DRAFT GENOME OF HAIRY SEA CUCUMBER *SCLERODACTYLA BRIAREUS*, A MODEL TO STUDY GENE FAMILIES FOR TISSUE REGENERATION AND HOST-VIRAL INTERACTIONS IN ECHINODERMS

by

Varnika Mittal

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2024

Approved by:

_____
Dr. Robert W. Reid

_____
Dr. Daniel A. Janies

_____
Dr. Jessica Schlueter

_____
Dr. Richard Allen White III

_____
Dr. Adam M. Reitzel

ABSTRACT

VARNIKA MITTAL. Draft Genome of Hairy Sea Cucumber *Sclerodactyla briareus*, A Model to Study Gene Families for Tissue Regeneration and Host-Viral Interactions in Echinoderms. (Under the direction of DR. ROBERT W. REID)

Echinoderms are highly regenerative animals that share a common ancestor with chordates, including humans. While the two phyla share a common ancestor, echinoderms defeat humans when it comes to regeneration. Regeneration is the replacement of damaged cells or regrowth of damaged tissues or organs naturally. Despite the significant differences in the body plan of echinoderms and humans, the similarities in their genome structure, the genes these two groups share, the phylogenetic relationship they have, and the simplicity of experimentation make echinoderms a valuable group to study regeneration. We expect that understanding tissue regeneration in echinoderms can set a stage for improved treatments and provide insights for developing therapeutic approaches to treat human injuries in the future.

Even within such a highly regenerative phylum as echinoderms, some species regenerate more readily than others. Holothurians, commonly known as sea cucumbers, occupy a special place in this regard, as they can fully and rapidly regenerate their body parts and major organs, including: the viscera, central nervous system, body wall, and muscles. However, the available genomic resources are very limited to implement holothuroids as animal models to study regeneration. Moreover, the available genomic resources do not represent diversity within the phylum. Hence, to fill this gap, I have updated an easy-to-use web-based application, EchinoDB, a database resource that includes the genomic and transcriptomic data on 42 unique echinoderm species, spanning the deepest divergences within the five extant classes of the phylum in addition to the 2 new major datasets: the RNA-Seq data of the brittle star *Ophioderma brevispinum* and the high-quality genomic assembly data of the green sea urchin *Lytechinus variegatus*.

Among sea cucumbers, the vast majority of molecular studies have been done on a single species, *Holothuria glaberrima* which do not represent the diversity of various regenerative events, including: regeneration of the gut luminal epithelium (mesodermal to endodermal) and regeneration of the pharyngeal bulb. However, other sea cucumber species, especially, those of the order Dendrochirotida are capable of such exceptional regeneration events. Therefore, I sequenced and annotated the draft genome of a dendrochirotid *Sclerodactyla briareus* to gain a deeper understanding of the regulatory molecular mechanisms controlling regeneration and genomic aspects behind the diversity of regeneration, seen in echinoderms.

To illustrate the practical utility of the dendrochirotid genome for regeneration studies, key components of the Notch and Wnt signaling pathways were selected and identified in the genome of hairy sea cucumber *S. briareus*. This is a biologically relevant example as these pathways are crucial for tissue regeneration in echinoderms. They are highly conserved across all multi-cellular animals and are known to coordinate many cellular events, including: cell proliferation, de-differentiation, cell division, and apoptosis. Therefore, I aimed to retrieve 29 selected genes of the Notch pathway and 25 selected genes of the canonical Wnt signaling pathway. Except for Mesp2 (a Notch pathway gene), all other genes were identified in the newly assembled draft genome of *S. briareus*.

I also studied *S. briareus* for primordial host-viral interactions and to learn about the evolution of their immune system by looking at the recombination activating genes (*RAG*) in relation to *Strongylocentrotus purpuratus* and other echinoderms. The objective was to discover and characterize novel viral sequences within *S. briareus* alongside the evolution of immune genes (*RAG-Like*) in marine environment. However, because of the gaps in the assembly, I was unable to find any evidence of viral markers in the genome of *S. briareus*. The paucity of full-length contigs in the genome assembly also resulted into only 3 protein sequences that may potentially

share a sequence homology with *RAG1-Like* gene, but further investigation is needed. The lack of results require improvements in the genome assembly and the availability of increased data for *RAG-Like* genes on echinoderms. Nevertheless, this work is still useful for regeneration studies on echinoderms.

Keywords: echinoderms, holothuroidea, dendrochirotid, hairy sea cucumber, database, EchinoDB, *de novo* assembly, notch pathway, wnt pathway, viruses, *RAG* genes

DEDICATION

I dedicate this work to my family and friends who have supported me throughout my academic career.

ACKNOWLEDGEMENTS

First, I express my gratitude to my doctoral advisor Robert W. Reid, who has served as a great mentor, challenging me to think critically. I would also like to extend my appreciation to the members of my committee, Dr. Daniel A. Janies, Dr. Rick White III, Dr. Jessica Schlueter, and Dr. Adam Reitzel for their valuable counsel during my academic journey. Special thanks to Dr. Denis Jacob Machado and Dr. Vladimir Mashanov who helped me and supported me throughout my doctorate. I would like to thank Dr. Cory Brouwer, Department of Bioinformatics and Genomics, the College of Computing and Informatics and the Graduate School, UNC Charlotte who generously funded my endeavors as a doctoral candidate.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A. forbesi: *Asterias forbesi*

A. japonicus: *Apostichopus japonicus*

A. planci: *Acanthaster planci*

A. rubens: *Asterias rubens*

C. atratus: *Colobocentrotus atratus*

C. frondosa: *Cucumaria frondosa*

D. melanogaster: *Drosophila melanogaster*

E. mathaei: *Echinometra mathaei*

H. forskali *Holothuria forskali*

H. glabberima: *Holothuria glabberima*

H. sapiens: *Homo sapiens*

L. variegatus: *Lytechinus variegatus*

O. brevispinum: *Ophioderma brevispinum*

P. miniata: *Patiria miniata*

S. briareus: *Sclerodactyla briareus*

S. purpuratus: *Strongylocentrotus purpuratus*

T. briareus: *Thyone briareus*

T. gratilla: *Tripneustes gratilla*

AWS: Amazon Web Services

BAM: Binary Alignment Map

BCM-HGSC: Baylor School of Medicine, Human Genome Sequencing Center

BLAST: Basic Local Alignment Search Tool

bp: base pair

BUSCO: Benchmarking Universal Single-Copy Ortholog

CDD: Conserved Domain Database

cDNA: complementary Deoxyribonucleic Acid

CPU: Central Processing Unit

CRC: Colorectal Cancers

CSS: Cascading Style Sheets

DBG: De Bruijn graph

DNA: Deoxyribonucleic Acid

FASTA: FAST All

FBG: Fuzzy Bruijn Graph

FD: Feulgen Densitometry

GB: GigaByte

gDNA: Genomic Deoxyribonucleic Acid

GEO: Gene Expression Omnibus

HMM: Hidden Markov Model

HMW: High Molecular Weight

HTML: HyperText Markup Language

Igs: Immunoglobulins

LINE: Long Interspersed Nuclear Elements

LTR: Long Terminal Repeats

MB: MegaByte

MCT: Mutable Collagenous Tissues

NCBI: National Center for Biotechnology Information

NM: NanoMeter

NR: RefSeq Non-Redundant Proteins

OLC: Overlap Layout Consensus

PacBio: Pacific Biosciences

PCR: Polymerase Chain Reaction

RAG: Recombination Activating Genes

RdRp: RNA-dependent RNA Polymerase

RefSeq: Reference Sequence Database

RNA: Ribonucleic Acid

RSS: Recombination Signal Sequence

SRA: Sequence Read Archive

SSWD: Sea Star Wasting Disease

TB: TeraByte

TCR: T-Cell Receptors

TE: Transposable Element

UniParc: UniProt Archive

UniProt: Universal Protein Resource

UniProtKB: UniProt Knowledgebase

UniRef: UniProt Reference Clusters

V(D)J: Variable Diversity Joining

CHAPTER 1: Introduction

1.1 What are Echinoderms?

The phylum Echinodermata is composed of marine invertebrate animals commonly known as echinoderms. It contains five classes: Asteroidea, Ophiuroidea, Holothuroidea, Echinoidea, and Crinoidea [34]. Together with hemichordates, the phylum Echinodermata (invertebrate, non-chordate deuterostome) a sister taxon to Chordata, which contains the phylum Vertebrata that includes humans (Figure 1.1) [34].

The body of an echinoderm has radial organization and is composed of multiple (usually five) units, arranged as sectors or rays around the central axis connecting the mouth and the anus [35]. Echinoderms share unique characteristics such as: pentaradial symmetry (or modifications thereof) in adults, a skeleton composed of many ossicles formed of stereom (a calcium carbonate material), a water vascular system, and a mutable collagenous tissue [36].

Apart from their fascinating body plan, echinoderms show some of the most impressive regenerative feats within the animal kingdom. Species of this phylum can regenerate most tissues and organs post-injury or self-induced autotomy [37, 38, 39, 40, 41, 42, 43, 44]. Even within a highly regenerative phylum such as echinoderms, some species have higher regeneration capacity than others. Holothurians, commonly known as sea cucumbers, hold a special place as they can rapidly regenerate most of their body parts. Sea cucumbers can undergo evisceration, a process in which the animal ejects parts of its gut, including some nervous tissue, to scare and defend against potential predators such as crabs and fish. Following evisceration, sea cucumbers can rapidly fill up the injury gap between the torn edges of their old organs to regrow

their missing body parts [38, 41, 45, 46, 47, 48].

Not only echinoderms have a wide range of regenerative capacities [49], but they also occupy a strategic phylogenetic position in relation to chordates (including humans). Because of their key phylogenetic position, reparative genes in echinoderms that might have been evolutionarily preserved, could potentially be re-activated in poorly regenerating vertebrates. The regenerative potential in echinoderms is impeccable as it does not decline with age [50]. Echinoderms are feasible and resistant to cancer. Even though the regeneration process in echinoderms involve extensive cell dedifferentiation and proliferation, these process do not go awry and result in tumors [35, 51]. They are also easy and inexpensive to keep in simple aquaria for extended periods of time. For all these reasons, echinoderms have been slowly gaining special attention as model systems for regeneration studies in recent decades [52, 53].

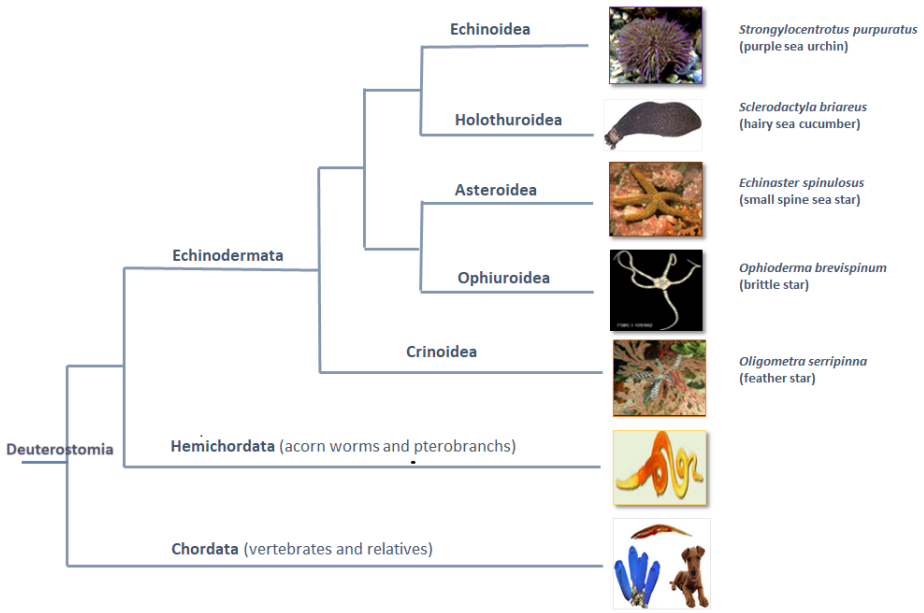

Figure 1.1: Phylogenetic positions of phylum Echinodermata, Hemichordata and Chordata (Deuterostomia). Modified from: TREE of LIFE web project (Echinodermata) [20, 21]

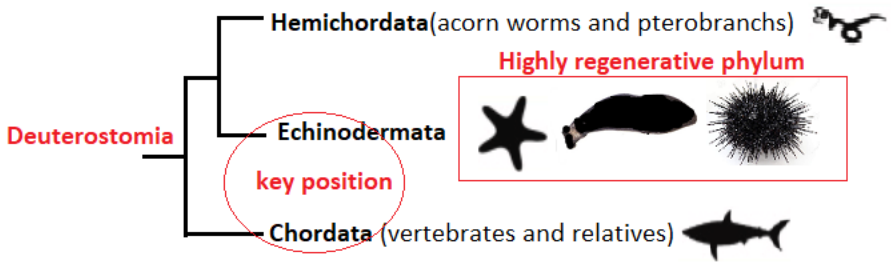

Figure 1.2: Phylogenetic tree showing relationships of selected phyla within the clade Deuterostomia. Modified from: Bromham et al., 1999 [22]

1.2 Advances in echinoderm genomics

Researchers motivated by the biomedical potential of echinoderms have assembled many resources to study these animals. These resources provide genomic information to accelerate research in molecular biology, developmental processes, gene regulatory networks, and regenerative biology.

