034091765.1
a-Zhihengliuella.WP_130448960.1
a-Brachybacterium.WP_012804931.1
a-Kitasatospora.WP_030394430.1
s-Streptomyces.WP_093802936.1
a-Actinoplanes.WP_043525061.1
b-Myxococcus.WP_090493059.1
b-Myxococcus.WP_140855929.1
b-Vitiosangium.WP_108069950.1
b-Cystobacter.WP_095990604.1
a-Saccharopolyspora.WP_009943607.1
Rhodophyta-Chondrus.XP_005717841.1
Rhodophyta-Gracilariopsis.PXF45697.1
Eukaryotes: Red algae
0.5

Figure 1.2. Phylogenetic relationship of GH6-containing proteins by Bayesian inference

(The figure legend is shown at the next page)

All tunicate sequences formed a cluster. The cluster was further divided into two subclusters of CesA-GH6 domains and GH6-1 proteins. However, the clustering of tunicate GH6 sequences with GH6 proteins of other organisms was not well-supported. Rooting was arbitrary. Numbers next to internal nodes or branches represent posterior probabilities of the neighboring branch. The same trimmed multiple sequence alignment was used as input for this analysis and the analysis by maximum likelihood method. Bayesian inference was performed with MrBayes using a mixed substitution model (aamodelpr = mixed). The analysis was terminated after 2,500,000 generations as the standard deviation of split frequencies remained as a stable 0.126917 after generation 1,830,000, although this analysis could not reach an ideal convergence due to short sequence lengths and divergent data. The starting part of sequence names represents its source organism category, in alphabetical order: a, Actinobacteria, excluding Streptomyces; b, Bacteria excluding Actinobacteria; f, fungi; s, genus Streptomyces; T, tunicates. Scales represent expected changes per site.

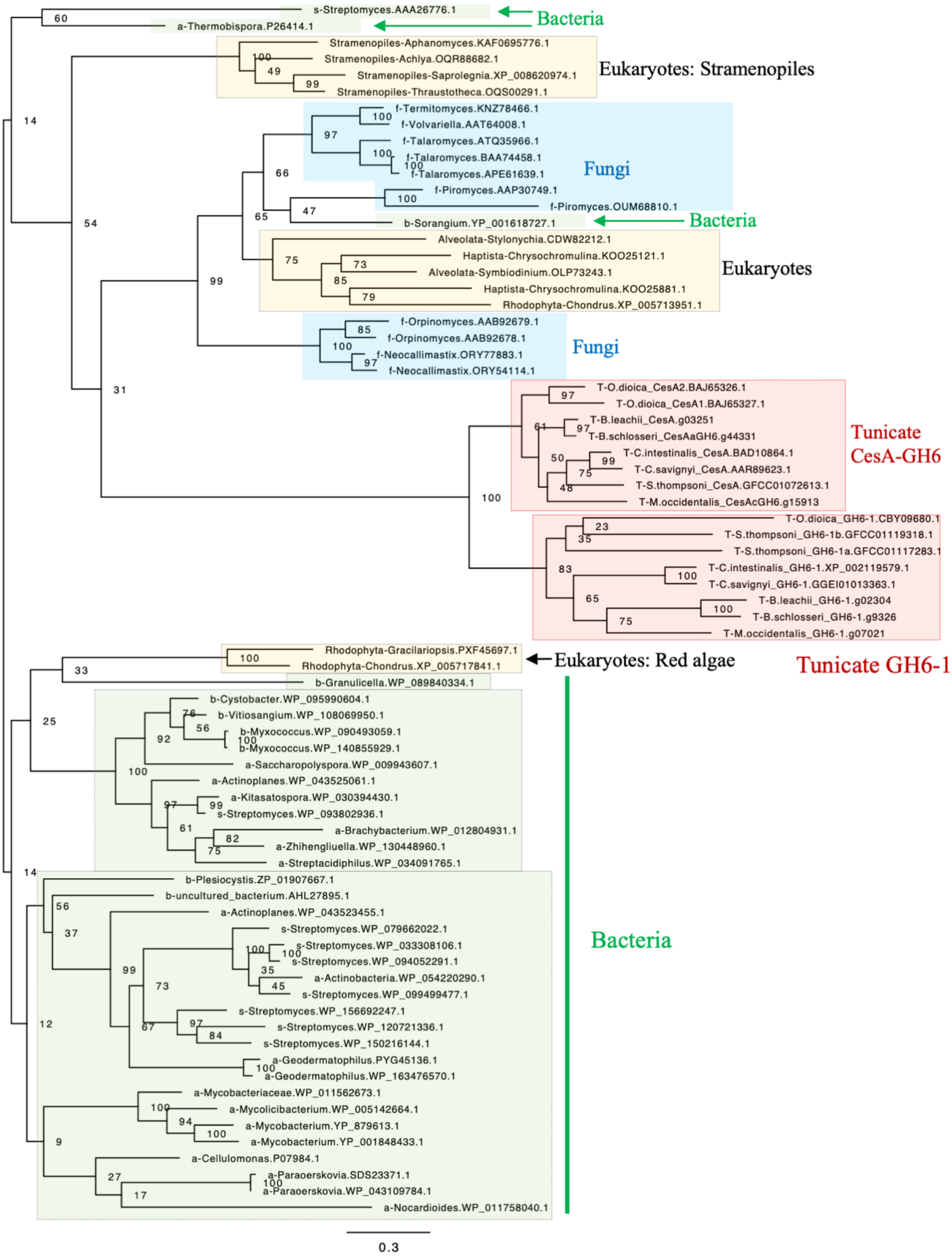
s-Streptomyces.AAA26776.1
a-Thermobispora.P26414.1
Bacteria
Stramenopiles-Aphanomyces.KAF0695776.1
Stramenopiles-Achlya.OQR88682.1
Stramenopiles-Saprolegnia.XP_008620974.1
Stramenopiles-Thraustotheca.OQS00291.1
Eukaryotes: Stramenopiles
f-Termitomyces.KNZ78466.1
f-Volvariella.AAT64008.1
f-Talaromyces.ATQ35966.1
f-Talaromyces.BAA74458.1
f-Talaromyces.APE61639.1
Fungi
f-Piromyces.AAP30749.1
f-Piromyces.OUM68810.1
b-Sorangium.YP_001618727.1
Bacteria
Alveolata-Stylonychia.CDW82212.1
Haptista-Chrysochromulina.KOO25121.1
Alveolata-Symbiodinium.OLP73243.1
Haptista-Chrysochromulina.KOO25881.1
Rhodophyta-Chondrus.XP_005713951.1
Eukaryotes
f-Orpinomyces.AAB92679.1
f-Orpinomyces.AAB92678.1
f-Neocallimastix.ORY77883.1
f-Neocallimastix.ORY54114.1
Fungi
T-O.dioica_CesA2.BAJ65326.1
T-O.dioica_CesA1.BAJ65327.1
T-B.leachii_CesA.g03251
T-B.schlosseri_CesAaGH6.g44331
T-C.intestinalis_CesA.BAD10864.1
T-C.savignyi_CesA.AAR89623.1
T-S.thompsoni_CesA.GFCC01072613.1
T-M.occidentalis_CesAcGH6.g15913
Tunicate
CesA-GH6
T-O.dioica_GH6-1.CBY09680.1
T-S.thompsoni_GH6-1b.GFCC01119318.1
T-S.thompsoni_GH6-1a.GFCC01117283.1
T-C.intestinalis_GH6-1.XP_002119579.1
T-C.savignyi_GH6-1.GGEI01013363.1
T-B.leachii_GH6-1.g02304
T-B.schlosseri_GH6-1.g9326
T-M.occidentalis_GH6-1.g07021
Rhodophyta-Gracilariopsis.PXF45697.1
Rhodophyta-Chondrus.XP_005717841.1
Eukaryotes: Red algae
Tunicate GH6-1
b-Granulicella.WP_089840334.1
b-Cystobacter.WP_095990604.1
b-Vitiosangium.WP_108069950.1
b-Myxococcus.WP_090493059.1
b-Myxococcus.WP_140855929.1
a-Saccharopolyspora.WP_009943607.1
a-Actinoplanes.WP_043525061.1
a-Kitasatospora.WP_030394430.1
s-Streptomyces.WP_093802936.1
a-Brachybacterium.WP_012804931.1
a-Zhihengliuella.WP_130448960.1
a-Streptacidiphilus.WP_034091765.1
b-Plesiocystis.ZP_01907667.1
b-uncultured_bacterium.AHL27895.1
a-Actinoplanes.WP_043523455.1
s-Streptomyces.WP_079662022.1
s-Streptomyces.WP_033308106.1
s-Streptomyces.WP_094052291.1
a-Actinobacteria.WP_054220290.1
s-Streptomyces.WP_099499477.1
s-Streptomyces.WP_156692247.1
s-Streptomyces.WP_120721336.1
s-Streptomyces.WP_150216144.1
a-Geodermatophilus.PYG45136.1
a-Geodermatophilus.WP_163476570.1
Bacteria
a-Mycobacteriaceae.WP_011562673.1
a-Mycolicibacterium.WP_005142664.1
a-Mycobacterium.YP_879613.1
a-Mycobacterium.YP_001848433.1
a-Cellulomonas.P07984.1
a-Paraoerskovia.SDS23371.1
a-Paraoerskovia.WP_043109784.1
a-Nocardioides.WP_011758040.1
0.3

Figure 1.3. Phylogenetic relationship of GH6-containing proteins by maximum likelihood method

(The figure legend is shown at the next page)

Similar to the previous Bayesian inference analysis, all tunicate sequences formed a cluster. The cluster was further divided into two subclusters of CesA-GH6 domains and GH6-1 proteins. However, the clustering of tunicate GH6 sequences with GH6 proteins of other organisms was not well-supported. Rooting was arbitrary. Numbers next to internal nodes or branches represent bootstrap support of the neighboring branch. The same trimmed multiple sequence alignment was used as input for the previous Bayesian inference analyses. This maximum likelihood analysis was performed with RAxML-HPC BlackBox on CIPRES Science Gateway. The WAG amino acid substitution model with empirical base frequencies was selected and bootstrapping was automatically stopped after 804 cycles. The starting part of sequence names represents its source organism category, in alphabetical order: a, Actinobacteria, excluding *Streptomyces*; b, Bacteria excluding Actinobacteria; f, fungi; s, genus *Streptomyces*; T, tunicates. Fully-expanded trees are shown as supplementary figures. Scales represent expected changes per site.

1.3.2 The origin of tunicate GH6 domains is hard to deduce

These analyses failed to determine the relationship of tunicate sequences among other GH6 proteins. Although in these trees, tunicate sequences were clustered with many fungal GH6 proteins, some other eukaryotic GH6s [from red algae (Rhodophyta), Haptista, and the SAR supergroup], and a proteobacterial GH6 (YP_001618727.1, *Sorangium cellulosum*), the Bayesian posterior probability (Figure 1.2) and maximum likelihood bootstrap support (Figure 1.3) for this clustering were low. Notably, branches leading to tunicate sequences were longer than branches to other sequence clusters.

1.3.3 Many tunicate GH6-1 proteins maintain the probable active site, in contrast to tunicate CesA proteins

With the sequence alignment of tunicate GH6-1 and CesA proteins, I also compared their sequence signatures to those of other GH6 proteins. The enzymatic activity of Exoglucanase 2 (Cel6A) of *Hypocrea jecorina* (formerly *Trichoderma reesei*) was well characterized (Koivula et al., 2002). The aspartic acid at position 221 of *H. jecorina* Cel6A (Hje-D221) serves as the catalytic center (Koivula et al., 2002). I found that in many tunicate GH6-1 proteins, an aspartic acid can be aligned to the catalytic *H. jecorina* D221 (Figure 1.4 A), except for SthGH6-1b (E197) and OdiGH6-1 (K211). However, the catalytic aspartic acid was not conserved in tunicate CesA-GH6 parts (Figure 1.4 A). These tunicate proteins also show a sequence environment that almost matches (8–9 out of 10 amino acids) the conserved 'signature 2' of GH6 (Figure 1.4 C: PROSITE PS00656: [LIVMYA]-[LIVA]-[LIVT]-[LIV]-E-P-D-[SAL]-[LI]-[PSAG]).

Another signature of GH6 that also contributes to catalytic ability is PS00655 (Figure 1.4 B, including another important aspartic acid, D175, in the *H. jecorina* protein (Koivula et al., 2002), but this signature was not conserved in tunicate GH6-containing proteins. In the aligned region, ≤40% of amino acids matched the signature pattern.

In addition, computational prediction by the SignalP-5.0 server (Almagro Armenteros et al., 2019) indicated that the first 24 amino acids of CinGH6-1 protein is a probable Sec/SPI type signal peptide (likelihood 0.5725, larger than the recommended threshold of 0.5), which is frequently seen in secreted proteins.

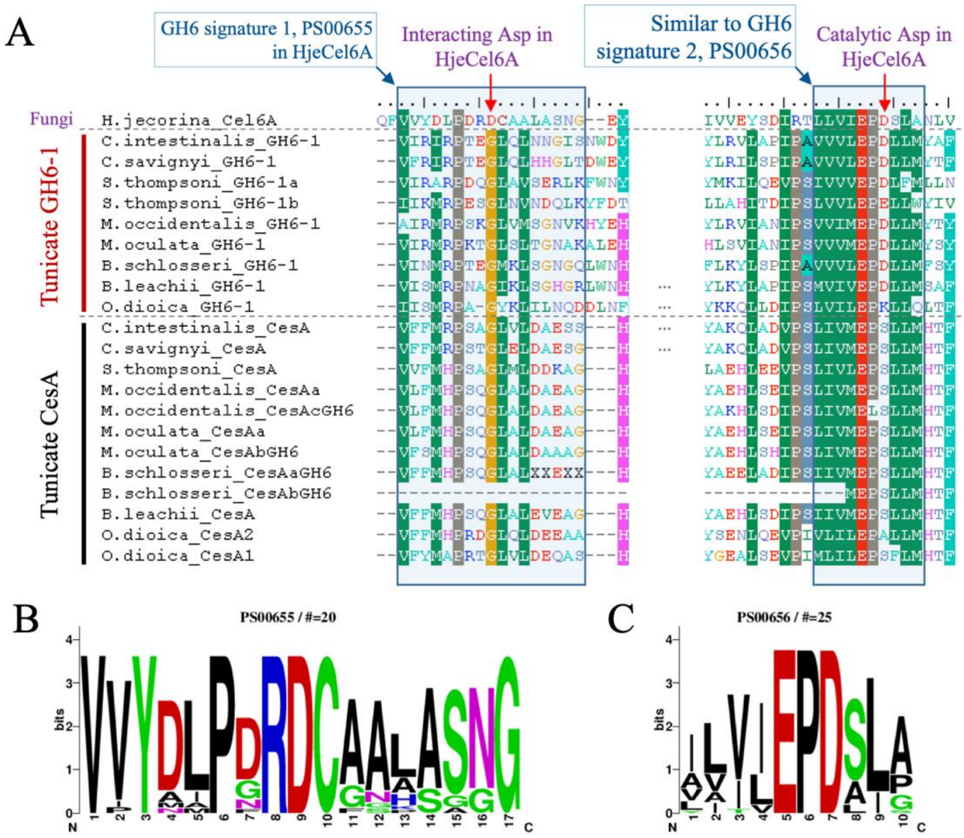

Figure 1.4. Amino acid conservation of tunicate GH6-domain-containing proteins

(**A**) GH6-1 proteins from ascidians and the GH6-1a from *Salpa thompsoni* have aspartic acids that correspond to the catalytic center of fungal Cel6A protein; however, another *S. thompsoni* GH6-1 protein (SthGH6-1b), an *Oikopleura* GH6-1 protein, and tunicate CesA proteins show other amino acids at this site. Similar amino acids under the BLOSUM62 matrix are color-shaded. HjeCel6A: *H. jecorina* Exoglucanase 2, UniProtKB P07987.1. (**B–C**) Sequence logos of Glycosyl Hydrolases Family 6 Signature 1 (PROSITE entry PS00655, panel **B**) and Glycosyl Hydrolases Family 6 Signature 2 (PROSITE entry PS00656, panel **C**), showing the amino acid frequency of each site.

1.3.4 Splice site conservation in tunicate *GH6-1* or *CesA* Genes

I arranged the positions of coding exon splice sites (exon boundaries) of tunicate GH6-1 (Table 1.5) and the GH6 domain of *CesA* genes (Table 1.6) and then registered all sites to an aligned amino acid sequence matrix for comparison. For example, the splice site V217.frame+2 of CinGH6-1 means that the last nucleotide of an exon locates at the second codon position for amino acid 217 (valine) of *C. intestinalis* type A GH6-1 protein. Similarly, the site K316.frame+3 means that the last nucleotide of an exon is the nucleotide of the third codon position for amino acid 316 (lysine).

Table 1.5. Splice site matches of tunicate GH6-1 proteins.

		Splice site name		
		Cin217	**Cin256**	**Cin316**
Protein	**Introns within coding region**	**Splice site residue & frame**		
CinGH6-1	3	V217, +2	G256, +1	K316, +3
CsaGH6-1	3	V223, +2	G262, +1	K322, +3
SthGH6-1a	6	E229, +2	G268, +1	P328, +3
SthGH6-1b	5	K230, +2	G269, +1	K329, +3
MoxGH6-1	3	R222, +2	G260, +1	A320, +3
MocGH6-1	2	n.s.*1 (R222)	G260, +1	A320, +3
BscGH6-1	5	K335, +2	G373, +1	A433, +3
BleGH6-1	4	K229, +2	n.s.*2 (G285)	A345, +3
OdiGH6-1	6	n.s.*1 (N244)	n.s.*2 (G282)	n.s.*1 (K343)
OdiCesA1*3	8	n.s.*1 (R1001)	n.s.*2 (G1040)	R1100*3, frame +2

All matching splice sites found in this study are C-terminal to the possible catalytic center: positions 178–187 in *C. intestinalis* type A GH6-1. *1: No splice (n.s.) site at the aligned amino acid and the amino acid is not conserved; *2: No splice (n.s.) site at the aligned amino acid, although this position encodes a conserved glycine; *3: The splice site OdiCesA1-R1100 could be aligned with splice site Cin316 of GH6-1 proteins at the amino acid level, but there is a one-nucleotide position difference and it may not represent a shared splice site.

Table 1.6. Shared splice sites in the GH6 domain of tunicate CesA proteins.

	Splice site name			
	Cin853	**Cin930**	**Cin976 *1**	**Cin1070**
Protein	**Splice site residue & frame**			
CinCesA	G853, +1	A930, +1	Q976, +3	R1070, +2
CsaCesA	G1137, +1	E1214, +1	Q1260, +3	R1354, +2
SthCesA	G870, +1	E947, +1	Q993, +3	M1087, +2
MoxCesAa	G938, +1	n.s.*2 (G1015)	gap *3	gap
MoxCesAcGH6	G38, +1	n.s. (G115)	L161, +3	R255, +2
MocCesAa	n.s. (G1505)	n.s. (G1582)	n.s. (Q1628)	R1722, +2
MocCesAbGH6	n.s. (G33)	n.s. (G110)	n.s. (Q156)	R250, +2
BscCesAaGH6	n.s. (G2)	G79, +1	Q125, +3	R219, +2
BscCesAbGH6	gap	gap	Q22, +3	R116, +2
BleCesA	G946, +1	G1023, +1	Q1069, +3	R1163, +2
OdiCesA2	n.s. (G848)	n.s. (E925)	Q971, +3	n.s. (R1065)
OdiCesA1	n.s. (G866)	n.s. (P943)	n.s. (Q989)	n.s. (R1083)

*1: This shared splice site has been reported in previous studies (Sagane et al., 2010; Nakashima et al., 2011) and used to infer evolutionary relationships of tunicate *CesA* genes. *2 n.s.: No splice site at this position/codon. *3 gap: A gap in this sequence in multiple alignment, possibly representing deletion of a segment or that the gene model is not complete.

Several splice sites matched among tunicate GH6-1 proteins (Table 1.5), and these matching splice sites also have the same frame as the exon-intron boundary. Therefore, I interpret them as genuine shared splice sites. For example, the site Cin316 was shared by eight *GH6-1* genes from seven tunicate species.

An obscure case was that of the *O. dioica CesA1* R1100 site. Although one *O. dioica* exon boundary was located in a codon for an arginine that could be aligned to the amino acids of splice site Cin316 of GH6-1 proteins, the location of the splice site was shifted by one nucleotide. The current results do not indicate that *O. dioica CesA1* shares this splice site with tunicate *GH6-1* genes.

Excluding the foregoing case, I found no splice site shared between tunicate *GH6-1* and *CesA*. However, several other splice sites are shared within CesA protein GH6 domains (Table 1.6).

1.4 Discussion

1.4.1 Two GH6-containing genes exist in tunicate genomes

In this chapter, I first tried to resolve the relationship of a recently discovered tunicate GH6-containing gene (*GH6-1*), the GH6 part of the tunicate *CesA* gene (called *CesA-GH6*), and GH6-containing genes from other organisms. The result was that tunicate CesA-GH6 and GH6-1 sequences form two clusters (Figure 1.2 & Figure 1.3), indicating that these are two monophyletic groups and that both were inherited from the tunicate common ancestor. On the other hand, in phylogenetic reconstructions, the grouping of tunicate GH6-containing proteins and other GH6s was not conclusive (Figure 1.2 & Figure 1.3). There were long branches that thwarted conclusive results regarding the relationship of tunicate GH6-containing proteins and those of other organisms. I suspect that the highly evolved tunicate GH6-containing-proteins cause long-branch attraction, adversely affecting tree topologies. Based on current phylogenetic trees, I could not confidently propose a non-tunicate GH6 protein(s) that represents the closest relative(s) to tunicate GH6-containing proteins. Considering branch lengths and the tree topology of GH6 proteins, it is possible that an ancient GH6 gene evolved highly, soon after it was transferred into an ancestral tunicate. After the transfer event, this GH6 gene likely duplicated in the tunicate genome. I drew this conclusion because of clustering of tunicate CesA-GH6 and GH6-1 groups, in which no genes of other organisms were inserted. Therefore, either scenario 2 or 3 in Figure 1.1 could explain the origin of tunicate GH6-containing genes. However, as I could not propose a candidate donor species/lineage of tunicate GH6s, I cannot directly evaluate the two possible scenarios further.

It could also be observed from the phylogenetic trees (Figure 1.2 & Figure 1.3) that the branch lengths of terminal branches are generally longer in tunicate GH6-1 sequences than those of CesA-GH6 parts. An interpretation is that CesA-GH6 domain may be under stabilizing selection. Although it may not have hydrolase activity, but it may have cellulose binding affinity and may be utilized to bind newly synthesized cellulose molecules. This conjecture would need affinity assays [for example, (Arola and Linder, 2016)] to validate. Another interpretation would be that different GH6-1 proteins in different tunicate species are under flexible evolutionary constraint or are positively selected to be divergent from the original form.

Assuming that tunicate GH6-genes were acquired via HGT event(s), no other tunicate genes would help to resolve the current, ambiguous tree topology. On the other hand, it is intriguing that many, but not all, GH6 proteins from other eukaryotes (including GH6s of fungi, the SAR supergroup, Haptista, and red algae) were clustered close to two tunicate

GH6 subclusters. Recently, it was shown that some fungi retain many genes acquired from bacteria (Murphy et al., 2019). Therefore, future disclosures of eukaryotic genes similar to tunicate GH6 genes may provide important information on possible horizontal gene transfer events. As I found no GH6 genes in Archaea (Table 1.3), GH6 genes may have been transferred from bacteria to multiple eukaryotes in parallel. Alternatively, GH6 genes could also have been transferred between different eukaryotic organisms.

The separate genomic locations of tunicate *CesA* genes and *GH6-1* genes (Table 1.2) imply that the two genes did not stem from recent tandem duplication events, so these genes have been regulated in different genomic contexts.

1.4.2 Lineage-specific gene content change and partial sequence signature conservation

I found multiple transcripts or gene models representing *GH6-1* (or *CesA-GH6*) in the genomes of some tunicate species (Table 1.1 and Table 1.2). Some of them may represent true lineage-specific duplications, as in the case of the two *CesA* genes of *O. dioica* (Sagane et al., 2010). For example, the two *S. thompsoni* GH6-1 proteins have only 35% identical amino acids when aligned and compared. They also showed long terminal branches in phylogenetic trees. In addition, although the current *S. thompsoni* genome had been assembled into sub-chromosome level scaffolds, these two *GH6-1* genes corresponded to different genomic scaffolds. However, some gene models and open reading frames are highly similar to (around 90% amino acid identity) and shorter than another gene model in the same genome. For example, one GH6-1 protein model of *B. schlosseri* (BscGH6-1b, g61144, chromosome unassigned) showed 97.5% identity to BscGH6-1 (g9326, chromosome 9). These could be more recently duplicated genes. Alternatively, these may just be different alleles annotated separately due to imperfections of software-based genome assembly and may not represent a true species-specific duplication. Some gene models contain the GH6 part, but not the CesA/GT2 part, of the tunicate *CesA* gene. Based on our knowledge that a typical, complete tunicate *CesA* gene contains a CesA/GT2 part and a GH6 part, it is possible that the CesA/GT2 part of a complete tunicate *CesA* gene was erroneously predicted as another gene model in the aforementioned cases, similar to a previous observation on a sea urchin genome (Tu et al., 2012) and several amphioxus gene models (Li et al., 2014). I also found that one *CesA* model of *M. oculata* (*MocCesAa*, Moocul.CG.ELv1_2. S71617.g04842.01.t) is obviously larger. It also encodes a rhodopsin-like G protein-coupled receptor domain (InterPro: IPR000276) at its upstream end. It would require further studies to confirm whether it is a true merged gene, a mistake in genome assembly and annotation, or a polycistronic operon. Polycistronic operons exist in *O. dioica* and *C. intestinalis* type A (Ganot et al., 2004; Satou et al., 2006), but the *CesA* genes of these species were not specifically verified to be a part of polycistronic transcripts.

This analysis of GH6-1 and CesA-GH6 sequence signatures shows that, although both tunicate GH6-1 proteins and CesA-GH6 domains contain a region that almost matches the conserved GH6-signature 2 (PROSITE PS00656), the probable catalytic aspartic acid exists only in GH6-1 proteins and not in CesA-GH6. This aspartic acid is conserved in most non-tunicate GH6 proteins (56 out of 58 sequences compared in this study). Mutation of this possible catalytic site in tunicate *CesA* probably occurred very early in an ancestral tunicate before the branching of the larvacean (Appendicularia) clade. Despite the loss of the aspartic acid of CesA-GH6, the conservation of other amino acids at the signature site may imply that this domain acquired novel function in tunicates. Nevertheless, whether tunicate GH6-1 proteins or CesA-GH6 domains possess any catalytic activity remains to be determined.

1.4.3 Shared splice sites indicate an ancient history of tunicate GH6 genes

In this study, I found several shared splice sites among tunicate *GH6-1* genes. I also extended the comparison of shared splice sites of *CesA* genes to other tunicate species. As previously reported (Sagane et al., 2010), 17 splice sites in *CesA* genes of *C. intestinalis* type A and *C. savignyi* are still conserved after about 100 million years of independent evolution (Delsuc et al., 2018). In addition, a splice site shared by *CesA2* of *O. dioica* and *O. longicauda*, *CesA* of *Halocynthia roretzi*, *Molgula tectiformis*, and two *Ciona* species was interpreted as support for common ancestry of all tunicate *CesA* genes (Nakashima et al., 2011). In this study, although I found no other sites shared between genes of *O. dioica* and other tunicates, I found that many shared splice sites are present among GH6-containing genes from three other major clades of tunicates (Thaliacea + Phlebobranchia + Stolidobranchia). It is reasonable to assume that many shared introns were acquired after the branching of larvaceans and before the subsequent divergence of major tunicate clades. The finding of shared introns in tunicate *GH6-1* genes is similar to the observation of another animal glycosyl hydrolase gene family, Glycosyl Hydrolase Family 9, in which orthologous genes share introns across animal phyla (Lo et al., 2003; Davison and Blaxter, 2005).

There was no well-supported splice site shared between *GH6-1* and *CesA-GH6*. Assuming that only one *GH6* gene was transferred horizontally into an ancestral tunicate genome, the lack of shared splice sites between *GH6-1* and *CesA-GH6* may indicate that the ancient *GH6* gene had no introns when it was transferred into the tunicate genome. This supports a previous interpretation about the tunicate *CesA* transfer event (Bhattachan and

Dong, 2017).

The obscure *O. dioica CesA1* splice site (R1100) differs by just one nucleotide from the Cin316 splice site of *GH6-1* genes. It may simply have resulted from an independent intron acquisition event. Alternatively, this could represent a shared splice site that experienced a one-nucleotide intron shift (Fekete et al., 2017), but this requires further investigation. Moreover, no other *CesA* genes examined show a splice site here. If the GH6 part of the ancient *CesA* gene contained that intron, other *CesA* genes must have undergone intron loss. Therefore, it is not a parsimonious explanation.

The presence of two *CesA* genes in *O. dioica* raised another question of whether tunicate *CesA* was duplicated before larvaceans diverged (Sagane et al., 2010) (see also Figure 1.2 and Figure 1.3). The observation that *Ciona CesA* genes share a splice site with *OdiCesA2*, but not *OdiCesA1*, may favor the scenario of early duplication (Sagane et al., 2010). In my analysis, the splice site discussed previously, Cin976, was found in the *OdiCesA2* and *CesA* genes of at least six other tunicate species (Table 1.6), but this splice site was not found in *M. oculata*. Therefore, it is possible that *O. dioica* had a lineage-specific duplication of the *CesA* gene and that one copy (*CesA1*) lost this intron. In addition, the phylogenetic trees shown in Figure 1.2 and Figure 1.3 also support the interpretation that larvaceans have their own lineage-specific *CesA* duplication.

1.4.4 Perspectives

It is likely that a GH6-containing gene was transferred to and duplicated in ancient tunicate genomes before major tunicate lineages diverged. The two tunicate GH6-containing genes acquired different introns and have preserved part of that sequence signature. Future larvacean transcriptomic studies that are complementary to recent larvacean genome projects [for example, Naville et al. (2019)] may provide a better understanding of tunicate GH6-containing genes and tunicate genome evolution.

In plants, activity of cellulase is required to regulate cellulose synthesis and growth of cell walls (Vain et al., 2014). The prediction of a signal peptide at the N-terminus of CinGH6-1 protein also raised the question of whether this protein is secreted to the extracellular space, where ascidians normally accumulate cellulose and other tunic-forming macromolecules. Therefore, it is important to examine whether tunicate GH6-containing proteins have any hydrolase activity and whether *GH6-1* genes would influence cellulose synthesis and physiology. To examine enzymatic activity of tunicate GH6-1, it is possible to use a proxy microorganism similar to a previous study (Matthysse et al., 2004) Alternatively, I may first synthesize GH6-1 protein with *in vitro* cell-free protein synthesis reagent kits or may prepare recombinant GH6-1 protein using bacterial or eukaryotic cells and purification (Spriestersbach et al., 2015). After obtaining purified GH6-1 proteins, enzymatic assays similar to described assays (Sharma et al., 2018) could clarify the enzymatic activity of tunicate GH6-1. Another approach to understand the biological function of tunicate GH6-1 would be to genetically manipulate animals using genome-editing methods (Sasaki et al., 2014; Treen et al., 2014). These questions were considered in my continued studies described in the next chapter.