In 2006, the draft genome of purple sea urchin *S. purpuratus* was sequenced and annotated by the Sea Urchin Genome Sequencing Consortium [54]. The draft assembly was completed using a whole-genome shotgun strategy at the Baylor School of Medicine, Human Genome Sequencing Center (BCM-HGSC). This was the first echinoderm genome sequenced, and it has facilitated many scientific discoveries using echinoderms as a research model [46, 55, 56]. After the original publication, the purple sea urchin genome has been progressively improved, resulting in its version 5.0. This version is estimated to encode about 23,500 genes, of which 7,700 genes are shared with humans [57].

In 2016, Janies et al. [3] released EchinoDB as a web-accessible information system designed to provide genomic, transcriptomic and amino acid sequence data on echinoderms. EchinoDB aims to serve research communities by providing diverse and rich, functional genomics data of echinoderm species. Mittal et al. [2] have recently updated EchinoDB and EchinoDB now includes RNA-seq data of the brittle star *O. brevispinum* [58], phylogenomic data of *Xyloplax* sp. [59, 60, 61] and high-quality

genomic assembly data of the green sea urchin *L. variegatus* [62] along with groups of 749,397 orthologous and paralogous transcripts arranged in orthoclusters by sequence similarity (previous version data from [3]). Therefore, EchinoDB not only facilitates data retrieval (annotated sequences) for various downstream projects, including regeneration, phylogeny, and gene family studies, but is also an excellent resource for studying diversity within the phylum.

Other echinoderm resources currently available on echinoderms are listed in Tables 1.1–1.3. These information resources are highly critical for the echinoderm research and their genome size varies from 0.1 GB to ∼3 GB. The differences in their body plans, which range from stalked flower-like sea lilies to ambulatory stellate starfish and brittle stars, soft-bodied sea cucumbers, spiked, armored, and globose sea urchins, to flat sand dollars [63], could be one of the causes of the variations in their genome sizes. However, there is no concrete evidence to support this theory. Table 1.1 highlights the genomic, transcriptomic, and chromosome-level data available for different species of echinoderms [64, 46, 65, 66, 67, 35, 54, 68, 62, 69, 70, 45, 69, 71]. On the other hand, table 1.2 presents web information systems that allows for querying and exploring genomic data on different species of echinoderms and table 1.3 outlines different databases containing curated and non-curated protein sequences for various species, including echinoderms.

Although these genomic resources facilitate the understanding of gene coding regions and provide a platform for comparative genomics among echinoderms, they are limited to only a small fraction of species within the phylum. For example, the understudied sea cucumbers (class Holothuroidea) regenerate most of their organs, including appendages and nerve cells through a complex process of evisceration, epidermal wound healing, cell division, morphogenesis, redifferentiation, and regeneration completion [46, 47, 48, 72, 66, 73, 41], whereas sea urchins (class Echinoidea), which have been the main focus of the sequencing and annotation efforts so far, are weak in

Table 1.1: Available resources for different species in the phylum Echinodermata. The table contains species, common name, class, family, genome type, and genome size in Giga base pairs (Gbp). The column resource type indicates the assembly state (complete, draft, transcriptome, or chromosome-level). The complete genome represents the full genomic content of the organism without gaps. Draft genome is the initial version that contains segments of contiguous base pairs with gaps. Transcriptome assembly reconstructs the complete set of full-length transcripts from RNA-seq data (or short-read sequences). Finally, assembly at the chromosome level is relatively complete and contains very few gaps.

Species name	**Common name**	**Class: Family**	**Genome Size (Gbp)**	**Resource type**	**Citation or GenBank Accession**
Acanthaster planci	Crown-of-thorns starfish	Asteroidea: Acanthasteridae	0.431	Complete Genome	[65]
Asterias rubens	Sugar sea star	Asteroidea: Asteriidae	0.41	Chromosome	GCA_902459465.3 [6]
Patricia miniata	Bat star	Asteroidea: Asterinidae	0.608	Complete Genome	GCF_015706575.1 [6]
Patiriella regularis	Cushion star	Asteroidea: Asterinidae	0.94	Draft Genome	[45]
Antedon mediterranea	Mediterranean feather star	Crinoidea: Antedonidae	14.2	Transcriptome	[69]
Anneissia japonica	Sea lily	Crinoidea: Comatulidae	0.58	Draft Genome	GCA_011630105.1 [6]
Evechinus chloroticus	Sea urchin Kina	Echinoidea: Echinometridae	0.17	Transcriptome	[64]
Strongylocentrotus purpuratus	Sea urchin	Echinoidea: Strongylocentrotidae	0.814	Complete Genome	[54]
Hemicentrotus pulcherrimus	Sea urchin	Echinoidea: Strongylocentrotidae	0.8	Draft Genome	[68]
Strongylocentrotus franciscanus	Red sea urchin	Echinoidea: Strongylocentrotidae	0.6	Draft Genome	[71]
Lytechinus variegatus	Green sea urchin	Echinoidea: Toxopneustidae	0.87	Chromosomal-Level Genome	[62]
Lytechinus variegatus	Green sea urchin	Echinoidea: Toxopneustidae	1.3	Draft Genome	[71]
Chiridota heheva	Sea cucumber	Holothuroidea: Chiridotidae	1.107	Draft Genome	[67]
Holothuria glaberrima	Brown rock sea cucumber	Holothuroidea: Holothuriidae	1.2	Draft Genome	[66]
Holothuria glaberrima	Brown rock sea cucumber	Holothuroidea: Holothuriidae	0.109	Transcriptome	[35]
Apostichopus japonicas	Japanese spiky sea cucumber	Holothuroidea: Stichopodidae	0.804	Complete Genome	[46]
Apostichopus parvimensis	Warty sea cucumber	Holothuroidea: Stichopodidae	0.87	Draft Genome	GCA_000934455.1 [6]
Australostichopus mollis	Brown sea cucumber	Holothuroidea: Stichopodidae	1.25	Draft Genome	[45]
Apostichopus californicus	Giant California sea cucumber	Holothuroidea: Stichopodidae	0.78	Draft Genome	GCA_020650975.1 [6]
Ophioderma brevispinum	Brittle star	Ophiuroidea: Ophiodermatiidae	2.68	Draft Genome	[70]
Ophionereis fasciata	Mottled brittle star	Ophiuroidea: Ophionereididae	1.18	Draft Genome	[45]
Ophionotus victoriae	Brittle star	Ophiuroidea: Ophiuridae	0.156	Transcriptome	[69]

regeneration and can regenerate parts of their broken or lost spines and pedicellariae [43, 74]. Since the potential for repair and regrowth varies widely within taxonomic

groups, available resources do not fully represent the diversity of echinoderms that have greater regenerative capacities.

Table 1.2: Online databases for echinoderm-specific genomic, transcriptomic, and peptide annotated sequences.

Name	URL	Details	Citation
SpBase	https://spbase.org/	SpBase contains reference sequence date from the purple sea urchin *Strongylocentrotus purpuratus*. In addition, it includes sequence information of three sea urchins: *Strongylocentrotus franciscanus*, *Allocentrotus fragilis*, and *Lytechinus variegatus*. Finally, the database houses a lower amount of data for two other sea urchins (*Arbacia punctulata* and *Eucidaris tribuloides*) as well as a sea star (*Asterina miniata*) and a hemichordate (*Ptychodera flava*).	[75]
EchinoDB	https://echinodb.uncc.edu	EchinoDB contains genomic and amino acid sequence data for 42 Echinoderm transcriptomes as well as the data for *Ophioderma brevispinum* (brittle star) and *Lytechinus variegatus* (green sea urchin). In addition, phylogenomic information is also available for *Xyloplax* species.	[2]
Echinobase	https://www.echinobase.org/entry/	Echinobase includes genome sequence and gene/CDS and protein sequences from two sea urchin species: *Strongylocentrotus purpuratus* and *Lytechinus variegatus*, three sea stars *Acanthaster planci*, *Patiria miniata* and *Asterias rubens* and a feather star, *Anneissia japonica*.	[76]
HpBase	https://cell-innovation.nig.ac.jp/Hpul/	The webpage provides information on *Hemicentrotus pulcherrimus* genome and transcriptome for a wide range of biologists.	[68]

Table 1.3: Available repositories of curated, non-curated protein sequences and annotation data.

Databases Name	Link	Curated Non-Curated	Database Details	Citation
UniProtKB (Universal Protein Knowledgebase)	www.uniprot.org	Both	UniProtKB is a protein database partially curated by experts.	[8]
UniProtKB/TrEMBL	www.bioinfo.pte.hu/more/TrEMBL.htm	Non-Curated (Low Quality)	TrEMBL contains high-quality computationally analyzed, unreviewed, automatically annotated entries.	[13]
RefSeq (NCBI Reference Sequence Database)	www.ncbi.nlm.nih.gov/refseq	Curated	RefSeq is an open access, annotated and curated collection of publicly available nucleotide sequences and their protein products.	[4]
PIR (Protein Information Resource)	proteininformationresource.org	Curated (High Quality)	Non-redundant annotated protein sequence database.	[77]
GenBank	www.ncbi.nlm.nih.gov/genbank	Non-Curated (Low Quality)	Unreviewed sequences submitted from individual laboratories and large-scale sequencing projects.	[6]
PDB (Protein Data Bank)	www.rcsb.org	Curated (High Quality)	Contains experimentally-determined structures of proteins, nucleic acids, and complex assemblies.	[78]
Nr (Non-Redundant Database)	www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins	Both	Nr database encompasses sequences from both non-curated and curated databases.	[4]

Even within the highly regenerative class as Holothuroidea, there is a wide diversity in regeneration potential. Sea cucumbers of the order Dendrochirotida are capable

of a variety of regenerative processes, such as the regeneration of gut luminal epithelium from mesenteric mesothelium and the regeneration of pharyngeal bulb, a very complex and important anatomical structure at the anterior end of the body that contains pharynx and unifies the radial branches of the nervous, water-vascular, and hemal systems [79, 25]. In contrast, the sea cucumber *H. glaberrima*, which has been the main focus of molecular studies [80, 81, 82, 35, 51, 83, 84, 85, 86] lacks such regenerative events and therefore, this species is not suitable for studying diversity in the regeneration phenomena.

Prior to this work, there was no dendrochirotid genome available which hindered studies focusing on the molecular mechanisms enabling regeneration of complex structures and for this reason, I assembled the draft genome of the dendrochirotid sea cucumber *S. briareus* to study genes involved in regeneration of complex structures.

1.3 *S. briareus* (Lesueur, 1824)

S. briareus (Lesueur, 1824) (Echinodermata: Echinozoa: Holothuroidea: Dendrochirotida: Sclerodactylidae), originally *T. briareus*, is a species of highly regenerative sea cucumber found from Nova Scotia south along the US Atlantic coast to the Gulf of Texas [87]. Because of the distributed hair-like epidermal structures across their whole body, they are sometimes referred to as "hairy sea cucumbers".

Hairy sea cucumbers enjoy muddy localities primarily buried in eelgrass [87]. Their color varies from green or brown to nearly black. They have a barrel-shaped body of up to 120 mm long and a burrowing phenotype with tapered anterior and posterior extremities. Their numerous feet, feeding tentacles at the anterior end, thin body wall with ossicles (composed of tables of 60-80 µm diameter), and no buttons [88] are some of their other physical traits.

In terms of their habitat, hairy sea cucumbers are widespread, unprotected, and inhabit shallow waters. They are easily accessible to interested researchers or they

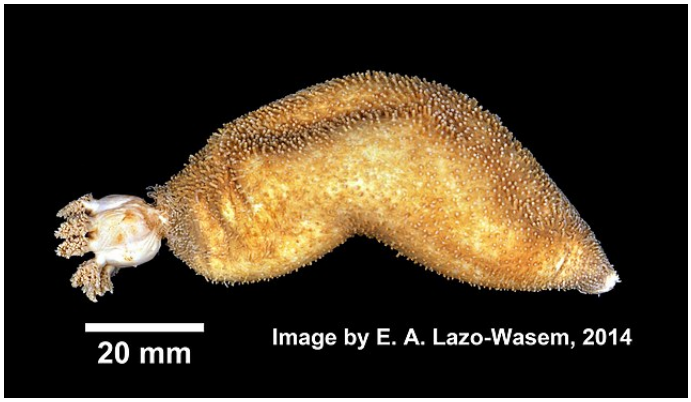

Figure 1.3: *S. briareus* (Lesueur, 1824). Source: GBIF | Global Biodiversity Information Facility [23, 24]

can be purchased from commercial suppliers. Hairy sea cucumbers are amenable to standard laboratory experimental procedures and they can be kept for weeks to months in a marine aquarium.