Chapter 2 Expression and possible functions of *GH6-1* gene in early *Ciona* development

2.1 Introduction

As revealed in the previous study by Inoue et al. (2019) and my study described in the last chapter and in Li et al. (2020), tunicate genomes contain an orthologous group of *GH6-1* genes. The *GH6-1* genes show sequence similarity to the GH6 part of tunicate *CesA* gene, these two gene groups are inherited from the common ancestor of extant tunicates and probably originated from horizontal gene transfer (HGT) and subsequent duplication. Although the sequences of these genes were revealed in genome projects and bioinformatic studies, the biological significance of these genes remained unknown.

For a horizontally transferred gene to be kept in a host genome, it usually provides extra advantage to either the gene itself or the host (Soucy et al., 2015). It was also proposed that a transferred gene has to cost minimal harm to the host at the early stage of 'domestication' before the host could recruit it to produce adaptive benefit later (Soucy et al., 2015). According to the 'complexity hypothesis', the host, having a different genomic content and cellular components than the donor, usually lack the molecular partners that would interact with the product of transferred gene in their original cellular context (Jain et al., 1999). In many known HGT cases in bacteria, the products of transferred genes are located at periphery of metabolic network (Pál et al., 2005) and may provide the host with additional abilities to utilize environmental resources in new or changing environments (Boto, 2014).

Among the genes that are considered to be originated from horizontal gene transfer, tunicate cellulose synthase (CesA) may be one of the examples that a transferred gene being recruited to affect more complex characters. As previously mentioned in the Introduction chapter, tunicate cellulose synthase provides the cellulose synthesis ability to tunicates. In addition, the study by Sasakura et al. (Sasakura et al., 2005) reported that the disruption of *Ci-CesA* expression in *Ciona* affected not only the tunic properties but also the larva-to-juvenile metamorphosis. Knockdown of one *CesA* copy (*Od-CesA1*) in larvacean tunicate *Oikopleura dioica* inhibited cellulose production at the larval tail and caused failures of notochord cell morphogenesis and of tail elongation (Sagane et al., 2010). The *Od-CesA1*-knockdown embryos also failed to hatch (Sagane et al., 2010). These examples showed that tunicate cellulose synthase gene and/or its product have been incorporated into an essential position of tunicate developmental program.

The epidermal expression of *C. intestinalis CesA* was interpreted as a location consistency of gene expression and cellulose biosynthesis (Nakashima et al., 2004; Sasakura et al., 2005). The study on *O. dioica CesA* genes also showed that different expression time and spaces of two *Od-CesA* genes closely match two genes' physiological functions: *Od-CesA1* is expressed at the lateral sides of the tail in tailbud embryos and is essential for tail development, whereas *Od-CesA2* is expressed at a specialized trunk epithelium (oikoplastic epithelium) and contributes to the secretion of the filter-feeding house of adults (Sagane et al., 2010).

If a transferred gene is not turned into a pseudogene, it needs to be expressed in the new host to provide any effects. However, after a gene was transferred between distant organisms, it usually needs corresponding evolution to fit into the gene expressing system of the new host. In the case of *Ciona intestinalis CesA* (*Ci-CesA*) gene, transcription factor AP-2 binds directly to a non-coding DNA segment upstream to *Ci-CesA* and is necessary for the epidermal expression of *Ci-CesA* (Sasakura et al., 2016). It was proposed that high guanine-cytosine actinobacterial genomic content could easily be transformed into AP-2 binding sites and contributed to the domestication of tunicate *CesA* precursor (Sasakura et al., 2016).

These studies inspired me to investigate the following aspects of the recently found tunicate *GH6-1* gene: (1) gene expression, (2) expression control mechanisms, and (3) physiological function. To investigate these aspects, I used *Ciona intestinalis*, a model ascidian tunicate widely used in evolutionary and developmental biological studies (Lemaire, 2011; Satoh, 2014). Many established methods are available for investigating gene expression and function in *C. intestinalis* (Christiaen et al., 2009; Satoh, 2014; Sasakura, 2018b). The genomic resources and previous expressed sequence tag (EST) studies of *C. intestinalis* also provided evidence of *GH6-1* (corresponding gene model KY.Chr3.452 in the HT version genome) expression in tailbud embryo, larva, juvenile, and adult: raw EST count were recorded and can be viewed on the Ghost database (Satou et al., 2003; Satou et al., 2019).

To understand how the *GH6-1* gene is utilized in *C. intestinalis* type A, firstly I examined expression profile of *GH6-1* in early developmental stages of. Secondly, I performed a reporter assay to check the existence of *GH6-1*-related enhancer in genomic DNA. Finally, I used Transcription Activator-Like Effector Nuclease (TALEN) technique to specifically knock out *GH6-1* and observe the phenotypes of embryos and larvae.

2.2 Methods

2.2.1 Animal acquisition, fertilization, and embryo culture

I obtained *Ciona intestinalis* type A adults, with the help from the National BioResource

Project (NBRP) under Japan Agency for Medical Research and Development. After arrival, these animals were kept in dark and fed with diatom *Chaetoceros calcitrans* in small seawater-containing bucket overnight. For each 20 animals, about 30–100 milliliter of diatom was given. The concentrated diatom product 'sun culture', was provided by Nisshin Marinetech (Yokohama, Kanagawa, Japan). Then the animals were transferred to seawater aquarium, which was maintained at 18 °C. This seawater aquarium was maintained in continuous lighting to stimulate gamete production (Joly et al., 2007) and prevent autonomous gamete spawning (Lambert and Brandt, 1967).

For embryo culture use, seawater was filtered through 0.22 micrometer pore polyethersulfone vacuum filter (Sartorius). Eggs and sperm were obtained surgically (Zeller, 2018). After insemination, eggs were dechorionated with solution [filtered seawater containing 0.1 % w/v Actinase E (protease E) and 1% sodium thioglycollate, adjusted to pH 10 before applying], washed with filtered seawater, and kept in filtered seawater containing 50 mg/L of streptomycin (streptomycin-seawater) in agarose-coated petri dishes and cultured at 18-20 °C.

2.2.2 Reverse transcription quantitative PCR

To understand the *CesA* and *GH6-1* expression in early developmental stages of *C. intestinalis* type A, I extracted total RNA from eggs, embryos, larvae, and young juveniles. Fifty to 100 individuals were first collected in microcentrifuge tubes with minimum remaining seawater (less than 100 µl), then depending on the sample volume, the samples were mixed with 600-1000 µl of TRIzol reagent (Invitrogen-Thermo Fisher Scientific). They were homogenized by 15-second vortex (for eggs and embryos) or manual grinding (hatched larvae and metamorphosing juveniles) with BioMasher II homogenizer (Nippi Inc., Tokyo, Japan). Later, the total RNA was extracted following manufacturer's protocol of TRIzol. The extracted total RNA was reverse-transcribed to complementary DNA (cDNA) with PrimeScript RT reagent Kit Perfect Real Time (Takara Bio, Kusatsu, Shiga, Japan).

Dual-labeled FAM-TAMRA probe and primer pairs for quantitative PCR (qPCR) were designed and synthesized by TaKaRa Bio based on these reference sequences: NCBI reference sequence XM_002131188 for *GAPDH*, NM_001047983 for *CesA*, XM_002119543 for *GH6-1*. For qPCR reaction, complementary DNA, probe-primer sets, and Premix Ex Taq polymerase mix (TaKaRa Bio) were prepared following the manufacturer's protocol and processed on a StepOnePlus thermal cycler (Applied Biosystems, Thermo Fisher Scientific). The expression level of *GAPDH* was used as the normalization standard.

2.2.3 Examining gene expression: *in situ* hybridization and microscopy

Fixation and *in situ* hybridization of *Ciona* samples follows published protocols (Satou et al., 1995) with minor modifications described below.

Embryos, larvae, and young juveniles of *C. intestinalis* type A were fixed in 4% paraformaldehyde dissolved in MOPS buffer (0.1M 3-Morpholinopropane-1-sulfonic acid, 0.5 M NaCl, pH 7.5) for either overnight (12-16 hours) at 4 °C or 1 hour at room temperature. After fixation, the samples were washed with phosphate-buffered saline then with 75% ethanol. They were stored in 75% ethanol at -30 °C until use. The embryo staging follows the tunicate anatomical and developmental ontology website TUNICANATO and previous publications (Hotta et al., 2007; Hotta et al., 2020).

Antisense ribonucleic acid probe (riboprobe) synthesis: total RNA mixture of neurula, tailbud, and larvae were extracted with TRIzol (Invitrogen-Thermo Fisher Scientific) and reverse-transcribed to complementary DNA using SuperScript® III First-Strand Synthesis SuperMix (Thermo Fisher Scientific). Then, the cDNA mixtures were used to specifically amplify a few cDNA fragments (size 800-1100 bp) of the hybridization target genes via PCR. The amplicons were cloned into either pCR^{TM} Blunt II TOPO® vector (Invitrogen-Thermo Fisher Scientific) or pGEM®T-Easy vector (Promega). The amplified cDNA were used as templates for digoxigenin-labeled riboprobe synthesis using DIG RNA labeling mix (Roche Diagnostics) and T7 or SP6 RNA polymerase (Roche Diagnostics). Sense strand riboprobes were also synthesized for control experiments.

In situ hybridization: embryos were rehydrated with PBS, treated with proteinase K, post-fixed, treated with triethanolamine and acetic anhydride following to the protocol (Satou et al., 1995) before 1 hour of pre-hybridization. Hybridization of riboprobes at 42 °C was extended to 18 to 30 hours. After anti-digoxigenin-antibody incubation, samples were washed for 10 times in PBST (0.1% Tween-20 in phosphate-buffered saline). The BM purple solution (Roche Diagnostics) was diluted [50% v/v, in alkaline-phosphatase buffer (Satou et al., 1995), pH 9.5] before applied as color development substrate. After color development, samples were washed in EDTA-PBS (10 mM EDTA in phosphate buffered saline) to stop enzyme reaction. Then, they were washed in 30%, 50%, and 75% ethanol, rehydrated with PBS, and mount in 70% Glycerol/30% PBS before imaging.

Images of the samples were recorded by an AxioCam HRc camera mounted on Axio Imager Z1 microscope (Zeiss). To ease identifying the signal, brightness of some images were adjusted, without creating image artefact, with Fiji/ImageJ software (Schindelin et al., 2012). For some embryos in which signals were distributed at different depths, the Extended

Depth of Field plugin (Forster et al., 2004) of Fiji was applied to combine images of different depths.

2.2.4 Single-cell transcriptome data analysis

A publicly available single-cell transcriptome dataset (accession number GSE131155 on Gene Expression Omnibus), which was originally reported in the work by Cao et al. (2019), was utilized to further confirm the expression of *GH6-1* and *CesA* genes in developing *Ciona* embryos. The initial data processing steps were done with help by Kanako Hisata, Marine Genomic Unit, OIST: the raw read FASTQ files were processed with the CellRanger software (10X Genomics) to generate cell identifier barcodes and gene reads matrix based on a reference genome (Satou et al., 2019).

Then, I used the Seurat toolkit version 3.2.1 (Stuart et al., 2019) in RStudio version 1.2.5019 (Rstudio Team, 2020) to analyze gene expression of a late tailbud stage I sample (corresponding to Gene Expression Omnibus accession: GSM3764780). Low-quality droplets/cells were excluded based on the published method (Cao et al., 2019): (1) cells with less than 1000 expressed genes were discarded, (2) only the genes that was expressed in at lease 3 cells were retained, (3) cells with unique molecular identifiers (UMIs) of five standard deviation above the mean were excluded. Cells passing this filtering (5034 cells) were kept. The expression of each cell was log-normalized. Genes with the top 1,000 highest standard deviations were denoted as highly variable genes and used in principal component analysis. I selected to use dimensions 1 to 20 of the principal component analysis result for running a graph-based clustering approach in the Seurat method and the clustering partitioned the cells into 30 clusters of cells. Gene expression was examined by various plotting methods in the Seurat package: violin plot, FeatureScatter, and FeaturePlot.

2.2.5 Reporter assay and electroporation

To investigate the existence of enhancers that may drive the endogenous expression of *GH6-1* in *Ciona*, I extracted parts of the genomic information of *Ciona intestinalis* type A, *C. savignyi*, and *Molgula occidentalis* from the databases mentioned in section 1.2.1. I used the VISTA comparative genomics computational tools (Frazer et al., 2004) to visualize genomic segments that share similar sequences. Conservation cutoffs (threshold) was set as: minimal 65% identity over each 50-nucleotides window. Next, I designed primers based on *Ciona intestinalis* genomic information to amplify a 2.8 thousand base pairs (kbp) genomic DNA segment upstream to the predicted *GH6-1* gene model KH.L63.12 (Table 2.1). The amplified DNA segment was cloned into a customized Kaede expression vector (a generous gift from Dr. Keisuke Nakashima, Marine Genomics Unit, OIST) with In-Fusion HD Cloning Kit (TaKaRa Bio) to construct a reporter plasmid.

Table 2.1. Primers used for *Ciona GH6-1* upstream genomic DNA cloning

Lower case letters represent homologous adaptor sequences used for In-Fusion recombination cloning.

GH6 upstream forward	5′-TTTCTAACTTTGTAAAATTTAAAATTGA
In-Fusion forward	5′-tgcctgcaggtcgacTTTCTAACTTTGTAAAATTTAAAATTGA
GH6 upstream reverse	5′- TTTGGTTCCTTGATCGAATTTTT
In-Fusion ATG-reverse	5′-aatcagactcaccatTTTGGTTCCTTGATCGAATTTTT

To introduce the customized reporter plasmids, electroporation experiments were performed based on a published protocol (Corbo et al., 1997) with minor modifications. Dechorionation steps were similar to the aforementioned method (section 2.2.1) but was applied to unfertilized eggs before the insemination step. After washing, eggs were fertilized, and then the dechorionated-fertilized *Ciona* eggs with minimal filtered seawater were added into mannitol solution (0.77 M mannitol in 10% v/v filtered seawater). Then, for each electroporation group, 15-60 µg of plasmid DNA (dissolved in 80 µl of TE buffer) were combined to 720 µl of abovementioned egg-mannitol mix, making final 800 µl volume for each electroporation group. Electroporation (50 volts, 20 milliseconds, in 4-millimeter cuvette) was performed by a GenePulser Xcell pulser (Bio-Rad). After pulsing, the eggs were carefully transferred into streptomycin-seawater in agarose-coated or gelatin-coated plastic petri dishes. After 12 hours of incubation at 18 °C, embryos were fixed with the same method described in section 2.2.3.

2.2.6 Gene knockout experiments by TALEN-mediated genome editing

To understand the physiological function of *GH6-1* in *C. intestinalis* type A, I assembled two sets of transcription activator-like effector nuclease (TALEN) pairs following the method described by Sakuma et al. (2013) and Treen et al. (2014) and the TALEN assembly protocols released on the NBRP-*Ciona intestinalis* Transgenic line RESources (CITRES) website . The Platinum Gate TALEN Kit was acquired from the Addgene plasmid repository. For expressing the TALEN in *Ciona*, EF1a>TALEN-NG::2A::mCherry vector provided by Dr. Yasunori Sasakura (see also the NBRP-CITRES website) was used for the second step assembly target. Optimal TALEN-binding targets in coding sequence of *Ciona GH6-1* gene were selected with the assistance of TAL Effector Nucleotide Targeter 2.0

(Doyle et al., 2012). The customized TALENs were designed to excise and disrupt the coding part of *GH6-1* gene. A TALEN pair (GH6-TL1) targets CGGC-CTAC-TGAA-GGTC-T and ATTT-CGAA-CTGG-GATT (spanning the 394-442nd nucleotide of assumed coding sequence); a second TALEN pair (GH6-TL2) targets TTCG-AACT-GGGA-TTAT and ATTT-CTAC-CTGG-ACAG (429-478th nucleotide). These targets are both upstream to the probable active site of *Ciona* GH6-1 protein (as described in Chapter 1). Therefore, these TALEN pairs were expected to disrupt the function of *Ciona* GH6-1 protein.

The TALEN-encoding plasmids were introduced to dechorionated-fertilized eggs by electroporation as described in section 2.2.5. For electroporation control, a plasmid containing promotor of *Ciona* forkhead gene (*Ci-fkh*) and monomer Venus fluorescent protein gene (a generous gift from Dr. Koki Nishitsuji, Marine Genomics Unit, OIST) was used. Twenty to 60 micrograms of DNA (10 to 30 micrograms for each plasmid of two units of a TALEN pair) were used in one experiment. After electroporation, the eggs were transferred to streptomycin-seawater and incubated at 18 °C.

A few embryos were used for checking the specific editing efficiency. Genomic DNA of 15-30 embryos (12-16 hours post fertilization) were extracted with Maxwell® RSC Blood DNA Kit on a Maxwell® RSC Instrument (Promega). The extracted genomic DNA were amplified with primers targeting a part of *GH6-1* gene; Forward: 5'-GCCTCGCTACAAGAACCACC and Reverse: 5'- ACACAATGACTTTTCGAGCGC. The amplicon was purified, cloned into pGEM®T-Easy vector (Promega), and sequenced with BigDye™ Terminator v3.1 kit on SeqStudio Genetic Analyzers (Thermo Fisher Scientific).

The electroporated embryos were examined after reaching neurula stage (about 10 hours after fertilization) under a Leica M205-FA microscope. The embryos that actually received and expressed the introduced plasmid showed red fluorescence of the mCherry protein. Only these embryos were used in later steps.

Crystallized cellulose of these *Ciona* larvae was examined by staining and microscopy. The staining method below was modifed from a previous study (Nakashima et al., 2011). Commercialized green fluorescent protein tagged carbohydrate binding module (Carbohydrate Binding Module 3A, GFP-CBM3, origin: *Bacteroides cellulosolvens*) was purchased from NZYtech. The protein was centrifuged and redissolved in an assay buffer (20 mM Tris-HCl, pH 7.5, 20 mM NaCl, 5 mM $CaCl_2$) according to manufacturer's protocol. It was diluted to 1/6 concentration in the assay buffer before application. The *Ciona* embryos or larvae were fixed and preserved as described in section 2.2.3. These samples were rehydrated in three washes (10 minutes each) of phosphate buffered saline containing 0.1% Tween-20 (PBSTw), incubated in blocking solution [PBS with 1% Blocking Reagent

(Roche)] for 1 hour at room temperature, washed two times in Tris-buffered saline containing 0.1% Tween-20, rinsed two times in the assay buffer, stained with the diluted GFP-CBM3 assay solution for 12-20 hours at 4 °C, washed 8 times in Tris-buffered saline + 0.1% Tween-20 buffer, and transferred to VECTASHIELD Antifade Mounting Medium (Vector Laboratories) before microscopic imaging.

2.3 Results

2.3.1 Expression of *GH6-1* and *CesA* at *Ciona* embryonic epidermis

The quantitative expression of *GH6-1* and *CesA* in early developmental stages was examined by RT-qPCR and the results are shown in Figure 2.1. The expression of *GH6-1* remained at low level until mid-tailbud stage (10 hpf at 18 °C). The late tailbud stages and hatching larvae (18 hpf) showed higher levels of *GH6-1* expression. However, the larvae after hatching (24 hpf) and settled juveniles showed lower expression.

In parallel, I also examined the temporal expression profile of the tunicate cellulose synthase gene, *CesA*. In accordance with results of a previous study (Nakashima et al., 2004), expression of *CesA* was detected in the tailbud embryos and larvae. The *CesA* expression level rose at the mid tailbud stage and became highest at late tailbud stage (about 16 hpf at 18 °C). Although the hatching larvae (18 hpf) also show high level of *CesA* expression, the expression was greatly reduced in the 24 hpf larvae and settled juveniles.

Interestingly, both *GH6-1* and *CesA* genes showed higher levels of expression at late tailbud stages and hatching larvae, and then showed decreased level of expression several hours after hatching. This observation shows that both the expression of genes is dynamically regulated, implying their functions in larval physiology.

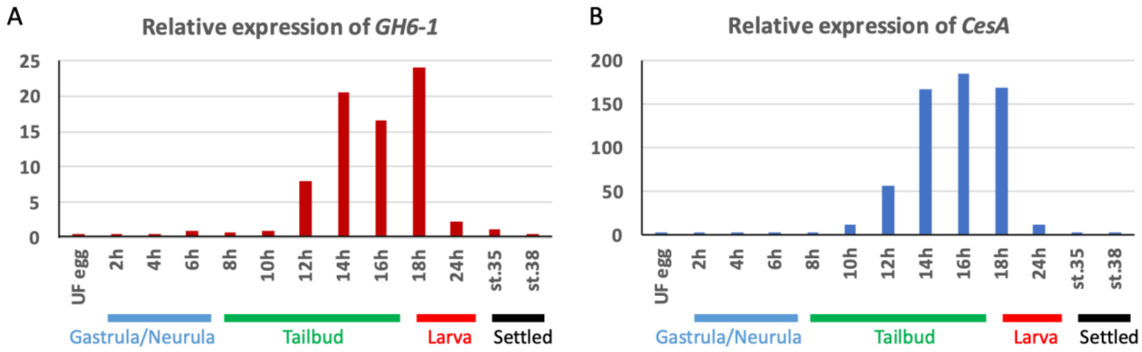

Figure 2.1. Quantitative level of the expression of *Ciona GH6-1* and *CesA* genes during early development of *Ciona intestinalis* type A

X-axis: developmental time and stages. Y-axis: the expression of each gene was first normalized to *GAPDH* and then the expression level of each gene at 6 hours post fertilization was set as 1 for normalization. The developmental staging follows the TUNICANATO website.

The spatial embryonic expression of *GH6-1* was determined by *in situ* hybridization (Figure 2.2). The *GH6-1* expression was not detected in gastrulae (Figure 2.2A), and first appeared at the epidermis of future tail tip of late neurula (Figure 2.2B). The tail tip expression persisted in early and mid tailbud stages (Figure 2.2C, D). Later, at the late tailbud stage I, dorsal and ventral midline epidermis of the tail and many of the trunk epidermal cells also showed expression (Figure 2.2E, Ea-Ec); specifically, three clusters of anterior trunk cells showed stronger signal (open arrows in Figure 2.2E, Ea). These locations may correspond to the future papillae. At the late tailbud stage II, while the anterior trunk and tail tip expression was still strong, the tail midline expression was decreased and hard to detect (Figure 2.2F). Control embryos treated with sense riboprobe showed no clear signal at all stages examined (Figure 2.2G-L).

This *GH6-1* expression was compared with the expression of *CesA* gene, another horizontally-transferred gene (Figure 2.3). Although these two genes are both expressed at the epidermis, the *CesA* expression appears ubiquitous in all of the epidermis (Figure 2.3) while the *GH6-1* expression is more localized.

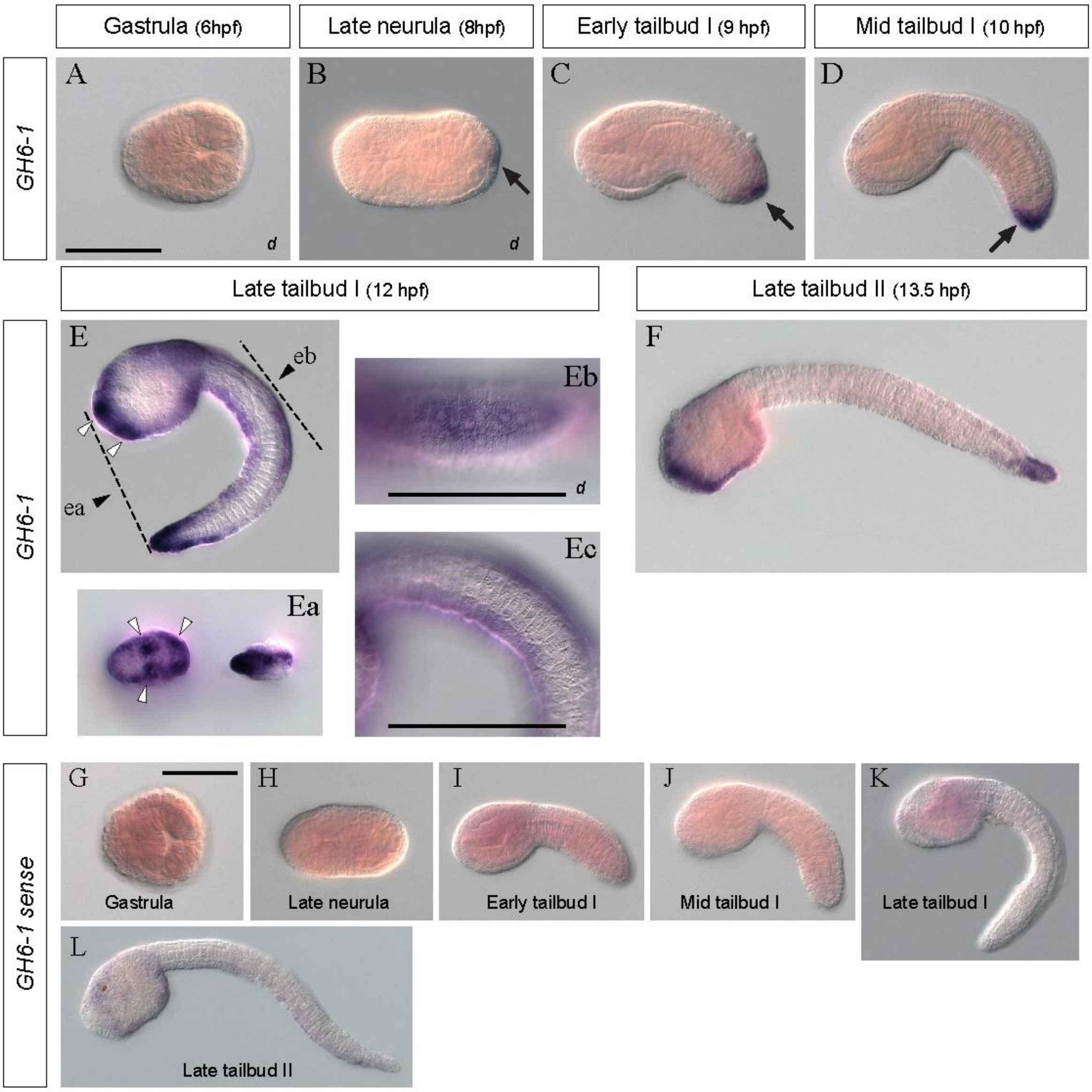

Figure 2.2. *In situ* hybridization of the *GH6-1* gene in *Ciona intestinalis* type A

No clear expression was observed in gastrula stage (**A**). Epidermal expression was first observed at tail-tip (arrows in panels **B**, **C**, and **D**). Expression was expanded in late tailbud stages (**E**, **Ea**-**Ec**, and **F**). In panel E, filled arrows ea and eb represent the viewing aspect of panel Ea and Eb, respectively. Control embryos (**G**-**L**) were treated with sense riboprobe and showed no clear signal. Panels A-E, Ea, and F are the same magnification as panel A, in which a scale bar represents 100 micrometers. Scale bars in Eb and Ec also represent 100 micrometers. Panel G shows a scale bar of 100 micrometers; panels H-L are in the same magnification as panel G. The *d* footnote in A, B, and Eb denotes dorsal-view.

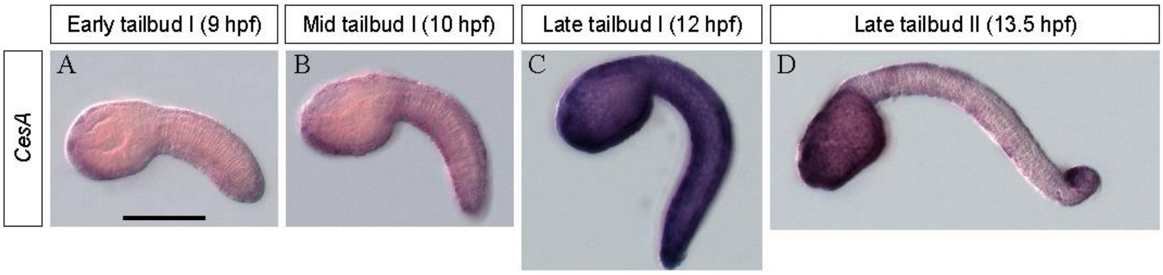

Figure 2.3. *In situ* hybridization of the *CesA* gene in *Ciona intestinalis* type A

The *CesA* appeared to show ubiquitous epidermal expression. The scale bar in panel A representing 100 micrometers applies to all panels.

In addition, analysis of the single-cell transcriptome also showed that the cells that express the *GH6-1* or *CesA* genes may be most likely identified as epidermal cells. In this analysis, data from late tailbud stage I embryos (Cao et al., 2019) was analyzed. Cells with different gene expression profiles were separated to 30 clusters (Figure 2.4A): cells in the same cluster shared similar gene expression patterns (Figure 2.4). The intermediate filament protein *IF-C* gene (KY.Chr3.1290 on the Ghost database HT version; see also LOC100175966 on the NCBI Gene database) (Wang et al., 2002; Cao et al., 2019) was selected as a marker of the epidermal cell identity. The *IF-C* gene was highly expressed at cell clusters number 0, 1, and 5 (Figure 2.4B). The same three clusters also contained many cells that expressed *GH6-1* (Figure 2.4C) and/or *CesA* (Figure 2.4D). This observation agreed with the *in situ* hybridization result as described in the previous text.

The gene expression of each cell was further analyzed by scatter plots, showing the normalized expression level of *IF-C*, *CesA*, and *GH6-1* genes (Figure 2.4E). Although in the cluster-level of grouping, *IF-C* expressing clusters matched with *GH6-1* expressing clusters and *CesA* expressing clusters, the individual cells (transcriptomes) may either show high level of both two genes or high level of one gene and no/low expression of another.

The *GH6-1* and *CesA* expression level were plotted onto the dimension reduction plots (Figure 2.4F). Within these target clusters, while many of the cells show higher level of either gene, only few cells showed high expression of both *GH6-1* and *CesA* (yellow dots, bottom-right inset). These results showed additional information, which are difficult to observe by *in situ* hybridization method alone, on cell heterogeneity within the same tissue type (the same cluster of cells). The *CesA* expression examined by traditional *in situ* hybridization method showed seemingly ubiquitous epidermal expression (Figure 2.3). On the other hand, GH6-1 expression was seen in the anterior-most and posterior-tip epidermal cells (Figure 2.2). Therefore, the *GH6-1* expressing cells are likely to co-express *CesA* in the two tip regions. In this single-cell transcriptome analysis, the quantified expression data showed that although there were cells co-expressing these two genes, usually only one gene was highly expressed. In other words, there is only a moderate correlation of expression strengths of these two genes.