Unlike most holothuroids, hairy sea cucumbers (Dendrochirotids) have the capacity to autotomize and rapidly regenerate their anterior end of the body, including the viscera [89], central nervous system [90], body wall [91, 92], and muscles [93]. Their pattern of evisceration is also unique. While other sea cucumbers that autotomize only the middle portion of the digestive tube, dendrochirotids discard the bulk of their digestive system, except for the cloaca, along with the anterior end of their body (Figure 1.4), which includes the oral integument and the pharyngeal bulb. The pharyngeal bulb is a very complex and important anatomical structure which contains the anterior portion of the digestive tube, as well as the circumoral nerve ring and the ring canal of the water-vascular system [25]. The nerve ring joins the five radial nerve cords into an anatomically continuous central nervous system, whereas the ring-canal of the water-vascular system serves the same purpose in the hydrocoel. This anterior evisceration results in the loss of the whole front of the body, leaving the animal with only a cloacal stump. Thus, the animal provides an opportunity to study regeneration of two structures. First, the autotomized anterior regions of the digestive tube, which in embryogenesis are generated by the endoderm, regenerate

from the mesodermally-derived coelomic epithelium through direct transdifferentiation [94]. Second, the nerve ring regeneration that occurs through complex branching morphogenesis of the severed anterior tips of the radial nerve cords [35]. This unique pattern of regeneration, nerve ring regeneration and the regeneration of anterior regions of the digestive tube in dendrochirotid hairy sea cucumbers, is of high biological interest in regenerative biology. Therefore, the main goal of the present study was to help understand the molecular basis underlying extensive regenerative capacity in the hairy sea cucumber (*S. briareus*) by providing an outline of its genomic landscape.

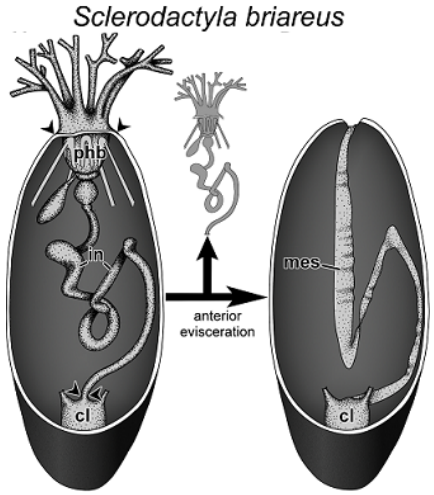

Figure 1.4: Anterior evisceration in *S. briareus*, including the intestine, mesentery, feeding tentacles, and the pharyngeal bulb. Abbreviations used are *cl* for cloaca, *in* for intestine, *mes* for mesentery (not shown in the diagram), and *phb* for the pharyngeal bulb. The arrowheads in the figure indicate the anterior planes of autotomy (evisceration). This figure was modified from Mashanov et al., 2011 [25])

1.4 Significance

We know that the potential of echinoderms in regeneration is the most substantial in the animal kingdom and not all species in the phylum are suitable for regeneration studies. For example, members of class Echinoidea possess the most limited regenerative capacities among echinoderms because of their body, which is planned on a hard shell-like test [74]. On the other hand, members of class Holothuroidea are highly re-

generative [73] and can regenerate their intestines and nerve tissues following injury or autotomy. In fact, sea cucumbers use evisceration to escape predators and it takes them only few weeks to regenerate their visceral organs completely [25, 95, 41]. This incredible regeneration capability makes sea cucumbers ideal for studying regeneration and evolutionary biology.

Among sea cucumbers, members of the order Dendrochirotida, are highly regenerative and display a unique pattern of evisceration. Unlike other sea cucumbers that autotomize only the middle portion of the digestive tube, dendrochirotids are capable of ejecting the bulk of their digestive system, including their intestinal epithelium and pharyngeal bulb, a very complex structure containing the anterior portion of the digestive tube, circumoral nerve ring, and the ring canal of the water-vascular system [79]. Therefore, using dendrochirotids in experiments will help us to understand deep spontaneous *in vivo* cell fate reprogramming and coordinated regeneration of multiple tissues (e.g., nervous, coelomic, and digestive epithelium) in complex structures, such as the pharyngeal bulb. These unique features of regeneration in dendrochirotid sea cucumbers make these taxa valuable models in regenerative biology.

With the advancement in high-throughput sequencing technologies, it is now possible to rapidly sequence the entire genome of any species. Therefore, to understand the regeneration of complex organs in echinoderms, we have sequenced and annotated the genome of *S. briareus*, a dendrochirotid species with incredible regeneration capacity. This advancement facilitates the increased data availability that allows us to identify genes involved in the regeneration and draw a cross-species comparison of gene functions, both within echinoderms and across other phyla.

Knowledge of gene signaling pathways and their interaction is important to understand the process of regeneration at the cellular and molecular level. Notch [96, 97, 98, 99] and Wnt [100, 101] signaling pathways are two examples of such pathways. These two pathways are highly conserved in the animal kingdom and

are mainly studied in the context of embryonic development because of which, little is known about their role in regeneration [102, 103, 104]. Therefore, my work (a draft genome of *S. briareus*), helps to establish the functional role of Notch and Wnt signaling pathways underpinning tissue regeneration in echinoderms.

To summarize, the draft genome of *S. briareus* serves as a valuable resource to conduct comparative studies and allow us to gain knowledge of gene signaling pathways and their interaction in regulating various cellular events, controlling different modes of regeneration within the the phylum.

1.5 Objectives

The overarching goal was to provide a high-definition genome resource of *S. briareus* that helps to obtain insights into molecular mechanisms underpinning visceral regeneration in sea cucumbers. Hence, in this dissertation, my goal was to generate a genomic resource of a highly regenerative species, *S. briareus* (Echinodermata: Holothuroidea: Dendrochirotida) to cater studies that are aimed to understand molecular and cellular events driving tissue regeneration in the phylum. Understanding tissue regeneration allows to capture the role of cell migration, de-differentiation, cell proliferation, and cell death in response to injury. This involves direct transdifferentiation of the existing coelomic epithelial cells into enterocytes in order to initiate regeneration post-injury [25, 41]. Hence, this work not only expands the available genomic resources on echinoderms but also enables us to facilitate an understanding of *in vivo* reprogramming of mature cells across germ layer boundaries to unfold the phenomenon involved in autotomization of anterior portion of the body, including the viscera and the pharyngeal bulb. Apart from illustrating the diversity in regeneration phenomenon, the genome of *S. briareus*, a dendrochirotid will provide an opportunity to explore an uncharted territory: the evolution of immune system of echinoderms against various pathogens in the marine environment. As a result, I have organized my dissertation into 4 chapters briefly discussed below.

In **Chapter 2**, I have released a new version of a database resource, EchinoDB, an open-source web application that contains the genomic and transcriptomic data of 42 unique echinoderm species in addition to the RNA-Seq data of brittle star *O. brevispinum*, phylogenomic data of *Xyloplax* species, and the genome assembly data of green sea urchin *L. variegatus* and hairy sea cucumber *S. briareus* to serve research communities by providing diverse and rich, functional genomics and transcriptomics data of different echinoderm species to facilitate pool of discoveries on Echinodermata, and the deuterostomes as a research model.

In **Chapter 3**, I have assembled and annotated a draft genome of the hairy sea cucumber, *S. briareus* to investigate genes essential for anterior regeneration to aid our understanding of the reparative mechanisms in echinoderms at the molecular level. In sea cucumbers, many events for tissue regeneration occur at the cellular level (cell migration, de-differentiation, proliferation and apoptosis) are known [73], but the underlying molecular mechanisms remain unclear [25]. Thus, the genome of *S.briareus* will serve as a useful reference genomic resource to identify gene regulatory networks supporting molecular regeneration.

In **Chapter 4**, I asked a couple of questions involving the genome assembly of *S. briareus*. How well-annotated my genome is? Can I retrieve the key components of Notch and Wnt signaling pathways, crucial for tissue regeneration process in echinoderms, in the draft genome of a hairy sea cucumber? As we know, sea cucumbers have the remarkable regeneration abilities and they can regrow their whole intestine after evisceration. In the regeneration process, many signaling pathways participate, including Notch [58] and Wnt pathways [101]. These genetic pathways are highly conserved across all multi-cellular animals. Therefore, in this chapter, we examined if we can retrieve the genes related to Notch and Wnt signaling pathways or not, to assess the usefulness of the genome.

In **Chapter 5**, I investigated *S. briareus* to untap its full potential not only in terms

of regeneration capacity but also in other areas, such as the discovery of novel viruses within echinoderms and the evolution of recombination activating genes (*RAG*), that provides support to the immune system by responding to a wide range of potential pathogens in the marine environment. The recent completion of the sea urchin genome, *S. purpuratus*, suggests that echinoderms have a sophisticated system of pathogen detection, mediated by a pair of recombination activating genes (*SpRAG1-Like* and *SpRAG2-Like*) [105, 106] to combat pathogens and other foreign substances, such as viruses present in the water. Therefore, in this chapter, the objective was to discover novel viruses infecting echinoderms by studying the transcriptomic data of *S. briareus* and to gain insights about immune-related genes (*SpRAG-Like*) in marine environment.

CHAPTER 2: EchinoDB v2.0

2.1 Overview

EchinoDB v2.0 is an open-source web-based application (`https://echinodb.uncc.edu`), designed to provide genomic, transcriptomic and amino acid sequence data on echinoderms. The objective of EchinoDB is to serve research communities by providing diverse and rich data for a wide diversity of echinoderm species. The previous version of EchinoDB was released in 2016 and consisted of amino acid sequence ortho-clusters (orthologous genes) from 42 echinoderm transcriptomes [3]. The new version was extended to incorporate new datasets that have been generated since the original release. These new datasets include RNA-Seq data for the brittle star *O. brevispinum* (Say, 1825) (Echinodermata: Ophiuroidea: Ophiacanthida: Ophiodermatidae) [58], genome assembly data of the green sea urchin *L. variegatus* (Lamarck, 1816) (Echinodermata: Echinoidea: Camarodonta: Toxopneustidae) [62], and phylogenomic data for *Xyloplax* sp. (Echinodermata: Asteroidea) [59, 60, 61]. The RNA-Seq data of the brittle star and the genome assembly data of the green sea urchin form the basis of two newly developed tools, OphiuroidDB [27] and EchinoidDB [28], respectively, integrated within the EchinoDB application (Figure 2.1).

To improve reliability and scale well with the increasing amount of data, the updated EchinoDB has been rewritten in R Shiny [107] and runs entirely in the cloud environment (AWS) [108]. R Shiny is highly extensible, easy to code and maintain, as compared to the previous implementation built using GO programming language in 2016. R Shiny supports faster development of user interfaces by providing a framework that requires no or little knowledge of scripting languages like HTML, CSS

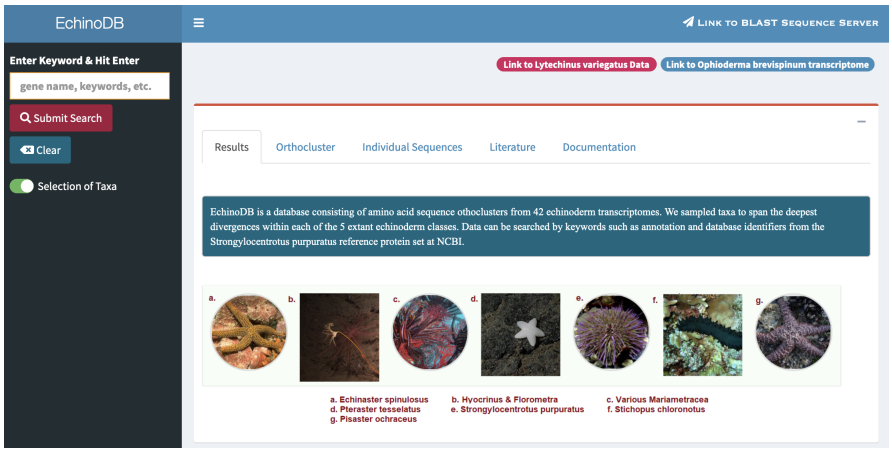

Figure 2.1: EchinoDB landing page [26], available at `https://echinodb.uncc.edu`. Users can search against all echinoderm classes, orders, and families or un-toggle to retrieve information for a particular taxon. Source: Mittal et al., 2022 [2]

or JavaScript. We had taken advantage of this feature to extend the application's capabilities to make new data (obtained from collaborations) easily available to the research community, for example, implementing the BLAST [5, 109, 33] search interface for the *Lytechinus* [28] and *Ophioderma* [27] sequences via Sequenceserver [110].

To demonstrate the practical utility of the new version of EchinoDB [26] and its associated resources - OphiuroidDB [27] and EchinoidDB [28] – we illustrated how EchinoDB is used in retrieving key components of the Notch and Wnt signaling pathways, that are crucial for tissue regeneration in echinoderms [111, 101, 112, 70, 113, 58]. In addition, we described the use of SequenceServer (BLAST tool) [33, 110, 114] integrated within EchinoDB to find the putative homologs of the skeleton matrix proteins [115, 116, 117, 118, 119] and tensilin (a protein that controls tensile strength of mutable collagenous tissues) [120, 121, 122, 123, 124, 125], previously reported in sea urchins (Echinodermata:Echinoidea) and sea cucumbers (Echinodermata: Holothuroidea).1

2.2 Materials and Methods

EchinoDB is re-factored in R Shiny and currently supports annotated transcriptomic data for 42 echinoderm species 2.1, functional transcriptomic data from a Notch pathway inhibition study in *O. brevispinum* [58], and protein sequences from a chromosome-level genome assembly of *L. variegatus* [62]. R Shiny is highly extensible, that is, code developed with R Shiny can be readily integrated with CSS themes, HTML widgets, and scripting languages (e.g. JavaScript). In addition, R Shiny is widely adopted and the code can be modified and tuned at later stages in the development cycle by many developers. EchinoDB v2.0 is hosted using the Nginx web server [126] in Amazon Web Services (AWS) [108]. AWS offers on-demand cloud computing services to build your own web-based applications independent of university information technology bureaus.

EchinoDB contains amino acid sequence clusters of orthologous genes, termed orthoclusters. These orthoclusters were generated by RNA-Seq profiling of adult tissues from 42 echinoderm specimens representing 24 orders and 37 families from all five extant classes [3]. The RNA-Seq data was assembled using Trinity [10] and translated into peptides using Transdecoder [127]. The *de novo* transcriptome assembly consisted of 1,198,706 amino acid sequences across 42 species. The data was clustered using OrthoMCL, an algorithm for grouping orthologous protein sequences based on sequence similarity [128]. The resulting orthoclusters database consisted of groups of 749,397 orthologous and paralogous transcripts. These orthoclusters were annotated through sequence similarity using the genome of purple sea urchin *S. purpuratus*, the best annotated echinoderm genome at the time of the origins of the project [129, 130]. Complete RNA-Seq analysis pipeline (from RNA sampling and isolation to sequencing, *de novo* transcriptome assembly, translation, orthoclustering and annotation) was described in [3]. These annotated orthoclusters now provide the basis for keyword searches in EchinoDB.

Table 2.1: Raw reads from the various echinoderm species that are available in NCBI's SRA [1] and Zenodo (doi: `https://doi.org/10.5281/zenodo.6985492`). Each line corresponds to transcriptome or gene expression data. Orthoclusters: number of orthoclusters. Sequences: number of amino acids or coding sequences. Length: sum of base pairs in all sequences. Source: Mittal et al., 2022 [2, 3]

Class: Order: Family	Species	Accesion	SRR	Title	Orthoclusters	Sequences	Length
Crinoidea: Comatulida: Zenometridae	*Psathyrometra fragilis*	PRJNA299480	SRR2846085	*Psathyrometra fragilis* Transcriptome or Gene expression	6651	9015	3.16E+07
Asteroidea: Velatida: Xyloplacidae	*Xyloplax* sp. Janetae (BJ2)	PRJNA299326	SRR2846120	*Xyloplax* sp. SIO-BIC E6809 Transcriptome or Gene expression	17993	24452	5.65E+07
Asteroidea: Spinulosida: Echinasteridae	*Echinaster spinulosus*	PRJNA300370	SRR2844624	*Echinaster spinulosus* Transcriptome or Gene expression	13844	18608	6.41E+07
Ophiuroidea: Ophiocomidea: Ophiocomidae	*Ophiocoma wendtii*	PRJNA299897	SRR2845427	*Ophiomastix wendtii* Transcriptome or Gene expression	3662	9783	8.82E+07
Ophiuroidea: Gnathophiuridea: Ophiotrichidae	*Ophiothrix spiculata*	PRJNA299898	SRR2845448	*Ophiothrix spiculata* Transcriptome or Gene expression	8118	18816	7.34E+07
Asteroidea: Velatida: Pterasteridae	*Pteraster tesselatus*	PRJNA299398	SRR2846094	*Pteraster tesselatus* Transcriptome or Gene expression	46531	51762	1.71E+08
Holothuroidea: Apodida: Synaptidae	*Synapta maculata*	PRJNA299890	SRR2846103	*Synapta maculata* Transcriptome or Gene expression	5309	11154	8.44E+07
Echinoidea: Echinoida: Strongylocentrotidae	*Strongylocentrotus purpuratus*	PRJNA299888	SRR2846101	*Strongylocentrotus purpuratus* Transcriptome or Gene expression	6885	11368	4.15E+07
Asteroidea: Forcipulatida: Asteriidae	*Pisaster ochraceus*	PRJNA299406	SRR2846074	*Pisaster ochraceus* Transcriptome or Gene expression	37807	43479	1.68E+08
Holothuroidea: Dendrochirotida: Psolidae	*Psolus* sp. (BJ11)	PRJNA299550	NA	*Psolus* sp. RWR-2015 Transcriptome or Gene expression	24634	35310	1.91E+08
Holothuroidea: Aspidochirotida: Stichopodidae	*Stichopus chloronotus*	PRJNA299896	SRR2846098	*Stichopus chloronotus* Transcriptome or Gene expression	17953	24854	1.09E+08
Crinoidea: Comatulida: Colobometridae	*Oligometra serripinna*	PRJNA299464	SRR2845419	*Oligometra serripinna* Transcriptome or Gene expression	55472	70278	2.11E+08
Crinoidea: Comatulida: Bourgueticrinidae	*Democrinus brevis*	PRJNA299465	SRR2844622	*Democrinus brevis* Transcriptome or Gene expression	6285	8287	4.72E+07
Asteroidea: Velatida: Korethrasteridae	*Peribolaster folliculatus* (BJ19)	PRJNA299409	SRR2845673	*Peribolaster folliculatus* Transcriptome or Gene expression	16927	20462	8.32E+07
Asteroidea: Paxillosida: Astropectinidae	*Psilaster charcoti*	PRJNA299410	SRR2846092	*Psilaster charcoti* Transcriptome or Gene expression	24055	28413	9.41E+07
Asteroidea: Forcipulatida: Labidiasteridae	*Labidiaster annulatus*	PRJNA299411	SRR2845003	*Labidiaster annulatus* Transcriptome or Gene expression	35615	40071	1.43E+08
Asteroidea: Velatida: Korethrasteridae	*Remaster gourdoni*	PRJNA299412	SRR2846097	*Remaster gourdoni* Transcriptome or Gene expression	18288	22056	8.21E+07
Crinoidea: Hyocrinida: Hyocrinidae	*Gephyrocrinus messingi*	PRJNA300546	SRR2859800	*Lamberticrinus messingi* Transcriptome or Gene expression	8950	12234	4.42E+07
Asteroidea: Paxillosida: Luidiidae	*Luidia clathrata*	PRJNA299414	SRR2845324	*Luidia clathrata* Transcriptome or Gene expression	36915	77487	9.42E+07
Asteroidea: Spinulosida: Echinasteridae	*Henricia leviuscula* A	PRJNA299415	SRR2844627	*Henricia leviuscula* Transcriptome or Gene expression	47492	76684	9.58E+07
Asteroidea: Paxillosida: Astropectinidae	*Astropecten duplicatus*	PRJNA299417	SRR2843238	*Astropecten duplicatus* Transcriptome or Gene expression	42051	73744	9.13E+07
Asteroidea: Valvatida: Poraniidae	*Glabraster antarctica* (BJ28)	PRJNA299418	SRR2844625	*Glabraster antarctica* Transcriptome or Gene expression	28408	54328	7.71E+07
Asteroidea: Valvatida: Asteropseidae	*Asteropsis carinifera*	PRJNA299419	SRR2843236	*Asteropsis carinifera* Transcriptome or Gene expression	25973	49607	6.51E+07
Asteroidea: Valvatida: Solasteridae	*Peribolaster folliculatus* (BJ30)	PRJNA299409	SRR2845673	*Peribolaster folliculatus* Transcriptome or Gene expression	22319	36551	5.25E+07
Asteroidea: Notomyotida: Benthopectinidae	*Cheiraster hirsutus*	PRJNA299420	SRR2844620	*Cheiraster hirsutus* Transcriptome or Gene expression	325	1271	6.85E+06
Asteroidea: Brisingida: Brisingidae	*Odinella nutrix*	PRJNA299463	SRR2845408	*Odinella nutrix* Transcriptome or Gene expression	312	1004	6.83E+06
Crinoidea: Comatulida: Ptilometridae	*Ptilometra australis*	PRJNA299466	SRR2846095	*Ptilometra australis* Transcriptome or Gene expression	33084	49470	7.31E+07
Crinoidea: Comatulida: Comasteridae	*Cenolia* new species	PRJNA299468	SRR2847917	*Cenolia trichoptera* Transcriptome or Gene expression	11658	18875	3.51E+07
Crinoidea: Comatulida: Antedonidae	*Isometra vivipara*	PRJNA299471	SRR2844835	*Isometra vivipara* Transcriptome or Gene expression	27204	43689	7.02E+07
Crinoidea: Comatulida: Antedonidae	*Phrixometra nutrix*	PRJNA299469	SRR2846073	*Phrixometra nutrix* Transcriptome or Gene expression	4923	12283	2.83E+07
Crinoidea: Comatulida: Antedonidae	*Promachocrinus kerguelensis*	PRJNA299478	SRR2846076	*Promachocrinus kerguelensis* Transcriptome or Gene expression	8011	12283	2.83E+07
Echinoidea: Arbacioida: Arbaciidae	*Arbacia punctulata*	PRJNA299547	SRR2843235	*Arbacia punctulata* Transcriptome or Gene expression	13324	33220	4.86E+07
Echinoidea: Cidaroida: Cidaridae	*Eucidaris tribuloides*	PRJNA299548	SRR2844624	*Eucidaris tribuloides* Transcriptome or Gene expression	6939	16512	2.97E+07
Echinoidea: Clypeasteroida: Dendrasteridae	*Dendraster excentricus*	PRJNA299549	SRR2844623	*Dendraster excentricus* Transcriptome or Gene expression	4619	12561	6.57E+07
Holothuroidea: Dendrochirotacea: Psolidae	*Psolus* sp. (BJ41)	PRJNA299550	NA	*Psolus* sp. RWR-2015 Transcriptome or Gene expression	16398	33062	7.32E+07
Holothuroidea: Aspidochirotida: Synallactidae	*Peniagone* sp. (BJ42)	PRJNA299551	NA	*Peniagone* sp. Transcriptome or Gene expression	12286	22457	5.25E+07
Holothuroidea: Dendrochirotacea: Cucumariidae	*Abyssocucumis* sp. (BJ43)	PRJNA299552	SRR2830762	*Abyssocucumis* albatrossi Transcriptome or Gene expression	12309	26171	5.47E+07
Holothuroidea: Aspidochirotida: Synallactidae	*Pseudostichopus* sp. (BJ44)	PRJNA299883	NA	*Pseudostichopus* sp. RWR-2015 Transcriptome or Gene expression	2464	5567	1.36E+07
Holothuroidea: Molpadida: Molpadidae	*Molpadia intermedia*	PRJNA299884	SRR2845419	*Molpadia granulata* Transcriptome or Gene expression	3793	6516	1.53E+07
Holothuroidea: Elasipodida: Laetmogonidae	*Pannychia moseleyi*	PRJNA299885	NA	*Pannychia moseleyi* Transcriptome or Gene expression	10124	20051	3.96E+07
Ophiuroidea: Euryalida: Gorgonocephalidae	*Astrophyton muricatum*	PRJNA299886	SRR2843239	*Astrophyton muricatum* Transcriptome or Gene expression	11730	26889	7.31E+07
Ophiuroidea: Ophiurida: Ophiodermatidae	*Ophioderma brevispinum*	PRJNA299887	SRR2845428	*Ophioderma brevispina* Transcriptome or Gene expression	11757	28450	6.52E+07

New data resources for ophiuroid and echinoid within the updated EchinoDB

We have added newly generated RNA-Seq data for *O. brevispinum* [58], a common brittle star found in shallow waters of the western Atlantic Ocean ranging from Canada to Venezuela. This resource can be found in EchinoDB under the name "OphiuroidDB". We have also added the "EchinoidDB" resource that contains the high-quality genome assembly data of *L. variegatus* [62], a sea urchin found in shallow waters throughout the western Atlantic Ocean ranging from the United States to Venezuela. The rationale for creating these two new data resources is that there has been a growing use of these two species in recent molecular studies in developmental and regenerative biology [58, 62, 70, 74, 131, 132, 133, 134].