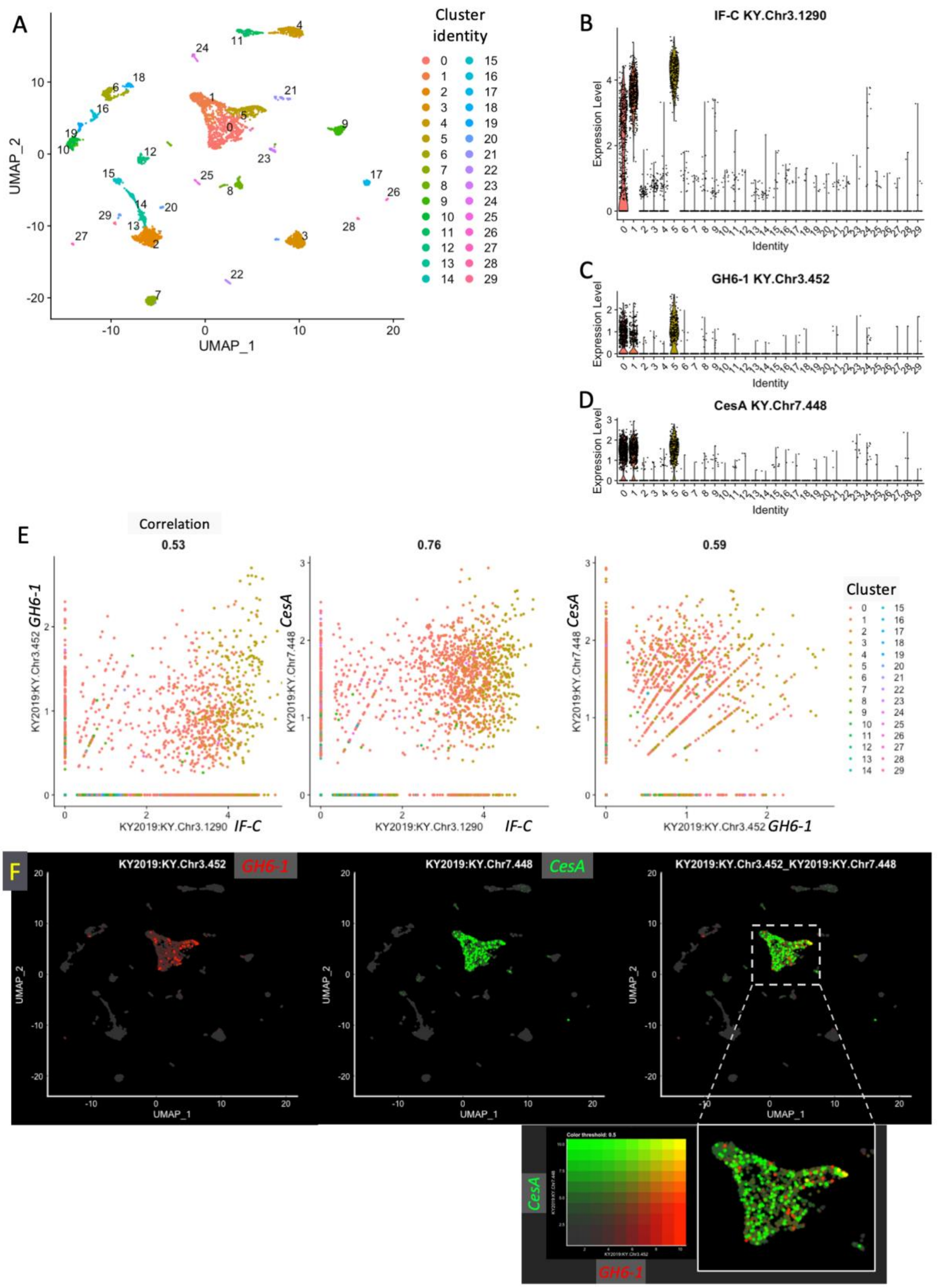

Figure 2.4. Single-cell transcriptome analysis showed *GH6-1* and *CesA* expressions correspond to epidermal cell identity

(The figure legend is shown at the next page)

(**A**) Dimension reduction plot shows representation of the late tailbud stage I embryonic cells, separated to 30 clusters (numbered 0 to 29). Dimensions were reduced by the uniform manifold approximation and projection (UMAP) technique in the Seurat package. Each dot represents the transcriptome of a single cell. Cells in the same cluster have similar gene expression profiles. Number and color labels denote the cluster identity. (**B**-**D**) Violin plots showing the expression level of three genes. X-axis: cell cluster identifier. Y-axis: normalized expression level of each gene. (B) The *IF-C* gene (KY.Chr3.1290) was selected as an epidermal marker of the cell identity of clusters. It was highly expressed at the cells in clusters No. 1 and 5 and to a lesser extent in cluster No. 0. (C, D) The *GH6-1*-expressing cells and *CesA*-expressing cells were mostly in the clusters No. 0, 1, and 5. (**E**) Scatter plots showing relationships of normalized expression of *IF-C*, *CesA*, and *GH6-1*. Each dot represents the transcriptome of a cell and are colored by cluster identity. The X- and Y-axes are normalized expression level. Pearson correlation between the two features is displayed above each plot. (**F**) UMAP-dimension reduction plots showing normalized expression level of *GH6-1* and *CesA* genes. Although expression of these two genes were shown mostly in the same cell clusters (no. 0, 1, and 5), only few cells show high expression of both genes (yellow dots). Enlarged dashed rectangle area is shown as inset at bottom-right.

2.3.2 Reporter assay for the existence of enhancers upstream to *Ciona GH6-1* gene

Non-coding genomic regions around orthologous genes from moderately distant species sometimes show sequence similarity, and this may correspond to the existence of conserved regulatory elements (Sean et al., 2004; Yue et al., 2016). In order to understand the mechanism regulating *GH6-1* expression in tunicates, I compared the genomic regions around *GH6-1* genes of three tunicates species (Figure 2.5). A few non-coding regions showed sequence similarity between *Ciona intestinalis* type A and *C. savignyi*; one small region of similarity was also found between *C. intestinalis* type A and the more distantly related *Mogula occidentalis*. Therefore, I hypothesized that the genomic region upstream to *C. intestinalis* type A *GH6-1* gene contains elements regulating *GH6-1* expression.

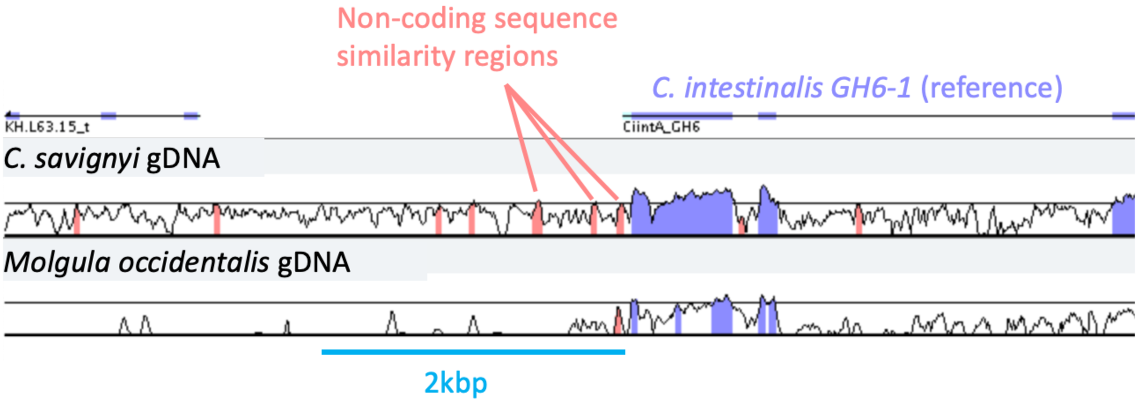

Figure 2.5. Comparison of genomic segments revealed regions of sequence similarity around tunicate *GH6-1* genes.

Genomic segments of three tunicate species were compared with the VISTA tools. The reference (*C. intestinalis* type A) is only represented by the gene-exon line symbols at the top. For *C. savignyi* and *M. occidentalis* genomic DNA (gDNA) segments, the regions of high sequence similarity are denoted as colored peaks: the blue-colored peaks show similarities of coding-regions, and the pink-colored peaks show non-coding sequence similarity regions. For clarity, only the upstream three exons of *Ciona GH6-1* gene (total: 4 exons) were shown. Under the same threshold setting, there is no region of high sequence similarity in the third intron.

In the reporter assay, I combined an about 2.8-kbp long *Ciona GH6-1* upstream genomic segment to a gene *Kaede*, which encodes a fluorescent reporter protein. This upstream genomic DNA segment successfully induced Kaede signal in *Ciona* embryos (Figure 2.6). The location of Kaede expression partially represented the endogenous expression of *GH6-1* gene. Therefore, I interpreted that this segment of non-coding genomic

DNA contains enhancers that can activate the endogenous *GH6-1* expression in normal embryogenesis. However, the location of the expression did not perfectly match the endogenous expression of *GH6-1*: the endogenous strong expression of *GH6-1* at anterior trunk epidermis was not clearly represented in the reporter experiment. Also, in the reporter experiment, the central nervous system, trunk mesenchyme, and tail muscle showed weak Kaede expression (Figure 2.6G). These fluorescence-expressing parts did not correspond to any endogenous expression of *GH6-1* gene.

In the 28 embryos observed, 26 individuals (92.9%) showed fluorescence at the tail-tip. The proportion of each tissue showing the Kaede fluorescence was summarized in Table 2.2.

Table 2.2. Percentage (and number) of embryos in reporter assay showing fluorescence at each tissue/structure

	Tail-tip	Tail-muscle	Mesenchyme (posterior trunk)	Central nervous system	Trunk epidermis
Percentage (%) of embryos show fluorescence at tissue/structure (n=28)	92.9 (26)	60.7 (17)	57.1 (16)	64.3 (18)	28.6 (8)

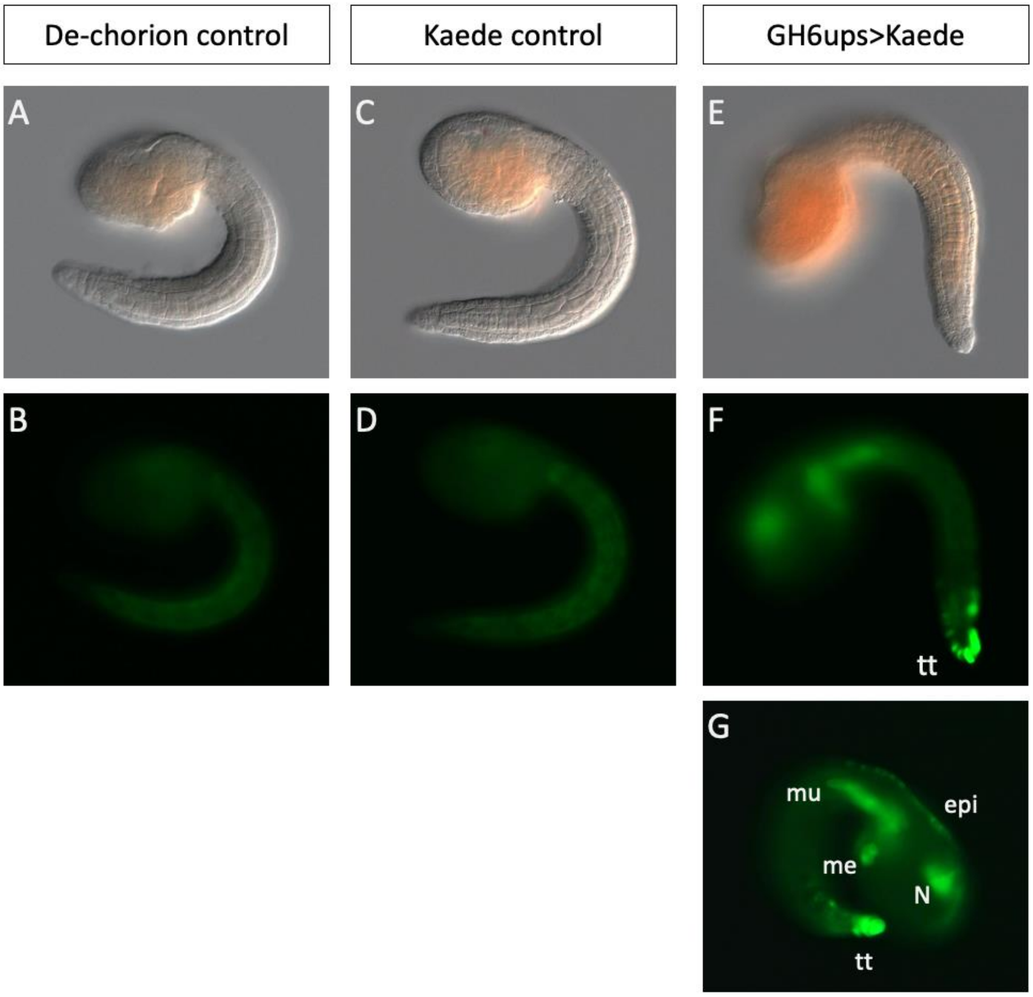

Figure 2.6. Fluorescent reporter gene expression was activated by *Ciona GH6-1* upstream genomic DNA

In the dechorionated control group (**A** and **B**), fertilized egg developed without any artificial DNA. In the 'Kaede control' group (**C** and **D**), eggs were electroporated with an 'empty' vector, which contains only the *Kaede* coding sequence but no enhancer/promotor. In both dechorionated and Kaede-only control groups, only dim background green light was shown. In the reporter group (GH6ups>Kaede) (**E**, **F**, and **G**), a plasmid construct, containing *Ciona GH6-1* upstream genomic sequence and *Kaede* gene, was introduced to the fertilized egg by electroporation. The *Kaede* expression was turned on and therefore a few parts of the tailbud embryo showed bright green light. The first row (A, C, E) and second row (B, D, F) shows the same embryo observed with DIC visible light and green fluorescence filtered light. Tissue/structure labels: epi, epidermis; me, trunk mesenchyme; mu, tail muscle; N, central nervous system; tt, tail-tip.

2.3.3 TALEN-mediated knockout of *GH6-1* affects adhesive papillae formation and metamorphosis

To understand the physiological significance of *GH6-1* gene, I also used TALEN-mediated genome editing to knockout the *GH6-1* gene. Customized TALEN-expressing plasmids were introduced to fertilized *Ciona* eggs by electroporation. In a pioneer experiment, the genome-editing efficiency of two TALEN pairs were evaluated by sequencing. The TALEN pair 1 had an editing efficiency of 57% and the TALEN pair 2 had an efficiency of 63%. Therefore, the TALEN pair 2 was selected for the rest of study. The target site was 429-478th nucleotide of assumed coding sequence of *GH6-1*, and I observed a higher proportion of phenotypically affected individuals (described below).

In electroporation-control experiments, a plasmid containing *Ciona* forkhead gene (*Ci-fkh*) promotor and monomer Venus fluorescent protein gene was used. When embryos develop to hatching larva stage (17.5-18 hours post fertilization at 18 °C), the larvae have three normally protruding adhesive papillae at the anterior trunk, as shown in the control group (Table 2.3 and Figure 2.7A). However, a majority (74%) of the TALEN-affected larvae did not develop the protruding papillae (Figure 2.7C).

As it is questioned whether the GH6-1 protein would participate in cellulose metabolism, I also examined the existence and abundance of cellulose of the larval tunic, which normally is formed shortly before *Ciona* larvae hatch. Larvae were stained with GFP-linked carbohydrate binding modules (Figure 2.7B, D). In the control animals, the larva showed cellulose signal at the larval tunic external to the epidermis. Notably, at the epidermis of anterior trunk, an area surrounding the future papillae, less cellulose signal was observed compared to other parts of trunk epidermis (arrows in Figure 2.7B). TALEN-affected animals also showed cellulose signal at larval tunic (Figure 2.7D). However, I found that anterior epidermis of these affected animals showed a signal strength similar to other parts of trunk epidermis (asterisks in Figure 2.7D). In other words, the anterior epidermis of affected animals showed a relatively stronger cellulose signal than the corresponding part of control animals.

For a normal *Ciona* larva, the adhesive papillae secrete mucus and are important for adhesion to substrate (Pennati and Rothbächer, 2015; Zeng et al., 2019). Although the TALEN-affected animals did not form the normal papillae, about two-thirds of these larvae could still adhere to Petri dish bottom at 2 days post fertilization. However, for all the settled animals at 3 days post fertilization, while a majority (40/44, 91%) of control group animals continued to metamorphose (axis rotation, adult structure formation), 78% (38/49) of the TALEN-treated larvae stopped further development required for metamorphosis after settlement: in these attached larvae, the tail was not resorbed, the trunk region shrank, and no adult organ were seen. Therefore, the functional GH6-1 protein contributes to essential physiology during the metamorphosis stages.

Table 2.3. Phenotypes of TALEN-affected *Ciona* larvae.