OphiuroidDB

We have provided the brittle star, *O. brevispinum* [27] transcriptome dataset, translated, and annotated using BLASTX [109] against the NCBI collection of predicted

proteins of *S. purpuratus* [129] and protein models from UniProt's Swiss-Prot [135] and NCBI's RefSeq [4, 136]. The application can be accessed via "Link to *O. brevispinum* transcriptome" in EchinoDB and is referred to as "OphiuroidDB".

The transcriptome data of *O. brevispinum* were first used to characterize the downstream genes controlled by the Notch signaling pathway, which plays an important role in brittle star arm regeneration [58]. The raw sequencing reads of *O. brevispinum* transcriptome were submitted to the NCBI as a GEO dataset under the accession number GSE142391 [58, 137], and these sequences can now be also downloaded directly from OphiuroidDB. A total of 30,149 genes were identified, annotated, and included in the application.

EchinoidDB

EchinoidDB facilitates access to a recently published annotated high-quality chromosome level genome assembly of *L. variegatus* [62, 28]. The data (*Lvar*_3.0) include 27,232 nucleotide and protein sequences, which were annotated using BLASTP [109] against UniProt Swiss-Prot [135], *S. purpuratus* [130] and non-*S. purpuratus* RefSeq invertebrate protein models [4, 136]. These annotations can be downloaded from EchinoidDB.

2.3 Results

The updated version of EchinoDB includes two new major datasets: the RNA-Seq data of the brittle star *O. brevispinum* and the high-quality genomic assembly data of the green sea urchin *L. variegatus*. In addition, we enabled keyword searches for annotated data and installed an updated version of Sequenceserver to allow Basic Local Alignment Search Tool (BLAST) searches. The data are downloadable in FASTA format. The first version of EchinoDB appeared in 2016 and was implemented in GO on a local server. The new version has been updated using R Shiny to include new features and improvements in the application. Furthermore, EchinoDB now runs

entirely in the cloud for increased reliability and scaling.

2.4 Discussion

The web information systems (Table 1.2) that are currently available on echinoderms, such as Echinobase [76], HpBase [68], and SpBase [75] only allows for querying and exploration of the biological data mostly related to sea urchins. Hence, these databases are not suitable for capturing much of the diversity of the phylum Echinodermata. In contrast, EchinoDB v2.0 contains biological data of 42 different echinoderm species representing all five echinoderm classes, in addition to transcriptomic and genomic data for *O. brevispinum* and*L. variegatus*. Therefore, EchinoDB can serve as a valuable information resource to represent the diversity within the phylum and facilitate studies of regenerative phenomenon that varies widely among echinoderms.

To demonstrate the utility of EchinoDB v2.0 and associated resources, we used EchinoDB to retrieve genes associated with the Notch [138] and Wnt [139] signaling pathways. This is a biologically relevant example, as both these pathways are required for regeneration in echinoderms [58, 70]. Knowledge of the Notch and Wnt signaling pathways is important because they are highly conserved in the animal kingdom and regulate a variety of cellular processes, including proliferation, differentiation, fate specification, and cell death [96, 97, 98, 99]. Recent studies indicate that inhibiting the Notch signaling pathway prevented the brittle stars from fully regenerating their arms [58, 70]. Furthermore, Wnt signaling pathway is a major regulator of development throughout the animal kingdom. This pathway plays an important role in early regenerative events, including cell division, cell dedifferentiation and apoptosis that contribute to intestinal regeneration in holothurians [140, 141, 142, 143, 144, 145, 92]. For example, in sea cucumber *A. japonicus*, *Wnt6*, *Wnt7* (*Wnt* gene family), *Fzd7* (*Frizzled* gene family), and *Dvl* (*Dishevelled* gene family) are all significantly upregulated during the early stages of intestinal regeneration [111, 101]. Similarly, in *H.*

glaberrima, *Wnt9* is upregulated in early intestinal primordium [112]. Expression knockdown of *Wnt7* and *Dvl* significantly inhibits intestinal regrowth in sea cucumbers, implying that the canonical Wnt signaling is essential for visceral regeneration [101]. Figure 2.2 illustrates the function of EchinoDB v2.0 where the user can search for the keyword "Notch" in our web resources to locate Notch-related sequences in brittle stars and other echinoderms. For the keyword "notch", a total of 432 amino sequences distributed throughout 7 orthoclusters were found in EchinoDB (amino acid sequence orthoclusters of 42 echinoderm transcriptomes), 54 in OphiuroidDB (transcriptomic data for the brittle star *O. brevispinum*), and 38 in EchinoidDB (genomic and peptide sequences for the green sea urchin *L. variegatus*). Similarly, Figure 2.3 illustrates the step-by-step process of obtaining the corresponding sequences and metadata for "dishevelled" gene (*Dvl*) associated with the Wnt signaling pathway from our web resources. A total of 68 amino acid sequences found for "dishevelled" gene, grouped into a single orthocluster (XP_789156.3) in EchinoDB, four sequences were retrieved from OphiuroidDB and one from EchinoidDB.

Another use case involved expanding our understanding of the clade-specific biology using BLAST search. For example, biomineralization contributes to the development of the stereome-type endoskeleton unique to echinoderms. Biomineralization is defined as the biologically controlled formation of mineral deposits resulting in structures that function as support, protection, or feeding anatomy [116]. Among echinoderms, biomineralization is best characterized in sea urchins [115]. Hence, we asked if we can use our database to obtain an insight on whether the biomineralization mechanisms described in echinoids are unique to that class or shared across the phylum. To this end, we leveraged the SequenceServer (BLAST search) functionality available within EchinoDB to retrieve spicule matrix proteins involved in biomineral-

alization. The spicule matrix protein family consists of nine members, including the most extensively studied SpSM50 and SpSM30B/C [115]. We used SequenceServer (BLAST) integrated in EchinoDB [114] with a cutoff e-value of 1e-06 to compare the amino acid sequences of the echinoid spicule matrix proteins against EchinoDB (42 species), OphiuroidDB (*O. brevispinum*) and EchinoidDB (*L. variegatus*). Table 2.2 lists a number of echinoid and non-echinoid species represented in EchinoDB that had a BLAST match to each of those nine reference echinoid spicule matrix proteins. All nine proteins had a putative ortholog in at least one non-echinoid class, which suggests that the skeletogenesis mechanisms discovered in sea urchins might be also shared by other members of the phylum.

Another echinoderm-specific phenomenon is the capacity of the connective tissue structures to rapidly change their tensile strength under the control of the central nervous system [120, 124, 125]. A subset of neurosecretory cells is thought to release proteins that can either stiffen or soften the extracellular collagenous matrix. One such protein characterized so far is Tensilin, which upon its release from the neurosecretory cells, stiffens the mutable collagenous tissue [122, 124, 125, 146]. Only three sequences are known thus far, all of them from members of the class Holothuroidea, including sea cucumbers *C. frondosa* [122], *A. japonicus* [147], and *H. forskali* [123]. We therefore asked if tensilin, and thus tensilin-induced stiffening mechanisms, are unique to holothurians or are they represented in other classes of the phylum. To this end, we used the published protein and nucleotide sequences of tensilin as a query to perform BLASTP (for amino acid sequence) and BLASTX (for the nucleotide sequences) searches with an e-value threshold of 1e-06 [114]. This allowed us to find potential homologs in species from all five echinoderm classes represented in our database, EchinoDB. The BLAST results are summarized in table 2.3 which suggest that the tensilin protein, and thus the molecular mechanisms controlling the tensile strength of the mutable collagenous tissue, might be conserved across the phylum.

In conclusion, EchinoDB v2.0 with two complementary strategies: BLAST search and keyword search, proved useful in facilitating better understanding of genomic underpinnings of phylum-specific biological phenomena.

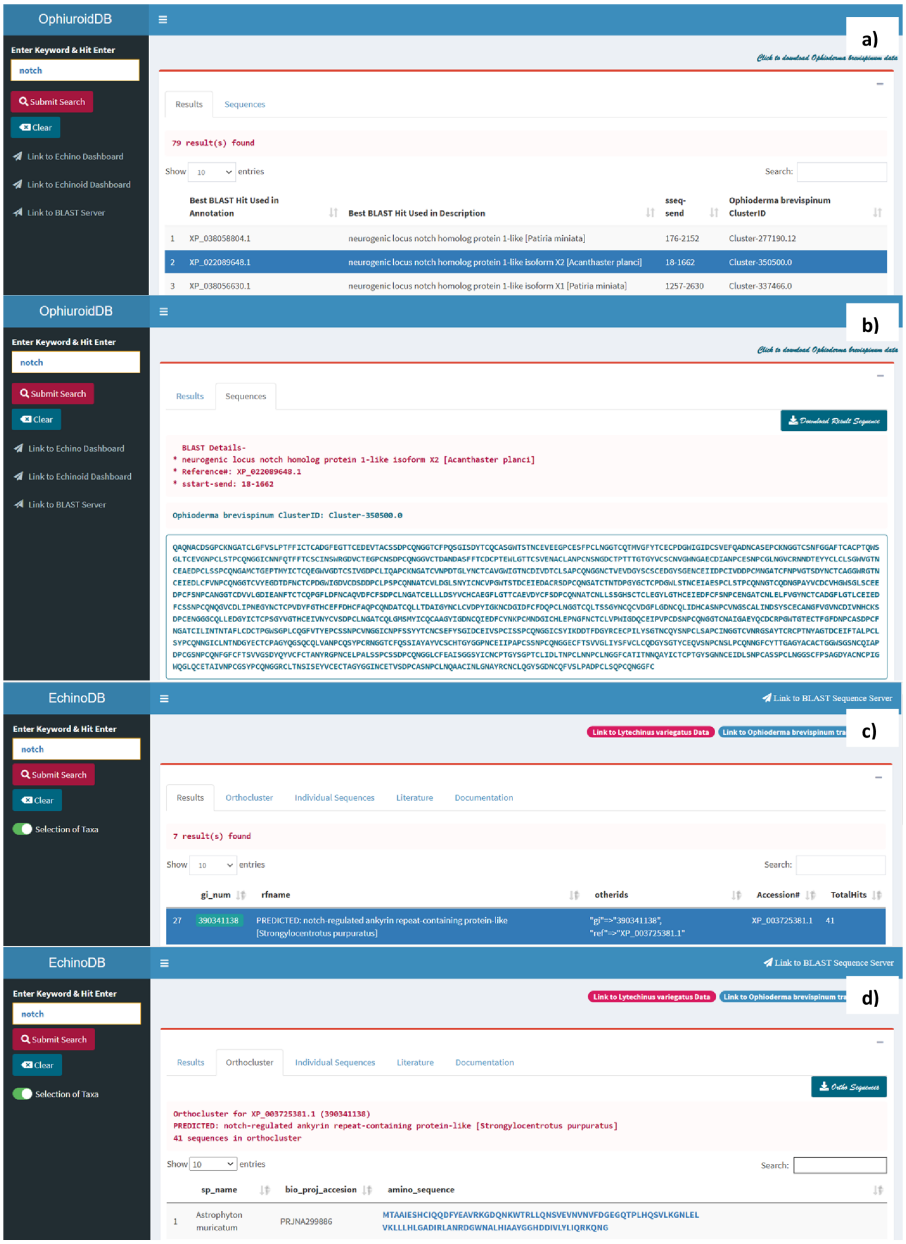

Figure 2.2: Usage example illustrating the search for Notch-related sequences in the brittle star *O. brevispinum* and other echinoderms. a) Screenshot of the OphiuroidDB main page [27]. The image shows the results after searching for the keyword "Notch" against the database of the brittle star *O. brevispinum*. The interface allows the selection of any record on the results page to view the sequence. b) Representative amino acid sequence from one selected Notch-related gene in OphiuroidDB. c) Results after searching for the keyword "Notch" in EchinoDB [26]. d) Amino acid sequence clusters of the selected orthologous record from the EchinoDB repository. Source: Mittal et al., 2022 [2]