	Control	**TALEN pair no.2**
Development of three adhesive papillae	100% (n=20)	26% (n=23)
Production of cellulose	100% (n=7)	100% (n=9)
Reduced cellulose at anterior trunk	100% (n=7)	29% (n=7)

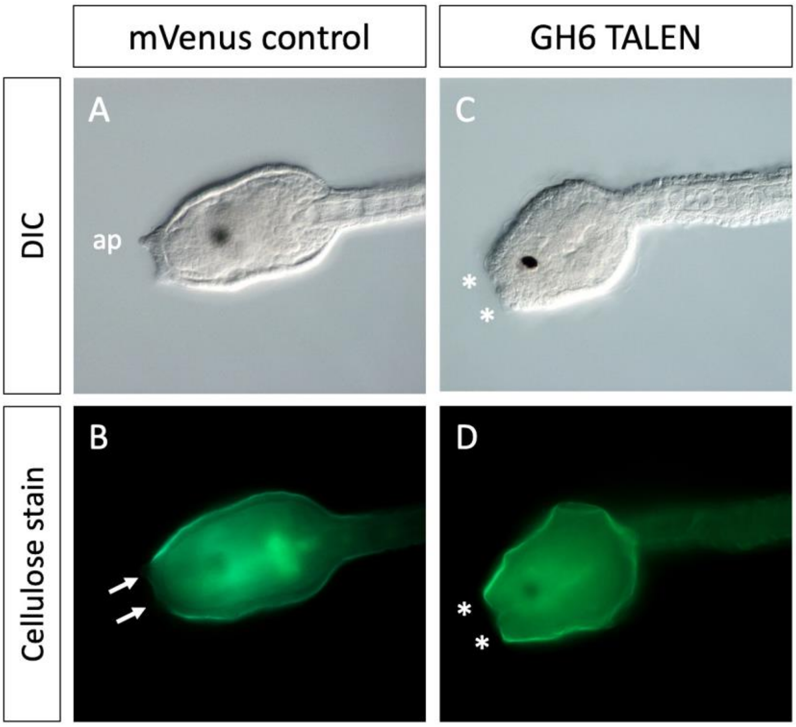

Figure 2.7. Knocking out of *GH6-1* by TALEN affects development of larval structures

(A, B) A hatching larva of electroporation control group, showing protruding adhesive papillae (ap in A) and reduced cellulose signal at the anterior trunk epidermis (arrows in B). **(C, D)** A hatching larva expressing TALENs targeting the *GH6-1* gene. The adhesive papillae fail in development (asterisks in C) and the cellulose signal at the anterior epidermis was slightly stronger than the corresponding part of control larvae.

2.4 Discussion

2.4.1 Dynamic expression of *GH6-1* gene in developing *Ciona* embryos and larvae and the implication to its possible function

In this chapter, firstly I examined and compared the quantitative and spatial expression profiles of *GH6-1* and *CesA* genes in *C. intestinalis* embryos and larvae. The stronger expression of *GH6-1* was observed in late embryonic stages before hatching. As CesA also showed strong expression at late embryonic stages, it is possible that the expression of two genes are temporally correlated and may be regulated by a similar set of upstream regulators.

The epidermal expression of *GH6-1* at late embryonic and early larval stages could be interpreted as that translated GH6-1 protein contributes to epidermal/larval structure formation or larval physiology. Alternatively, *GH6-1* could merely share a similar regulation status to other genes that are controlled by an upstream developmental regulatory network (Davidson, 2010). For example, the pathways patterning epidermal sensory neurons include tail midlines-specific transcription factor (Pasini et al., 2006).

The epidermal expression also implies that GH6-1 has a function different from nutritious digestion. Based the conserved signature analysis in section 1.3.3, tunicate GH6-1 may be able to catalyze cellulose hydrolysis. However, the epidermal expression is clearly different from the expression profiles of many 'digestive cellulases' of other animals, which are expressed at digestive systems: salivary glands (termites), midgut (insects), or hepatopancreas (crayfish and snail) (Watanabe and Tokuda, 2001). Also, the larva of *Ciona* does not have a functional mouth/oral siphon and does not eat food particles (Chiba et al., 2004; Hotta et al., 2020).

A *Ciona* hatching larva has protruding papillae at the anterior of the trunk and extracellular larval tunic covering its surface. The larval tunic at the tail expands outwardly in the dorsal and ventral directions and the tail-tip to form a tail-fin (Sasakura et al., 2005). The cells showing stronger expression of *GH6-1* coincides with the location of these protruding structures.

The papillae include protruding, elongated cells and hyaline caps (Zeng et al., 2019). On the other hand, the larval tunic itself is acellular and external to the epidermis, although in the wild type animals the test cells are attaching to the exterior of the larval tunic (Sato et al., 1997). If the prediction of signal peptide in *Ciona* GH6-1 (section 1.3.3) holds true, and if the *Ciona* GH6-1 protein is able to hydrolyze or interact with cellulose or other polysaccharides, the secreted GH6-1 proteins could regulate or alter the composition of local polysaccharide macromolecules. Since the extracellular matrix can affect the shape of cells, it is possible that the GH6-1 participated in the formation of both kinds of protruding structures. If *Ciona* GH6-1 is able to digest cellulose or other extracellular polysaccharides, the tunic at the site of GH6-1 secretion would be softened and allow the adhesive papillae cells to grow outward.

2.4.2 The *GH6-1* gene obtained expression regulation after horizontal gene transfer

In my reporter assay, the position of reporter fluorescent protein signal partially represented the endogenous *GH6-1* expression in *Ciona*. Possible explanations are listed below. (1) The intergenic, *GH6-1* upstream genomic segment I used to construct the reporter plasmid probably contains not only the regulatory *cis*-elements of the *GH6-1* gene but also *cis*-elements of the neighbor gene. This 2.8 kbp non-coding genomic segment includes regulatory components of two genes: the *GH6-1* gene and the neighboring gene model KY.Chr3.453 (the genome HT version), which is in the opposite direction. *Ciona intestinalis* has a compact genome, and most enhancers are located within 1.5 kbp upstream to the transcription start sites of each gene (Satoh, 2003; Q. Irvine, 2013). Therefore, it is likely that the *cis*-elements of KY.Chr3.453 gene also induced expression of the reporter at the locations that do not correspond to *GH6-1* expression. (2) The *GH6-1* gene may have other enhancers and silencers at other genomic locations. One candidate would be the introns of *GH6-1* gene. In the comparison of genomic sequences, regions of sequence similarity between *C. intestinalis* and *C. savignyi* were also found at the first and the second introns (Figure 2.5). While the first intron is small (170 bases), the second intron is the largest intron of this gene and has a size of about 2.2 kbp. These introns may also contain *cis*-elements that regulate the expression of *GH6-1* gene, and the endogenous expression of *GH6-1* is under the integrative regulation from all relevant *cis*-elements. Therefore, without these elements, my reporter assay represented only part of the endogenous *GH6-1* expression.

Athough I have not clarified the detailed locations and properties of all *cis*-elements relevant to *GH6-1* gene, this gene is expressed at the embryonic epidermis with local regulation. This shows that the *GH6-1* gene, after horizontally transferred to tunicate genomic context, obtained expression regulation in the cellular environment of the new host. This is similar to the incorporation of *CesA* gene into expression regulation, in which an endogenous epidermal transcription factor AP-2 could turn on the horizontally transferred *Ci-CesA* gene at epidermis (Sasakura et al., 2016). However, the details of expression regulation of *GH6-1* is different from that of *CesA*. Both *in situ* hybridization and single-cell transcriptome analysis results showed differences of the two genes: different expression strength at different locations. The current observation would lead to future studies on clarifying what the exact regulation mechanisms are and when these mechanisms were linked to these transferred genes.

2.4.3 *Ciona GH6-1* is likely incorporated into metamorphosis regulation

In this chapter, it has been revealed that knocking out *GH6-1* gene in *Ciona intestinalis* type A affects three biological processes: the adhesive papillae development, cellulose accumulation, and larva-to-juvenile metamorphosis. These phenotypes of larvae suggest significant roles of *GH6-1* gene in early *Ciona* development.

Metamorphosis in ascidians, in which a swimming larva starts to transform into its sessile life, is a complex set of events (Cloney, 1982; Chambon et al., 2002; Sasakura et al., 2005; Nakayama-Ishimura et al., 2009). Among the *Ciona* larval organs, papillae are thought to be a key player in metamorphosis (Nakayama-Ishimura et al., 2009; Wakai et al., 2021). The mucous secretion and adhering to a surface of substrate by papillae are considered as the first events in metamorphosis (Cloney, 1982; Sasakura, 2018a). The *GH6-1* knockout larvae could still attach to the bottom of Petri dishes, implying that the GH6-1 is not necessary in mucus-secretion and adhesion.

Knocking out of *Ciona GH6-1* gene affects the cellulose distribution of the anterior face of trunk. This is somewhat an expected outcome as I had proposed that GH6-1 protein might have catalytic activity, based on the conserved signature analysis in Chapter 1.

More surprisingly, the knockout animals also showed failure of establishing papillae protrusion and failed to continue metamorphic events after settlement. The abnormal papillae may be related to my hypothesis that GH6-1 could digest/affect extracellular components and ease the growth of protruding structures (section 2.4.1). The mechanisms of how GH6-1 or abnormal papillae caused the failure of later metamorphosis is unknown for now, but it is possible that the neurons in abnormal adhesive papillae were also affected. Metamorphosis in *Ciona* includes the collaborative works of various types of cells and cellular processes: adhesion of papillae to substrate and subsequent neuronal activities (Matsunobu and Sasakura, 2015; Wakai et al., 2021), neurotransmitter and neuropeptide signaling (Kimura et al., 2003; Kamiya et al., 2014; Hozumi et al., 2020), and apoptosis (Chambon et al., 2002). Either of the process could be directly (or indirectly) affected when GH6-1 is lost. Although the *GH6-1* knockout larvae still have mucus-secreting or adhering ability at the papillae, it will be important to examine whether the neurons in these adhesive papillae cannot sense the attachment physical stimuli or cannot send neuronal signals.

Different metamorphic events of *Ciona* depend on different pathways (Nakayama-Ishimura et al., 2009; Hozumi et al., 2020). For example, growth of adult organs requires neurotransmitter gamma-aminobutyric acid (GABA) but does not rely on existence of *Ci-*

CesA/cellulose nor gonadotropin-releasing hormone (GnRH) (Sasakura et al., 2005; Hozumi et al., 2020). Clarifying the detailed mechanism of how *GH6-1* affects downstream metamorphic events will be valuable for understanding ascidian metamorphosis.

2.4.4 Perspectives

Although my current investigation did not include the biochemical analysis of enzymatic activity, that analysis would be necessary to reveal whether the tunicate GH6-1 can catalyze cellulose hydrolysis or interact with any other carbohydrates. Possible methods were described in section 1.4.4. In addition, as described in previous sections, another family of genes, Glycoside Hydrolase Family 9, in tunicate genomes (Lo et al., 2003; Davison and Blaxter, 2005) is another endogenous family of possibly active cellulase, but the character of this gene family in tunicate biology is not well known. Confirming the characters of all possible cellulases with experimental evidence would be helpful to understand how tunicates utilize and regulate extracellular cellulose. This knowledge would also shed light on the evolution of tunic and unique tunicate life forms.

After metamorphosis, the adult ascidians may still express *CesA* gene and synthesize cellulose to build and maintain their tunic. This is supported by a low level of *CesA* expression in an EST dataset of *Ciona* on the Ghost database (Satou et al., 2003; Satou et al., 2019). However, the same dataset showed very low or no expression of *GH6-1* gene in adult samples. It would be valuable to know the detailed expression of *CesA* and *GH6-1* genes in adult *Ciona* and how adult ascidians maintain their tunic structure.

In other animals, some of the horizontally transferred genes also play important roles in development. The insect *oskar* gene, originated from combining bacterial and eukaryotic genetic information, has multiple roles in insect development and is critical for fruit fly (*Drosophila melanogaster*) germ plasm formation (Ewen-Campen et al., 2012; Blondel et al., 2020). Mammalian syncytins, originated from multiple integrated retroviral envelope genes, are influential to placenta development (Frendo et al., 2003; Feschotte and Gilbert, 2012). Following the observation that *Ciona CesA* affects metamorphosis, this study showed that *Ciona GH6-1* gene is also important to *Ciona* early development. I expect future studies on *GH6-1* would provide a more comprehensive view on tunicate development, physiology, and evolution.

Conclusion

In this study, I first investigated the phylogenetic relationships, sequence signatures, and exon-intron structure of tunicate *GH6-1* genes. I found that (1) *GH6-1* genes have no homologs in non-tunicate animal taxa and (2) GH6-1 sequences, forming an independent orthologous group, show close relationship with the GH6 domain part of tunicate CesA (CesA-GH6). Also considering the findings of splice sites (intron location) analysis, I propose that a GH6-encoding gene of a microorganism was transferred to an ancestral tunicate and this ancient *GH6* gene duplicated in tunicate genome and later became *GH6-1* and *CesA-GH6*. Based on analysis of the predicted GH6-1 protein sequences, tunicate GH6-1 proteins may be secreted cellulases, but biochemical evidence is yet lacking.

To investigate the expression of tunicate *GH6-1* gene, I used *Ciona intestinalis* type A as an experiment model. Quantitative expression analysis by RT-qPCR showed *GH6-1* expression at late embryonic and hatching larval stages. This temporal expression profile is reminiscent of *Ci-CesA* gene expression profile. Spatial expression analysis by *in situ* hybridization showed epidermal expression of *GH6-1* in *Ciona* embryos and showed locally enhanced expression in a few locations: tail tip, tail midlines, and anterior trunk epidermis. Analysis of single-cell transcriptome of a late tailbud stage I dataset showed that both *GH6-1*- and *CesA*-expressing cells are mostly in cell clusters of epidermal identity, but not many cells show strong co-expression for both genes. Localized signal in the reporter assay partially represented the endogenous *GH6-1* expression, suggesting the existence of specific enhancers upstream to *Ciona GH6-1* gene.

As a preliminary study on the function of *Ciona GH6-1*, I prepared *GH6-1* knock out *Ciona* by TALEN-mediated genome editing. The affected embryos show perturbed papillae formation and metamorphosis, as well as altered cellulose amount. Most of these affected larvae could settle to a surface but did not continue the metamorphic events.

My study provided an example of a horizontally transferred gene in tunicate genome being expressed in early development and utilized in metamorphic regulation. Tunicate *CesA* gene, another horizontally transferred gene, provides tunicates the ability to utilize cellulose and also contributes to metamorphic regulation in *Ciona* (Sasakura et al., 2005). As cellulose of tunicates and the sessile life form of ascidians are among the most known characters of tunicates, my current observations and future studies on *GH6-1* would help us understand how horizontally transferred genes influenced tunicate evolution.

References

ANISEED [Online]. Available: https://www.aniseed.cnrs.fr/ [Accessed 2019-12-20].

Botryllus schlosseri Genome Project [Online]. Available: http://botryllus.stanford.edu/botryllusgenome/ [Accessed 2020-02-28].

Carbohydrate Active Enzymes database [Online]. Available: http://www.cazy.org/ [Accessed 2021-04-30].

Ciona Intestinalis Transgenic line RESources, CITRES [Online]. Available: https://marinebio.nbrp.jp/ciona/forwardToProtocolsAction.do [Accessed 2020-10-14].

Conserved Domain Database (CDD), NCBI. https://www.ncbi.nlm.nih.gov/cdd/

Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/

Ghost Database [Online]. Available: http://ghost.zool.kyoto-u.ac.jp/download_kh.html [Accessed 2019-12-17].

HMMER hmmscan [Online]. Available: https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan [Accessed 2020-06-08].

InterPro [Online]. Available: https://www.ebi.ac.uk/interpro/ [Accessed 2020-03-01].

OikoBase [Online]. Available: http://oikoarrays.biology.uiowa.edu/Oiko/index.html [Accessed 2019-12-25].

TAL Effector Nucleotide Targeter 2.0 [Online]. Available: https://tale-nt.cac.cornell.edu/ [Accessed 2020-10-14].

TUNICANATO tunicate anatomical and developmental ontology. https://www.bpni.bio.keio.ac.jp/tunicanato/3.0/

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., Von Heijne, G. & Nielsen, H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology,* 37**,** 420-423.

Appeltans, W., Ahyong, S. T., Anderson, G., Angel, M. V., Artois, T., Bailly, N., Bamber, R., Barber, A., Bartsch, I., Berta, A.*, et al.* 2012. The magnitude of global marine species diversity. *Curr Biol,* 22**,** 2189-202.

Arola, S. & Linder, M. B. 2016. Binding of cellulose binding modules reveal differences between cellulose substrates. *Scientific Reports,* 6.

Batta-Lona, P. G., Maas, A. E., O'neill, R. J., Wiebe, P. H. & Bucklin, A. 2017. Transcriptomic profiles of spring and summer populations of the Southern Ocean salp, *Salpa thompsoni*, in the Western Antarctic Peninsula region. *Polar Biology,* 40**,** 1261-1276.

Bhattachan, P. & Dong, B. 2017. Origin and evolutionary implications of introns from analysis of cellulose synthase gene. *Journal of Systematics and Evolution,* 55**,** 142-148.

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. 1965. Structure of Hen Egg-White Lysozyme: A Three-dimensional Fourier Synthesis at 2 Å Resolution. *Nature,* 206**,** 757-761.

Blondel, L., Jones, T. E. M. & Extavour, C. G. 2020. Bacterial contribution to genesis of the novel germ line determinant *oskar*. *eLife,* 9.

Boehlke, C., Zierau, O. & Hannig, C. 2015. Salivary amylase – The enzyme of unspecialized euryphagous animals. *Archives of Oral Biology,* 60**,** 1162-1176.

Boto, L. 2014. Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proc Biol Sci,* 281**,** 20132450.

Brown, R. M. 2006. The Biosynthesis of Cellulose. *Journal of Macromolecular Science, Part A,* 33**,** 1345-1373.

Brown, R. M. & Saxena, I. M. 2007. *Cellulose: Molecular and Structural Biology*, Springer Netherlands, 978-1-4020-5380-1(e-book).

Brunetti, R., Gissi, C., Pennati, R., Caicci, F., Gasparini, F. & Manni, L. 2015. Morphological evidence that the molecularly determined *Ciona intestinalis* type A and type B are different species: *Ciona robusta* and *Ciona intestinalis*. *Journal of Zoological Systematics and Evolutionary Research,* 53**,** 186-193.

Buisson, G., Duée, E., Haser, R. & Payan, F. 1987. Three dimensional structure of porcine pancreatic alpha-amylase at 2.9 A resolution. Role of calcium in structure and activity. *The EMBO journal,* 6**,** 3909-3916.

Cao, C., Lemaire, L. A., Wang, W., Yoon, P. H., Choi, Y. A., Parsons, L. R., Matese, J. C., Wang, W., Levine, M. & Chen, K. 2019. Comprehensive single-cell transcriptome lineages of a proto-vertebrate. *Nature,* 571**,** 349-354.

Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics,* 25**,** 1972-3.

Chambon, J.-P., Soule, J., Pomies, P., Fort, P., Sahuquet, A., Alexandre, D., Mangeat, P.-H. & Baghdiguian, S. 2002. Tail regression in *Ciona intestinalis* (Prochordate) involves a Caspase-dependent apoptosis event associated with ERK activation. *Development,* 129**,** 3105-3114.

Chiba, S., Sasaki, A., Nakayama, A., Takamura, K. & Satoh, N. 2004. Development of *Ciona* *intestinalis* Juveniles (Through 2nd Ascidian Stage). *Zoological Science,* 21**,** 285-298.

Christiaen, L., Wagner, E., Shi, W. & Levine, M. 2009. The Sea Squirt *Ciona intestinalis*. *Cold Spring Harbor Protocols,* 2009**,** pdb.emo138-pdb.emo138.

Cloney, R. A. 1982. Ascidian Larvae and the Events of Metamorphosis. *American Zoologist,* 22**,** 817-826.

Corbo, J. C., Levine, M. & Zeller, R. W. 1997. Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, *Ciona intestinalis*. *Development,* 124**,** 589-602.

Coughlan, M. P. 1985. The Properties of Fungal and Bacterial Cellulases with Comment on their Production and Application. *Biotechnology and Genetic Engineering Reviews,* 3**,** 39-110.

Danks, G., Campsteijn, C., Parida, M., Butcher, S., Doddapaneni, H., Fu, B., Petrin, R., Metpally, R., Lenhard, B., Wincker, P*., et al.* 2013. OikoBase: a genomics and developmental transcriptomics resource for the urochordate *Oikopleura dioica*. *Nucleic Acids Res,* 41**,** D845-53.

Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics,* 27**,** 1164-5.

Daugavet, M. A., Shabelnikov, S., Shumeev, A., Shaposhnikova, T., Adonin, L. S. & Podgornaya, O. 2019. Features of a novel protein, rusticalin, from the ascidian *Styela rustica* reveal ancestral horizontal gene transfer event. *Mobile DNA,* 10.

Davidson, E. H. 2010. Emerging properties of animal gene regulatory networks. *Nature,* 468**,** 911-920.

Davison, A. & Blaxter, M. 2005. Ancient Origin of Glycosyl Hydrolase Family 9 Cellulase Genes. *Molecular Biology and Evolution,* 22**,** 1273-1284.

Dehal, P., Satou, Y., Campbell, R. K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D. M*., et al.* 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science,* 298**,** 2157-67.

Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature,* 439**,** 965-968.

Delsuc, F., Philippe, H., Tsagkogeorga, G., Simion, P., Tilak, M. K., Turon, X., Lopez-Legentil, S., Piette, J., Lemaire, P. & Douzery, E. J. P. 2018. A phylogenomic framework and timescale for comparative studies of tunicates. *BMC Biol,* 16**,** 39.

Delsuc, F., Tsagkogeorga, G., Lartillot, N. & Philippe, H. 2008. Additional molecular support for the new chordate phylogeny. *Genesis,* 46**,** 592-604.

Denoeud, F., Henriet, S., Mungpakdee, S., Aury, J. M., Da Silva, C., Brinkmann, H., Mikhaleva, J., Olsen, L. C., Jubin, C., Canestro, C.*, et al.* 2010. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science,* 330**,** 1381-5.

Doyle, E. L., Booher, N. J., Standage, D. S., Voytas, D. F., Brendel, V. P., Vandyk, J. K. & Bogdanove, A. J. 2012. TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic Acids Research,* 40**,** W117-W122.

Dunning Hotopp, J. C. 2011. Horizontal gene transfer between bacteria and animals. *Trends Genet,* 27**,** 157-63.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A.*, et al.* 2019. The Pfam protein families database in 2019. *Nucleic Acids Research,* 47**,** D427-D432.

Endean, R. 1961. The Test of the Ascidian, *Phallusia mammillata*. *Quarterly Journal of Microscopical Science,* 102**,** 107-117.

Endler, A. & Persson, S. 2011. Cellulose synthases and synthesis in *Arabidopsis*. *Mol Plant,* 4**,** 199-211.

Ewen-Campen, B., Srouji, John r., Schwager, Evelyn e. & Extavour, Cassandra g. 2012. *oskar* Predates the Evolution of Germ Plasm in Insects. *Current Biology,* 22**,** 2278-2283.

Fekete, E., Flipphi, M., Ag, N., Kavalecz, N., Cerqueira, G., Scazzocchio, C. & Karaffa, L. 2017. A mechanism for a single nucleotide intron shift. *Nucleic Acids Res,* 45**,** 9085-9092.

Feschotte, C. & Gilbert, C. 2012. Endogenous viruses: insights into viral evolution and impact on host biology. *Nature Reviews Genetics,* 13**,** 283-296.

Forster, B., Van De Ville, D., Berent, J., Sage, D. & Unser, M. 2004. Complex wavelets for extended depth-of-field: a new method for the fusion of multichannel microscopy images. *Microsc Res Tech,* 65**,** 33-42.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Research,* 32**,** W273-W279.

Frendo, J.-L., Olivier, D., Cheynet, V. R., Blond, J.-L., Bouton, O., Vidaud, M., Rabreau, M. L., Evain-Brion, D. L. & Mallet, F. O. 2003. Direct Involvement of HERV-W Env Glycoprotein in Human Trophoblast Cell Fusion and Differentiation. *Molecular and Cellular Biology,* 23**,** 3566-3574.

Ganot, P., Kallesoe, T., Reinhardt, R., Chourrout, D. & Thompson, E. M. 2004. Spliced-leader RNA trans splicing in a chordate, *Oikopleura dioica*, with a compact genome. *Mol Cell Biol,* 24**,** 7795-805.

Garcia-Vallve, S., Romeu, A. & Palau, J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res,* 10**,** 1719-25.

Gmachl, M. & Kreil, G. 1993. Bee venom hyaluronidase is homologous to a membrane protein of mammalian sperm. *Proceedings of the National Academy of Sciences,* 90**,** 3569-3573.

Govindarajan, A. F., Bucklin, A. & Madin, L. P. 2010. A molecular phylogeny of the Thaliacea. *Journal of Plankton Research,* 33**,** 843-853.

Griffiths, J. S., Šola, K., Kushwaha, R., Lam, P., Tateno, M., Young, R., Voiniciuc, C., Dean, G., Mansfield, S. D., Debolt, S.*, et al.* 2015. Unidirectional movement of cellulose synthase complexes in Arabidopsis seed coat epidermal cells deposit cellulose involved in mucilage extrusion, adherence, and ray formation. *Plant Physiology,* 168**,** 502-520.

Henrissat, B. 1991. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochemical Journal,* 280**,** 309-316.

Henrissat, B. & Davies, G. 1997. Structural and sequence-based classification of glycoside hydrolases. *Current Opinion in Structural Biology,* 7**,** 637-644.

Henrissat, B., Teeri, T. T. & Warren, R. a. J. 1998. A scheme for designating enzymes that hydrolyse the polysaccharides in the cell walls of plants. *FEBS Letters,* 425**,** 352-354.

Holland, L. Z. 2016. Tunicates. *Current Biology,* 26**,** R146-R152.

Hotta, K., Dauga, D. & Manni, L. 2020. The ontology of the anatomy and development of the solitary ascidian *Ciona*: the swimming larva and its metamorphosis. *Scientific Reports,* 10.

Hotta, K., Mitsuhara, K., Takahashi, H., Inaba, K., Oka, K., Gojobori, T. & Ikeo, K. 2007. A web-based interactive developmental table for the ascidian *Ciona intestinalis*, including 3D real-image embryo reconstructions: I. From fertilized egg to hatching larva. *Dev Dyn,* 236**,** 1790-805.

Hozumi, A., Matsunobu, S., Mita, K., Treen, N., Sugihara, T., Horie, T., Sakuma, T., Yamamoto, T., Shiraishi, A., Hamada, M.*, et al.* 2020. GABA-Induced GnRH Release Triggers Chordate Metamorphosis. *Current Biology,* 30**,** 1555-1561.e4.

Huber, J. L., Da Silva, K. B., Bates, W. R. & Swalla, B. J. 2000. The evolution of anural larvae in molgulid ascidians. *Seminars in Cell & Developmental Biology,* 11**,** 419-426.

Husnik, F. & Mccutcheon, J. P. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat Rev Microbiol,* 16**,** 67-79.

Imai, K. S., Stolfi, A., Levine, M. & Satou, Y. 2009. Gene regulatory networks underlying the compartmentalization of the *Ciona* central nervous system. *Development,* 136**,** 285-293.

Inoue, J., Nakashima, K. & Satoh, N. 2019. ORTHOSCOPE analysis reveals the presence of the cellulose synthase gene in all tunicate genomes but not in other animal genomes. *Genes (Basel),* 10**,** 294.

Intra, J., Pavesi, G. & Horner, D. S. 2008. Phylogenetic analyses suggest multiple changes of substrate specificity within the Glycosyl hydrolase 20 family. *BMC Evolutionary Biology,* 8**,** 214.

Jain, R., Rivera, M. C. & Lake, J. A. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences,* 96**,** 3801-3806.

Jeffery, W. R. 2007. Chordate ancestry of the neural crest: New insights from ascidians. *Seminars in Cell & Developmental Biology,* 18**,** 481-491.

Jeffery, W. R., Swalla, B. J., Ewing, N. & Kusakabe, T. 1999. Evolution of the ascidian anural larva: evidence from embryos and molecules. *Molecular Biology and Evolution,* 16**,** 646-654.

Joly, J.-S., Kano, S., Matsuoka, T., Auger, H., Hirayama, K., Satoh, N., Awazu, S., Legendre, L. & Sasakura, Y. 2007. Culture of *Ciona intestinalis* in closed systems. *Developmental Dynamics,* 236**,** 1832-1840.

Jue, N. K., Batta-Lona, P. G., Trusiak, S., Obergfell, C., Bucklin, A., O'neill, M. J. & O'neill, R. J. 2016. Rapid evolutionary rates and unique genomic signatures discovered in the first reference genome for the Southern Ocean salp, *Salpa thompsoni* (Urochordata, Thaliacea). *Genome Biol Evol,* 8**,** 3171-3186.

Kamiya, C., Ohta, N., Ogura, Y., Yoshida, K., Horie, T., Kusakabe, T. G., Satake, H. & Sasakura, Y. 2014. Nonreproductive role of gonadotropin-releasing hormone in the control of ascidian metamorphosis. *Developmental Dynamics,* 243**,** 1524-1535.

Kapustin, Y., Souvorov, A., Tatusova, T. & Lipman, D. 2008. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol Direct,* 3**,** 20.

Katoh, K., Rozewicki, J. & Yamada, K. D. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform,* 20**,** 1160-1166.

Katoh, K. & Standley, D. M. 2013. MAFFT multiple sequence alignment software version

7: improvements in performance and usability. *Mol Biol Evol,* 30**,** 772-80.

Kimura, S. & Itoh, T. 1996. New cellulose synthesizing complexes (terminal complexes) involved in animal cellulose biosynthesis in the tunicate *Metandrocarpa uedai*. *Protoplasma,* 194**,** 151-163.

Kimura, Y., Yoshida, M. & Morisawa, M. 2003. Interaction between noradrenaline or adrenaline and the β1-adrenergic receptor in the nervous system triggers early metamorphosis of larvae in the ascidian, *Ciona savignyi*. *Developmental Biology,* 258**,** 129-140.

Koivula, A., Ruohonen, L., Wohlfahrt, G., Reinikainen, T., Teeri, T. T., Piens, K., Claeyssens, M., Weber, M., Vasella, A., Becker, D.*, et al.* 2002. The active site of cellobiohydrolase Cel6A from *Trichoderma reesei*: the roles of aspartic acids D221 and D175. *J Am Chem Soc,* 124**,** 10015-24.

Koonin, E. V., Makarova, K. S. & Aravind, L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol,* 55**,** 709-42.

Kowalewski, A. O. 1866. Entwicklungsgeschichte der einfachen Ascidien. *Mémoires de l'Académie Impériale des Sciences de Saint Pétersbourg,* 7**,** 1-19.

Lambert, C. C. & Brandt, C. L. 1967. The Effect of Light on the Spawning of *Ciona Intestinalis*. *The Biological Bulletin,* 132**,** 222-228.

Lemaire, P. 2011. Evolutionary crossroads in developmental biology: the tunicates. *Development,* 138**,** 2143-52.

Li, K.-L., Lu, T.-M. & Yu, J.-K. 2014. Genome-wide survey and expression analysis of the *bHLH-PAS* genes in the amphioxus *Branchiostoma floridae* reveal both conserved and diverged expression patterns between cephalochordates and vertebrates. *Evodevo,* 5**,** 20.

Li, K.-L., Nakashima, K., Inoue, J. & Satoh, N. 2020. Phylogenetic analyses of Glycosyl Hydrolase Family 6 genes in tunicates: possible horizontal transfer. *Genes,* 11**,** 937.

Lin, C. C. & Aronson, J. M. 1970. Chitin and cellulose in the cell walls of the oomycete, *Apodachlya* sp. *Archiv fur Mikrobiologie,* 72**,** 111-114.