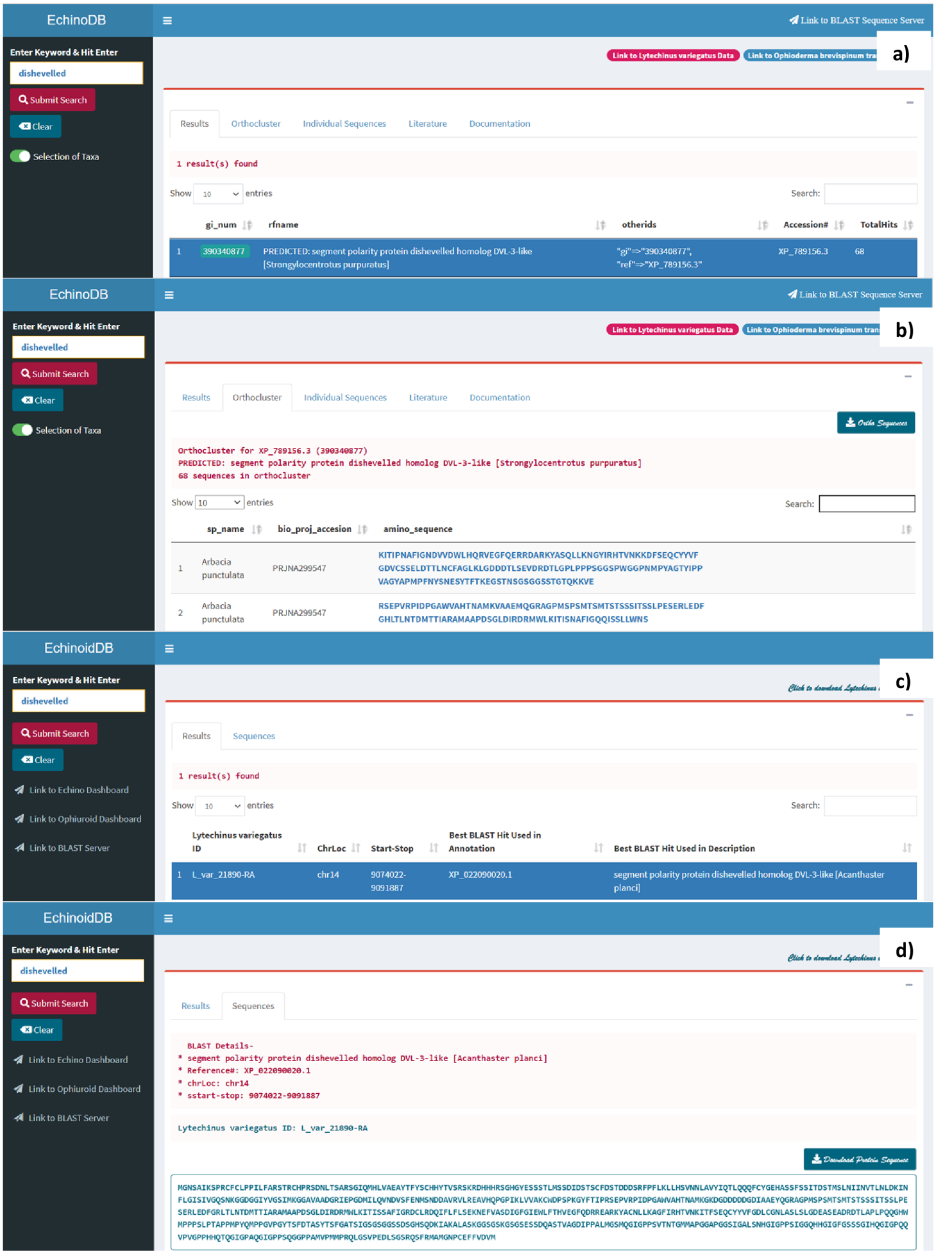

Figure 2.3: A use case illustrating the retrieval of the "dishevelled" gene (a principal component of the Wnt signaling pathway that governs several cellular processes, including cell proliferation, cell differentiation, and cell death) from EchinoDB and EchinoidDB. a) Results after searching for the keyword "dishevelled" in EchinoDB [26]. b) Amino acid sequence clusters of the selected orthologous record from EchinoDB. c) Screenshot of the EchinoidDB main page [28] that displays the results for the keyword "dishevelled". d) Example amino acid sequence of selected record in EchinoidDB. Source: Mittal et al., 2022 [2]

Table 2.2: Spicule matrix proteins retrieved from the database. Each line corresponds to individual proteins, for which we list accession numbers of corresponding reference sequences from the NCBI [4, 5], GenBank [6] or UniProt [7, 8] databases. The numerical values in the table represent the number of species in each class of the phylum that had a BLAST match to the reference sequence. Source: Mittal et al., 2022 [2]

DataBase (Accession)	**Protein**	**Description**	**Asteroidea**	**Ophiuroidea**	**Echinoidea**	**Holothuroidea**	**Crinoidea**
NCBI (NP_999775.2)	SpSM50	50 kDa spicule matrix protein precursor [*Strongylocentrotus purpuratus*]	2	1	4	0	0
NCBI (NP_999776.1)	SpSM37	spicule matrix protein SM37 precursor [*Strongylocentrotus purpuratus*]	0	1	3	6	1
NCBI (NP_999803.1)	SpSM32	spicule matrix protein SM32 precursor [*Strongylocentrotus purpuratus*]	2	2	4	3	1
UniProt (P28163/SM30_STRPU)	SpSM30B/C	30 kDa spicule matrix protein precursor [*Strongylocentrotus purpuratus*]	4	1	4	1	0
NCBI (NP_999804.1)	SpSM29	spicule matrix protein SM29 precursor [*Strongylocentrotus purpuratus*]	2	0	4	0	0
GenBank (CAA42179.1)	LSM34	spicule matrix 34 kd protein [*Lytechinus pictus*]	2	2	4	1	0
UniProt (Q25116)	HSM30	30 kDa spicule matrix protein [*Hemicentrotus pulcherrimus*]	1	1	4	0	0
UniProt (Q26264)	HSM41	41 kDa spicule matrix protein [*Hemicentrotus pulcherrimus*]	2	2	4	0	0
UniProt (Q95W96)	PM27	Primary mesenchyme-specific protein [*Heliocidaris erythrogramma*]	1	3	3	3	0

Table 2.3: Tensilin proteins. The first row corresponds to protein accession number from UniProt database [7, 8] whereas, second and third row depict nucleotide accession numbers from NCBI databases [4, 5] The numerical values in the table represent the number of species in each class of the phylum that had a BLAST match to the reference sequence. Source: Mittal et al., 2022 [2]

DataBase (Accession)	Description	Asteroidea	Ophiuroidea	Echinoidea	Holothuroidea	Crinoidea
UniProt (Q962H0)	Tensilin [*Cucumaria frondosa*]	8	1	1	9	3
NCBI (KR002726.1)	*Apostichopus japonicus* tensilin mRNA, complete cds	5	1	2	9	0
NCBI (KY609179.1)	*Holothuria forskali* tensilin mRNA, complete cds	9	1	2	9	0

2.5 Conclusion

EchinoDB serves a user base drawn from the fields of phylogenetics, developmental biology, genomics, physiology, neurobiology, and regeneration. As use cases, I illustrated the function of EchinoDB in retrieving components of signaling pathways involved in the tissue regeneration process of different echinoderms, including the emerging model species *O. brevispinum*. Moreover, I used EchinoDB to shed light on the conservation of the molecular components involved in two echinoderm-specific phenomena: spicule matrix proteins involved in the formation of stereom endoskeleton and the tensilin protein that contributes to the capacity of the connective tissues to quickly change its mechanical properties. The genes involved in the former had been previously studied in echinoids, while gene sequences involved in the latter had been previously described in holothuroids. Specifically, I asked (a) if the biomineralization-related proteins previously reported only in sea urchins are also present in other, non-echinoid, echinoderms and (b) if tensilin, the protein responsible for the control of stiffness of the mutable collagenous tissue, previously described in sea cucumbers, is conserved across the phylum. One of the focal points in the future is to extend the genomic, transcriptomic, and orthocluster contents of EchinoDB.

CHAPTER 3: Assembly and Annotation of *Sclerodactyla briareus*

3.1 Overview

Echinoderms show the most impressive capacities for adult organ and tissue regeneration among deuterostomes. In spite of the close phylogenetic relationship with chordates and the biomedical relevance, regeneration in echinoderms has received much less attention than in more distant groups, such as flatworms and cnidarians. In part, this is due to a relative scarcity of genomic resources, the situation that makes state-of-the-art molecular work difficult. Therefore, to understand regeneration of complex organs in echinoderms, I have sequenced and annotated the genome of *S. briareus*, a dendrochirotid species widely available along the US Atlantic Coast.

Like many other sea cucumber species, dendrochirotids are capable of visceral autotomy (evisceration), which can be experimentally induced by injecting of a few milliliters of distilled water or KCl solution into the main body cavity. However, pattern of evisceration in dendrochirotids is unique. Unlike other sea cucumbers that autotomize only the middle portion of the digestive tube, dendrochirotids discard the bulk of the digestive system, except for the cloaca, along with the anterior end of their body (Figure 3.1), which includes the oral integument and the pharyngeal bulb. These unique features of regeneration in dendrochirotid sea cucumbers make these taxa valuable models in regenerative biology.

Using dendrochirotids in experiments can provide important insights into evolution of the cellular and molecular events involved in the repair of multiple tissues (e.g., nervous, coelomic, and digestive epithelium) in complex structures, such as the pharyngeal bulb which are not represented in other sea cucumber species, including the most studied species *H. glaberrima*. Most sea cucumber species are incapable of re-

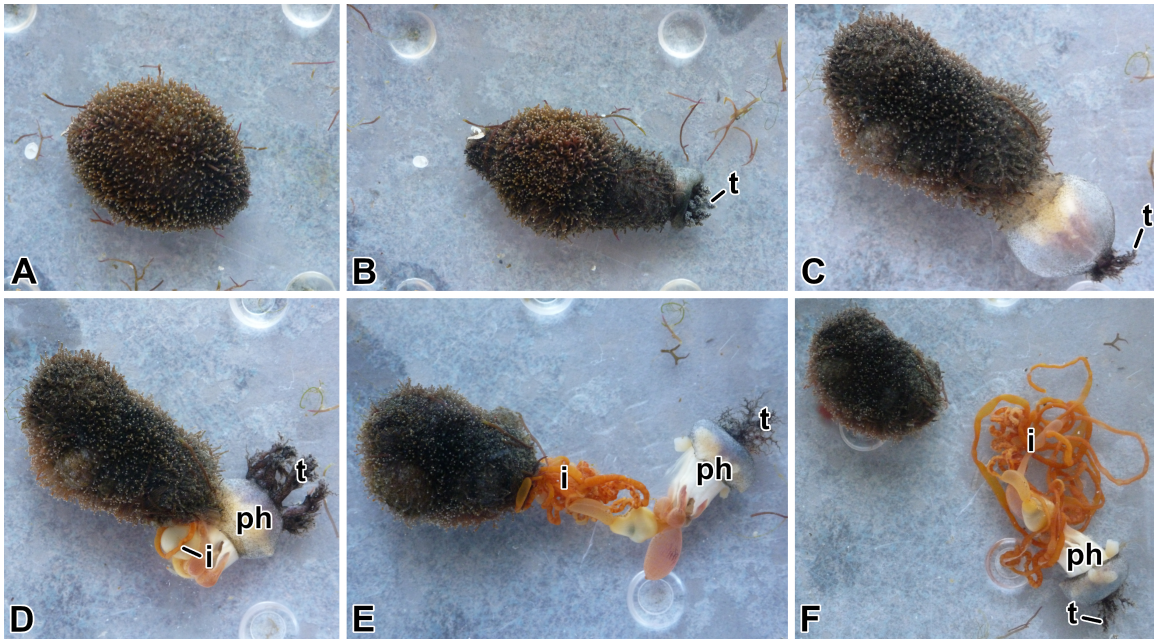

Figure 3.1: Time-lapse images of the evisceration (visceral autotomy) sequence in the sea cucumber *S. briareus*. *ph* – rudiment of the pharyngeal bulb; *t* – tentacle; *i* – intestine. Source: Mittal et al., in preparation

generation of the gut luminal epithelium from the mesentheric mesothelium through direct tissue transdifferentiation across the germ layer boundaries (mesodermal to endodermal) [79] and the regeneration of the pharyngeal bulb, a very complex and important anatomical structure at the anterior end of the body. Therefore, the purpose of this study was to create a genomic resource of a dendrochirotid, which will not only represent the diversity of regenerative events in Holuthuroids but also allow us to explore the *in vivo* cell fate reprogramming and coordinated regeneration of complex organs in echinoderms. Moreover, the availability of a dendrochirotid genome will help untap the full potential of *S. briareus* as a model organism.

3.2 Materials and Methods

Computational resources

All analytical processing was performed using computer clusters (Red Hat Enterprise Linux 7.5 with 64 CPUs and 512 GB to 1.5 TB of memory) and high-memory machines (Red Hat Enterprise Linux 7.5 with 16 CPUs and 512 GB to 4 TB of

memory) at the University of North Carolina at Charlotte.

Genomic DNA extraction

Adult individuals of hairy sea cucumber *S. briareus* were obtained from the Marine Biological Laboratory (Woods Hole, MA, USA). Specimens were received on April 13, 2016. Immediately after delivery, the package was opened and left overnight to slowly allow the seawater to warm up to room temperature. The animals were then kept in aquaria with aerated artificial seawater. A total of 100 mg of tissue was collected from the retractor muscles, longitudinal muscle bands, and interradial coelomic epithelium of the body wall of a single adult individual. The tissue samples were washed in filter-sterilized (0.2 μm) seawater, cut into small pieces with a sterile blade, and put into the lysis buffer. The high molecular weight nuclear genome DNA was then extracted using the Qiagen MagAttract HMW DNA kit according to the manufacturer's instructions with the following modifications to increase the yield and the molecular weight of the resulting DNA.

1. The speed of shaking in the MPS-1 machine was reduced from 1,400 rpm to 1,200 rpm.

2. After the first shake (in Buffer MB), the samples were allowed to sit at RT for additional 5 minutes before being placed into the magnetic rack. This was done to increase the amount of DNA bound to the magnetic beads to improve the final yield.

The concentration of the extracted DNA was assessed using the Qubit dsDNA Broad Range Kit (ThermoFisher) and the total yield was determined at approximately 50 μg. The DNA integrity in the sample was verified by agarose (0.6%) gel electrophoresis (at 2 V/cm, 4 hours).

Genome size estimate from the sequencing data

I used two complementary approaches: Feulgen densitometry (FD) assay [148, 149, 150] and calculation of $k$mer frequency within the sequence data using Jellyfish. For

the FD assay, soft uncalcified tissues (stomach and podia) from a single individual were finely minced with a razor blade, fixed in methanol:acetic acid (3:1) for ten minutes and pressed in a drop of 45% acetic acid onto a gelatin-coated slide. From the samples from air-dried to hydrolyzed in 5N HCl to brief washes in a 0.5% sodium metabisulfite solution to dehydration in an ethanol series, microscopic images were taken at a consistent light intensity in the green monochromatic channel to measure the absolute values of DNA mass per nucleus for the genome size estimation. The other approach used to estimate the genome size is the calculation of $k$mer frequency using Jellyfish [151] from the unassembled sequence data in GenomeScope [152].

DNA library preparation and sequencing

All DNA sequencing was performed at Eremid Genomic Services located in David H. Murdock Research Institute, Kannapolis, NC, USA. Two different technologies were leveraged to obtain short and long sequence reads from the high molecular weight genomic DNA (HMW gDNA) sample extracted as previously described.

For Illumina sequencing, subsamples of ≥ 8 μg of HMW gDNA were used to produce short sequence reads on the Illumina HiSeq2500 platform. Two paired-end libraries of different sizes ("short" and "long") were constructed using the standard protocol provided by Illumina. A "short" library was generated using TrueSeq DNA PCR-free library with the read length of 250 bp and the insert size of $\sim$ 450 bp. A "long" library was generated using Illumina Nextera Mate Pair Sample Preparation Kit with the read length of 250 bp and the insert size of $\sim$ 3 Kbp. The short and long libraries were combined onto their respective pools and sequenced in the Rapid Mode to produce 2 $\times$ 250 bp reads.

With the goal of improving the assembly, I took advantage of the Pacific Biosciences Single Molecule Real-Time (SMRT) platform to produce long sequence reads. The aim was to generate $\sim$50 $\times$ coverage to reduce and close the gaps, considering the estimated genome size of $\sim$1.7 Gbp. Four SMRTbell libraries were created with an

average fragment length of ≥60 Kbp using over 10 μg of HMW gDNA. These libraries were generated using the SMRTbell Template Prep Kit 1.0 following the PacBio >20 Kbp Template Preparation using BluePippin Size-Selection System (15-20 Kbp) for Sequel Systems procedures and checklist (catalog no. 100-286-000-07; Pacific Biosciences, USA) and sequenced in a cumulative total of 35 SMRTcells. Finally, reads from Illumina-based sequencing and PacBio sequencing were used to produce hybrid assembly.

DNA assembly

To facilitate the *de novo* assembly into contigs, several pre-processing steps such as removal of adapter sequences, poly-N sequences, and low-quality bases, were performed on the raw DNA sequence reads from *S. briareus*. The Illumina reads from the short insert library were processed with Trimmomatic v0.39 [153] to remove the adapters and leading and trailing bases with quality below 3. Trimmomatic also scanned each read with a 4-base sliding window cutting when the average quality per base dropped below 15. Reads shorter than 36 bp were also discarded. The long insert library Illumina reads were processed with NxTrim v0.3.0-alpha [154] using default parameters to separate reads into four different categories according to the adapter position: mate pairs, unidentified mate pairs, paired-end, and single-end sequence reads. The quality of the sequence data before and after each step was determined by FastQC v0.11.9 [29]. The tools used in the assembly are summarized in figure 3.2.

All sequence files produced by Trimmomatic v0.39 [153] and NxTrim v0.3.0-alpha [154] were evaluated with HTQC toolkit v0.90.8 [155] to produce the quality stats per file (using ht-stat) and perform the final read trimming and filtering (with ht-trim and ht-filter, respectively). FastUniq v1.1 [156] was used to remove duplicates introduced by PCR amplification from paired-end short reads.

All cleaned Illumina reads were used as input for the *de novo* assembly with ABySS

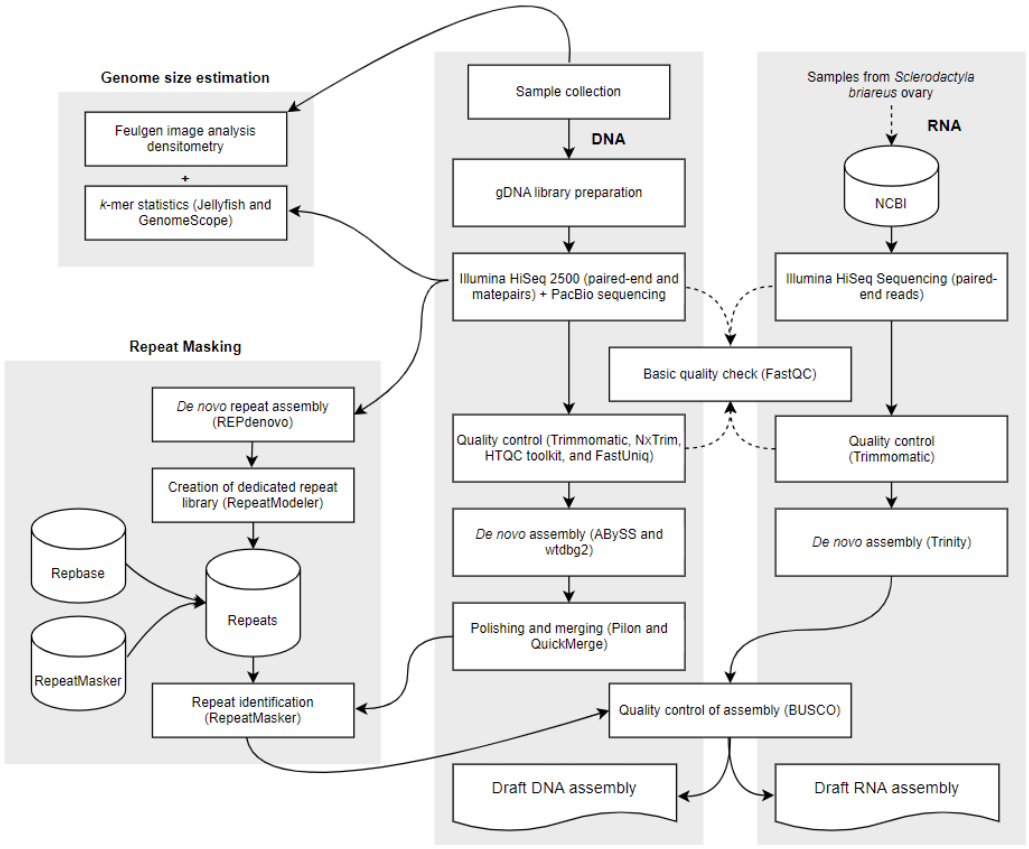

Figure 3.2: Schematic workflow of the *S. briareus de novo* DNA and RNA assembly. The main software tools used at each step of the workflow are shown in parenthesis. Grey boxes indicate four main components of the workflow. DNA: library preparation, quality control, high-throughput sequencing, and *de novo* assembly of gDNA and RNA. Genome size estimation: two different strategies used to estimate the haploid genome size. Repeat Masking: identification and categorization of repetitive DNA in the gDNA assembly. I used FastQC v0.11.9 [29] at different workflow steps to track the effect of quality control procedures on the sequence reads (see dashed arrowhead lines). Adapted from: Mashanov et al., 2022 for Mittal et al., in preparation

v2.11 [157] using *k*mers ranging from 23 to 61 with a step of 2 and 71 to 111 with a step of 10. The individual assemblies generated at each *k*mer value were ranked using several metrics, including the number of sequences, total assembly length, L50, and N50 [158]. I then polished the resulting best assembly in Pilon v1.2.3 [159] to improve base calling and detect sequence variation.

The long PacBio sequence reads were *de novo* assembled with wtdbg2 v2.5 [160] that uses a hybrid approach and implements the efficiency of fuzzy-Bruijn graph

(FBG), akin to De Bruijn graph (DBG), and flexibility of overlap–layout–consensus (OLC) to assemble whole-genome shotgun sequencing data. The contigs generated by wtdbg2 v2.5 [160] were merged with the above ABySS [157] assembly of Illumina reads using quickmerge v0.3 [161].

Assembly statistics were calculated using the "assembly-stats" [162] (developed at the Wellcome Sanger Institute) and "assemblathon stats.pl" [163] (developed at the UC Davis Bioinformatics Core) tools. The completeness of protein-coding gene representation in the assembled contigs of the draft genome assembly was assessed with BUSCO v4.0.6 [11] run in "protein mode" against the evolutionary conserved metazoan gene set (metazoa odb10, creation date: 2021-02-17, number of species: 65, number of BUSCOs: 954) and eukaryota gene set (eukaryota odb10, creation date: 2020-09-10, number of species: 70, number of BUSCOs: 255). BUSCO uses known genes to measure genome assembly and annotation completeness. Since *S. briareus* is not listed among the species available for AUGUSTUS [30] training, I tested three other species with BUSCO v4.0.6 [11]: *Homo sapiens*, *D. melanogaster*, and *S. purpuratus* in "genome" and "protein" mode. The BUSCO results are included in table 3.3.

Repetitive DNA analyses

Prior to the downstream analysis, repetitive DNA elements in the genome of *S. briareus* were separately *de novo* assembled using REPdenovo v0.0 [164]. The cleaned paired-end and single-end short sequence reads (resulted from the quality control step discussed above) using different *k*mer sizes ranging from 25 to 50 with a step size of 2 were used as an input for the assembly. The contigs assembled with REPdenovo v0.0 [164] were fed to RepeatModeler v1.0.11 [165] to build a library of repetitive genomic elements in the genome of *S. briareus*. The resulting hairy sea cucumber repeat library was then combined with repeat libraries from the 2018 version of Repbase [166, 167, 168, 169] and RepeatMasker v4.0.8 [170]. A final custom repeat library

containing unique entries were generated to screen the draft genome of *S. briareus* with RepeatMasker v4.0.8 [170] to identify interspersed repeats and low complexity DNA sequences in the genome.

De novo transcriptome assembly

The *de novo* transcriptome assembly was performed using the reads deposited to NCBI by Brown University (SRA# SRR1139189) in 2015 [171, 1] as input. These reads were generated using complete ovary of gravid females and served the transcriptome assembly.

I used Trimmomatic v0.39 [153] to remove adapters, low-quality bases, and calls marked as N. The quality of the data before and after filtering was evaluated using FastQC v0.11.9 [29]. This generated 29,414,825 HiSeq quality filtered and adapter trimmed reads. All cleaned RNA-Seq reads were pooled and assembled with Trinity v2.13.0 [9]. The assembly yielded 2,463,269 contigs with the N50 of 2,364 bp. The key statistics of the transcriptome assembly are listed in table 3.2.

To assess the quality of the assembled transcriptome, I performed a series of benchmark tests. First, to assess the representation of reads in the transcriptome, all cleaned RNA-Seq reads were aligned back to the assembled contigs using the STAR alignment tool v2.7.9a [172]. Nearly, half of the total reads (44.17%) mapped back to the assembly. Of these, 41.90% aligned uniquely as proper pairs and the remaining 2.27% mapped to multiple loci.

Second, to determine the completeness of the assembly in terms of protein-coding gene content, I ran BUSCO v4.0.6 [11, 12] in "transcriptome" and "protein" mode against the metazoan and eukaryota gene sets. The BUSCO results of the transcriptome assembly are included in table 3.4.

Gene prediction and annotation

A combined approach was used for the identification and annotation of protein-coding genes, including homology-based prediction, *de novo* prediction, and transcriptome based prediction. The gene prediction and annotation workflow is summarized in figure 3.3.