Lo, N., Watanabe, H. & Sugimura, M. 2003. Evidence for the presence of a cellulase gene in the last common ancestor of bilaterian animals. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 270.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research,* 42**,** D490-D495.

Manni, L., Lane, N. J., Joly, J.-S. P., Gasparini, F., Tiozzo, S., Caicci, F., Zaniolo, G. &

Burighel, P. 2004. Neurogenic and non-neurogenic placodes in ascidians. *Journal of Experimental Zoology,* 302B**,** 483-504.

Martin, W. F. 2017. Too much eukaryote LGT. *Bioessays,* 39.

Matsunobu, S. & Sasakura, Y. 2015. Time course for tail regression during metamorphosis of the ascidian *Ciona intestinalis*. *Dev Biol,* 405**,** 71-81.

Matthysse, A. G. 1983. Role of bacterial cellulose fibrils in *Agrobacterium tumefaciens* infection. *Journal of Bacteriology,* 154**,** 906.

Matthysse, A. G., Deschet, K., Williams, M., Marry, M., White, A. R. & Smith, W. C. 2004. A functional cellulose synthase from ascidian epidermis. *Proceedings of the National Academy of Sciences,* 101**,** 986-91.

Mazet, F., Hutt, J. A., Milloz, J., Millard, J., Graham, A. & Shimeld, S. M. 2005. Molecular evidence from *Ciona intestinalis* for the evolutionary origin of vertebrate sensory placodes. *Developmental Biology,* 282**,** 494-508.

Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE), 14 November 2010 2010 New Orleans, LA, USA. 1 - 8.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, Gustavo a., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J*., et al.* 2021. Pfam: The protein families database in 2021. *Nucleic Acids Research,* 49**,** D412-D419.

Modelski, M. J., Menlah, G., Wang, Y., Dash, S., Wu, K., Galileo, D. S. & Martin-Deleon, P. A. 2014. Hyaluronidase 2: A Novel Germ Cell Hyaluronidase with Epididymal Expression and Functional Roles in Mammalian Sperm. *Biology of Reproduction,* 91.

Murphy, C. L., Youssef, N. H., Hanafy, R. A., Couger, M. B., Stajich, J. E., Wang, Y., Baker, K., Dagar, S. S., Griffith, G. W., Farag, I. F*., et al.* 2019. Horizontal gene transfer as an indispensable driver for evolution of Neocallimastigomycota into a distinct gut-dwelling fungal lineage. *Appl Environ Microbiol,* 85.

Nakashima, K., Nishino, A., Horikawa, Y., Hirose, E., Sugiyama, J. & Satoh, N. 2011. The crystalline phase of cellulose changes under developmental control in a marine chordate. *Cell Mol Life Sci,* 68**,** 1623-31.

Nakashima, K., Yamada, L., Satou, Y., Azuma, J. & Satoh, N. 2004. The evolutionary origin of animal cellulose synthase. *Dev Genes Evol,* 214**,** 81-8.

Nakayama-Ishimura, A., Chambon, J. P., Horie, T., Satoh, N. & Sasakura, Y. 2009. Delineating metamorphic pathways in the ascidian *Ciona intestinalis*. *Dev Biol,* 326**,** 357-67.

Naville, M., Henriet, S., Warren, I., Sumic, S., Reeve, M., Volff, J. N. & Chourrout, D. 2019. Massive changes of genome size driven by expansions of non-autonomous transposable elements. *Curr Biol,* 29**,** 1161-1168.

Nixon, J. E., Wang, A., Morrison, H. G., Mcarthur, A. G., Sogin, M. L., Loftus, B. J. & Samuelson, J. 2002. A spliceosomal intron in *Giardia lamblia*. *Proc Natl Acad Sci U S A,* 99**,** 3701-5.

Ohno, M., Kimura, M., Miyazaki, H., Okawa, K., Onuki, R., Nemoto, C., Tabata, E., Wakita, S., Kashimura, A., Sakaguchi, M.*, et al.* 2016. Acidic mammalian chitinase is a proteases-resistant glycosidase in mouse digestive system. *Scientific Reports,* 6.

Ohta, N., Kaplan, N., Ng, J. T., Gravez, B. J. & Christiaen, L. 2020. Asymmetric Fitness of Second-Generation Interspecific Hybrids Between *Ciona robusta* and *Ciona intestinalis*. *G3 Genes|Genomes|Genetics,* 10**,** 2697-2711.

Pál, C., Papp, B. & Lercher, M. J. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics,* 37**,** 1372-1375.

Pasini, A., Amiel, A., Rothbächer, U., Roure, A., Lemaire, P. & Darras, S. 2006. Formation of the Ascidian Epidermal Sensory Neurons: Insights into the Origin of the Chordate Peripheral Nervous System. *PLoS Biology,* 4**,** e225.

Patel, A. A. & Steitz, J. A. 2003. Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol,* 4**,** 960-70.

Patel, S. & Goyal, A. 2017. Chitin and chitinase: Role in pathogenicity, allergenicity and health. *International Journal of Biological Macromolecules,* 97**,** 331-338.

Pennati, R. & Rothbächer, U. 2015. Bioadhesion in ascidians: a developmental and functional genomics perspective. *Interface Focus,* 5**,** 20140061.

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R. & Finn, R. D. 2018. HMMER web server: 2018 update. *Nucleic Acids Res,* 46**,** W200-W204.

Putnam, N. H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V. V.*, et al.* 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science,* 317**,** 86-94.

Q. Irvine, S. 2013. Study of *Cis*-regulatory Elements in the Ascidian *Ciona intestinalis*. *Current Genomics,* 14**,** 56-67.

Rambaut, A. *FigTree v1.4* [Online]. Available: http://tree.bio.ed.ac.uk/software/figtree/ [Accessed 26/11/2018].

Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G. & Koonin, E. V. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Current Biology,* 13**,** 1512-1517.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol,* 61**,** 539-42.

Rstudio Team 2020. RStudio: Integrated Development for R. . http://www.rstudio.com/

Sagane, Y., Zech, K., Bouquet, J. M., Schmid, M., Bal, U. & Thompson, E. M. 2010. Functional specialization of cellulose synthase genes of prokaryotic origin in chordate larvaceans. *Development,* 137**,** 1483-92.

Sakuma, T., Ochiai, H., Kaneko, T., Mashimo, T., Tokumasu, D., Sakane, Y., Suzuki, K., Miyamoto, T., Sakamoto, N., Matsuura, S.*, et al.* 2013. Repeating pattern of non-RVD variations in DNA-binding modules enhances TALEN activity. *Sci Rep,* 3**,** 3379.

Sasaki, H., Yoshida, K., Hozumi, A. & Sasakura, Y. 2014. CRISPR/Cas9-mediated gene knockout in the ascidian *Ciona intestinalis*. *Dev Growth Differ,* 56**,** 499-510.

Sasakura, Y. 2018a. Cellulose production and the evolution of the sessile lifestyle in ascidians. *Sessile Organisms,* 35**,** 21-29.

Sasakura, Y. 2018b. *Transgenic Ascidians*, Springer Nature Singapore, 978-981-10-7544-5.

Sasakura, Y., Nakashima, K., Awazu, S., Matsuoka, T., Nakayama, A., Azuma, J. & Satoh, N. 2005. Transposon-mediated insertional mutagenesis revealed the functions of animal cellulose synthase in the ascidian *Ciona intestinalis*. *Proc Natl Acad Sci U S A,* 102**,** 15134-9.

Sasakura, Y., Ogura, Y., Treen, N., Yokomori, R., Park, S. J., Nakai, K., Saiga, H., Sakuma, T., Yamamoto, T., Fujiwara, S.*, et al.* 2016. Transcriptional regulation of a horizontally transferred gene from bacterium to chordate. *Proc Biol Sci,* 283.

Sato, Y., Terakado, K. & Morisawa, M. 1997. Test cell migration and tunic formation during post-hatching development of the larva of the ascidian, *Ciona intestinalis*. *Dev Growth Differ,* 39**,** 117-26.

Satoh, N. 2003. The ascidian tadpole larva: comparative molecular development and genomics. *Nature Reviews Genetics,* 4**,** 285-295.

Satoh, N. 2014. *Developmental Genomics of Ascidians,* Hoboken, New Jersey, Wiley-Blackwell, 9781118656181.

Satoh, N. 2016. *Origin and Evolution of Chordates,* New York, Academic Press, 9780128099346.

Satou, Y., Hamaguchi, M., Takeuchi, K., Hastings, K. E. & Satoh, N. 2006. Genomic overview of mRNA 5'-leader trans-splicing in the ascidian *Ciona intestinalis*. *Nucleic Acids Res,* 34**,** 3378-88.

Satou, Y., Kawashima, T., Kohara, Y. & Satoh, N. 2003. Large scale EST analyses in *Ciona intestinalis*. *Development Genes and Evolution,* 213**,** 314-318.

Satou, Y., Kawashima, T., Shoguchi, E., Nakayama, A. & Satoh, N. 2005. An integrated database of the ascidian, *Ciona intestinalis*: towards functional genomics. *Zoolog Sci,* 22**,** 837-43.

Satou, Y., Kusakabe, T., Araki, L. & Satoh, N. 1995. Timing of initiation of muscle-specific gene expression in the ascidian embryo precedes that of developmental fate restriction in lineage cells. *Development, Growth & Differentiation,* 37**,** 319-327.

Satou, Y., Mineta, K., Ogasawara, M., Sasakura, Y., Shoguchi, E., Ueno, K., Yamada, L., Matsumoto, J., Wasserscheid, J., Dewar, K.*, et al.* 2008. Improved genome assembly and evidence-based global gene model set for the chordate *Ciona intestinalis*: new insight into intron and operon populations. *Genome Biol,* 9**,** R152.

Satou, Y., Nakamura, R., Yu, D., Yoshida, R., Hamada, M., Fujie, M., Hisata, K., Takeda, H. & Satoh, N. 2019. A nearly complete genome of *Ciona intestinalis* type A (*C. robusta*) reveals the contribution of inversion to chromosomal evolution in the genus *Ciona*. *Genome Biol Evol,* 11**,** 3144-3157.

Satou, Y., Sato, A., Yasuo, H., Mihirogi, Y., Bishop, J., Fujie, M., Kawamitsu, M., Hisata, K., Satoh, N. & O'neill, R. 2021. Chromosomal inversion polymorphisms in two sympatric ascidian lineages. *Genome Biology and Evolution*.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B.*, et al.* 2012. Fiji: an open-source platform for biological-image analysis. *Nat Methods,* 9**,** 676-82.

Sean, E., Woolfe, A., Goodson, M., Goode, D. K., Snell, P., Mcewen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H.*, et al.* 2004. Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLoS Biology,* 3**,** e7.

Sharma, H. K., Qin, W. & Xu, C. 2018. Cellobiohydrolase (CBH) Activity Assays. *In:* Lübeck, M. (ed.) *Cellulases: Methods and Protocols.* New York, NY: Springer New York.

Sigrist, C. J., De Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L. & Xenarios, I. 2012. New and continuing developments at PROSITE. *Nucleic Acids Res,* 41**,** D344-7.

Soucy, S. M., Huang, J. & Gogarten, J. P. 2015. Horizontal gene transfer: building the web of life. *Nat Rev Genet,* 16**,** 472-82.

Spriestersbach, A., Kubicek, J., Schäfer, F., Block, H. & Maertens, B. 2015. Purification of His-Tagged Proteins. 559**,** 1-15.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30**,** 1312-3.

Stolfi, A., Ryan, K., Meinertzhagen, I. A. & Christiaen, L. 2015. Migratory neuronal progenitors arise from the neural plate borders in tunicates. *Nature,* 527**,** 371-374.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P. & Satija, R. 2019. Comprehensive Integration of Single-Cell Data. *Cell,* 177**,** 1888-1902.e21.

Tassy, O., Dauga, D., Daian, F., Sobral, D., Robin, F., Khoueiry, P., Salgado, D., Fox, V., Caillol, D., Schiappa, R.*, et al.* 2010. The ANISEED database: Digital representation, formalization, and elucidation of a chordate developmental program. *Genome Research,* 20**,** 1459-1468.

Treen, N., Yoshida, K., Sakuma, T., Sasaki, H., Kawai, N., Yamamoto, T. & Sasakura, Y. 2014. Tissue-specific and ubiquitous gene knockouts by TALEN electroporation provide new approaches to investigating gene function in *Ciona*. *Development,* 141**,** 481-7.

Tsagkogeorga, G., Turon, X., Hopcroft, R. R., Tilak, M.-K., Feldstein, T., Shenkar, N., Loya, Y., Huchon, D., Douzery, E. J. P. & Delsuc, F. 2009. An updated 18S rRNA phylogeny of tunicates based on mixture and secondary structure models. *BMC Evolutionary Biology,* 9**,** 187.

Tu, Q., Cameron, R. A., Worley, K. C., Gibbs, R. A. & Davidson, E. H. 2012. Gene structure in the sea urchin *Strongylocentrotus purpuratus* based on transcriptome analysis. *Genome Res,* 22**,** 2079-87.

Uversky, V. N., Wohlkönig, A., Huet, J., Looze, Y. & Wintjens, R. 2010. Structural relationships in the lysozyme superfamily: Significant evidence for glycoside hydrolase signature motifs. *PLoS ONE,* 5**,** e15388.

Vain, T., Crowell, E. F., Timpano, H., Biot, E., Desprez, T., Mansoori, N., Trindade, L. M., Pagant, S., Robert, S., Hofte, H.*, et al.* 2014. The cellulase KORRIGAN is part of the cellulose synthase complex. *Plant Physiol,* 165**,** 1521-1532.

Van Daele, Y., Revol, J.-F., Gaill, F. & Goffinet, G. 1992. Characterization and supramolecular architecture of the cellulose-protein fibrils in the tunic of the sea peach (*Halocynthia papillosa*, Ascidiacea, Urochordata). *Biology of the Cell,* 76**,** 87-96.

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

Wakai, M. K., Nakamura, M. J., Sawai, S., Hotta, K. & Oka, K. 2021. Two-Round Ca 2+ transient in papillae by mechanical stimulation induces metamorphosis in the ascidian *Ciona intestinalis* type A. *Proceedings of the Royal Society B: Biological Sciences,* 288**,** 20203207.

Wang, J., Karabinos, A., Zimek, A., Meyer, M., Riemer, D., Hudson, C., Lemaire, P. & Weber, K. 2002. Cytoplasmic intermediate filament protein expression in tunicate development: a specific marker for the test cells. *European Journal of Cell Biology,* 81**,** 302-311.

Watanabe, H. & Tokuda, G. 2001. Animal cellulases. *Cellular and Molecular Life Sciences,* 58**,** 1167-1178.

Wei, J. & Dong, B. 2018. Identification and expression analysis of long noncoding RNAs in embryogenesis and larval metamorphosis of *Ciona savignyi*. *Mar Genomics,* 40**,** 64-72.

Wei, J., Zhang, J., Lu, Q., Ren, P., Guo, X., Wang, J., Li, X., Chang, Y., Duan, S., Wang, S.*, et al.* 2020. Genomic basis of environmental adaptation in the leathery sea squirt (*Styela clava*). *Molecular Ecology Resources,* 20**,** 1414-1431.

Yue, J.-X., Kozmikova, I., Ono, H., Nossa, C. W., Kozmik, Z., Putnam, N. H., Yu, J.-K. & Holland, L. Z. 2016. Conserved Noncoding Elements in the Most Distant Genera of Cephalochordates: The Goldilocks Principle. *Genome Biology and Evolution,* 8**,** 2387-2405.

Zeller, R. W. 2018. Electroporation in Ascidians: History, Theory and Protocols. *In:* Sasakura, Y. (ed.) *Transgenic Ascidians* Springer Nature Singapore.

Zeng, F., Wunderer, J., Salvenmoser, W., Hess, M. W., Ladurner, P. & Rothbächer, U. 2019. Papillae revisited and the nature of the adhesive secreting collocytes. *Developmental Biology,* 448**,** 183-198.