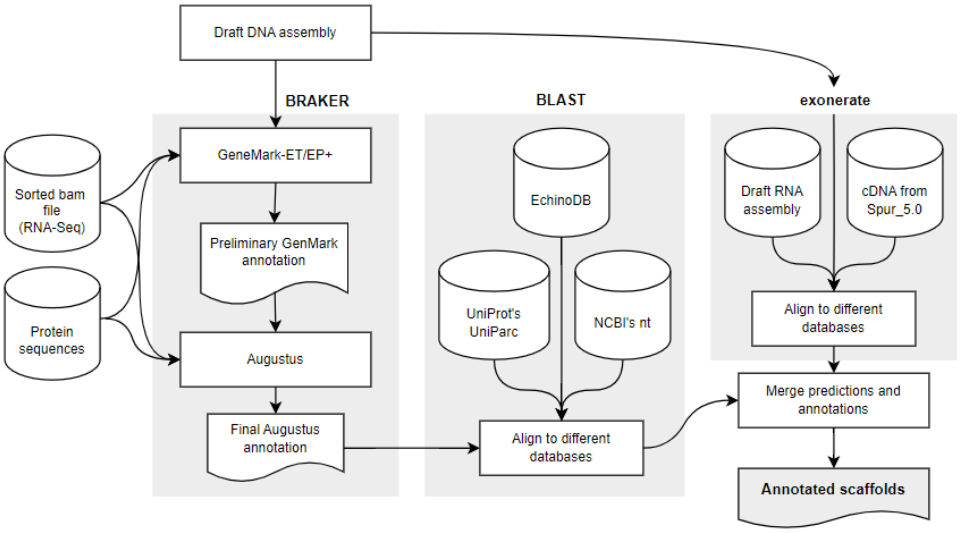

Figure 3.3: Schematic workflow of the procedures used for gene prediction and annotation of the *S. briareus* draft genome. Main steps (indicated by the gray boxes) are named according to the leading software used at each stage (BRAKER [30, 31, 32], BLAST [33], exonerate [15], and CD-HIT [16]). Adapted from: Mashanov et al., 2022 for Mittal et al., in preparation

First, I used BRAKER v2.1.5 [173, 32] to predict full gene structures and their annotations. The BRAKER pipeline used a sorted bam file from RNA-Seq reads and known amino acid sequences for three well-annotated echinoderm species (the sea urchins *S. purpuratus* and *L. variegatus*, and the sea cucumber *A. japonicus*) sourced from various databases such as Ensembl [54, 174, 175], NCBI [171, 4], UniProt [13, 7], and EchinoDB [26]. The sorted bam file was generated by aligning cleaned paired-end RNA-Seq reads against the draft genome of *S. briareus* using the STAR v2.7.9a alignment tool [172] and used as an input to BRAKER [173, 32] for gene prediction.

Next, I used exonerate v2.4.0 [15] to align the cDNA sequences from *S. purpuratus* (Spur_5.0) [175] and the newly assembled *S. briareus* transcriptome to the draft genome of the hairy sea cucumber.

Lastly, the predicted gene models were aligned with BLAST v.2.11.0+ [5] against the following databases each downloaded on April 11, 2022): the UniProt Archive (UniParc; `https://www.uniprot.org/help/uniparc`), the NCBI's non-redundant nucleotide database ("nt"; `https://ftp.ncbi.nlm.nih.gov/`), and the complete EchinoDB database of protein coding genes (`https://echinodb.uncc.edu/`) to annotate the predicted gene models.

3.3 Results

DNA sequencing

I employed three different genomic DNA library preparation and sequencing strategies. First, a PCR-free library preparation ("short" libraries) followed by sequencing on an Illumina HiSeq 2500 machine yielded 2 × 128,431,928 250 nt paired-end reads with an overall GC content of 41%. Second, mate-paired libraries ("long" libraries) with approximately 3 Kbp insert size were sequenced as above and resulted in 2 × 4,882,919 250 nt paired-end reads with an overall GC content of 45%. Third, a PacBio long-read sequencing on PacBio Sequel instrument generated approximately 12 million reads with a total yield of 88 billion bp (53.3 × coverage). The data are submitted under the GenBank's accession number "JAODTK000000000.1".

Nuclear DNA assembly statistics

The draft assembly of the *S. briareus* genome generated 14,969 contigs with the total assembly length of 1,383,787,993 bp (i.e., ∼1.38 GB) (Table 3.1). This value is close to the haploid genome size estimated by a densitometry assay (1.66 GB) [148, 149, 150] and by an analysis of the *k-mer* profile of the unassembled reads (1.11 GB) [151, 152], thus suggesting that the complexity of the genome is adequatly

captured in the assembly. Scaffolds range from 3,241 bp to 1,577,589 bp, with a mean scaffold size of 92,444 bp. The N50 scaffold length and L50 scaffold size are 223,505 bp and 1,793, respectively. The scaffold nucleotide content is 28.53%, 21.45%, 21.43%, and 28.59% for A, C, G, and T, respectively.

Table 3.1: Summary metrics of the *S. briareus* genome assembly. Source: Mittal et al., in preparation

Metrics	**Quantification**
Total assembly length (bp)	1,383,787,993
Number of scaffolds	14,969
Shortest scaffold	3,241
Longest scaffold	1,577,589
Mean scaffold length	92,444
N50 scaffold length	223,505
L50 (scaffolds)	1,793
Assembly GC content	42.88%
Assembly coverage or depth	40X
Repetitive DNA (bp)	564,595,743

An independent *de novo* assembly of repetitive DNA elements with REPdenovo v0.0 [164] resulted in 14,969 individual sequences with a total length of 1,383,820,095 bp. The average length and N50 of the repetitive DNA segments assembled this way were 1,468 bp and 2,755 bp, respectively. These sequences were then used to aid in the repeat identification and masking in the draft genome [165, 170]. A total of 564,595,743 bp (40.80%) of the draft assembly were classified as repetitive DNA and masked. Most DNA repeats (36.80% of the repetitive DNA sequence length) were classified as interspersed elements. However, a significant number of repeats (28.34% of the repetitive DNA) were marked as unclassified. The most common (6.37% of the repetitive DNA) transposable element (TE) in the classification of repeats corresponds to long interspersed nuclear elements (LINEs). Repetitive DNA elements, and long terminal repeats (LTRs) amounted to 1.48% and 0.61% of the total sequence length, respectively.

The gene annotation pipeline consisting of BRAKER v2.1.5 [32] and exonerate v2.4.0 [15] predicted 159,717 exons in 43,845 genes, each corresponding to a different

transcript in a total of 40,807 nuclear DNA scaffolds. According to position information, at least 864 of those 43,845 genes could represent gene isoforms. These putative isoforms are found in 317 contigs. The degree of fragmentation of this draft genome and limitations in resolving gene isoforms may partially account for the overestimation of the number of protein-coding genes in *S. briareus*. These caveats should be addressed in future efforts to increase the contiguity of this genomic resource.

RNA assembly statistics

In order to facilitate the annotation efforts of the genomic scaffolds, I took advantage of the *S. briareus* whole ovary RNA-Seq reads (accession# SRR1139189) available at the National Center of Biotechnology Information (NCBI) [1]. The reads were quality filtered and adapter trimmed (using trimmomatic [153]) to improve the overall quality followed by *de novo* assembly with Trinity [9] to reconstruct the full-length transcripts from the RNA-Seq data. The assembly yielded 2,463,269 contigs with a total length of approximately 0.093 GB. The key assembly metrics are shown in table 3.2.

Table 3.2: Key metrics of the *S. briareus* transcriptome assembly using Trinity [9, 10]. Source: Mittal et al., in preparation

Metric	**Value**
Total number of bases assembled	93,040,756
Total number of reads assembled	29,414,825
Total number of assembled contigs	2,463,269
Total number of Trinity “genes"	61,396
Average contig length (bp)	1,113
Median contig length (bp)	505 nt
N50 (bp)	2,364 nt
Assembly coverage	45X
GC content for the complete assembly (%)	44.94
Overall read alignment rate (%)	95.21

Gene content completeness

The gene space completeness of the draft genome assembly (at the scaffold level) and that of the transcriptome assembly was evaluated using the Benchmarking Uni-

versal Single-Copy Ortholog (BUSCO) v4.0.6 tool [11, 12]. BUSCO assesses the representation of single-copy marker genes in newly generated genomic and transcriptomic datasets as a proxy for the whole-genome gene representation. The results of the draft genome and transcriptome assembly are summarized in table 3.3 and table 3.4, respectively.

Table 3.3: Summary of BUSCO v4.0.6 [11, 12] results for the scaffolds of the draft genome assembly. The database column names each odb10 BUSCO database used. The species column indicates the Augustus training parameter. Ditto marks (") indicate values identical to the cell above. The names "Human", "Fly" and "Spur" correspond to *H. sapiens*, *D. melanogaster*, and *S. purpuratus*, respectively. The e-value used for the analysis was 1.00E-03. Source: Mittal et al., in preparation

Database	Species	Complete (All)	Single-copy	Duplicated	Fragmented	Missing	N
Metazoa	Human	27.6%	27.6%	0.0%	7.8%	64.6%	954
"	Fly	28.1%	28.1%	0.0%	8.0%	63.9%	954
"	Spur	21.9%	21.9%	0.0%	10.8%	67.3%	954
Eukaryota	Human	22.4%	22.4%	0.0%	12.9%	64.7%	255
"	Fly	23.9%	23.9%	0.0%	12.2%	63.9%	255
"	Spur	20.0%	20.0%	0.0%	8.2%	71.8%	255

Table 3.4: Summary of BUSCO v4.0.6 [11, 12] results for the scaffolds of the transcriptome assembly. The database column names each odb10 BUSCO database used. The species column indicates the Augustus training parameter. Ditto marks (") indicate values identical to the cell above. The names "Human", "Fly" and "Spur" correspond to *H. sapiens*, *D. melanogaster*, and *S. purpuratus*, respectively. The e-value used for the analysis was 1.00E-03. Source: Mittal et al., in preparation

Database	Species	Complete (All)	Single-copy	Duplicated	Fragmented	Missing	N
Metazoa	Human	95.0%	57.1%	37.9%	2.4%	2.6%	954
"	Fly	95.0%	57.1%	37.9%	2.4%	2.6%	954
"	Spur	95.0%	57.1%	37.9%	2.4%	2.6%	954
Eukaryota	Human	97.2%	53.3%	43.9%	1.2%	1.6%	255
"	Fly	97.2%	53.3%	43.9%	1.2%	1.6%	255
"	Spur	97.2%	53.3%	43.9%	1.2%	1.6%	255

3.4 Discussion

I have assembled a draft genome, including non-coding and regulatory regions. Despite the low BUSCO score and fragmentation of the genome assembly, the assembly can be served as a valuable resource and allow the analysis of protein-coding genes which are relevant to echinoderm regeneration for example, components related to Notch and Wnt signaling pathways, including ligands, transcription factors, recep-

tors, regulators, and target genes but how? I have talked about this in detail in Chapter 4 by identifying the key components of Notch and Wnt signaling pathways, crucial for regeneration in echinoderms.

3.5 Conclusion

Here, I present the first draft nuclear genome of the sea cucumber *S. briareus* (Lesueur, 1824) (Echinodermata: Echinozoa: Holothuroidea: Dendrochirotida: Sclerodactylidae), an emerging model for regenerative studies (e.g., [176, 25, 177, 178]). The draft nuclear DNA assembly has yielded 14,969 contigs summing up to ∼1.38 Gb, corresponding to ∼83% of the expected haploid genome size independently determined by the densitometry analysis [148, 149, 150]. Despite the high degree of fragmentation of the assembly, which is partially caused by a high frequency of the repetitive DNA elements (∼40.8% of the assembly), the current draft genome of *S. briareus* will be fundamental towards chromosome-level assembly of the entire genome of *S. briareus* and facilitate state-of-the-art functional genomic studies in this sea cucumber species in the future.

CHAPTER 4: Case Study - Notch and Wnt Signaling Pathways

4.1 Overview

Regeneration occurs widely in animals. However, their regenerative capacities vary among species and this ability plummets as development proceeds in most vertebrates. Echinoderms are exceptional and display spectacular capacities to fully regenerate certain organs or appendages after injury as adults. This remarkable regeneration requires proper regulation of key signaling pathways, including Notch and Wnt. In this chapter, I will discuss the importance of Notch and Wnt signals and the cross-talking between these signals, essential for organs regeneration, including the development of small intestines.

The Notch and Wnt signaling pathways are highly conserved across all multicellular animals and function as major regulators of morphogenesis throughout the animal kingdom [96, 97, 98, 99, 100, 101]. These pathways coordinate many cellular events, including proliferation, differentiation, cell-fate specification, and cell death [138, 179, 180, 97, 98, 181, 99, 182, 145, 112].

The Wnt signaling pathway has long been recognized as a critical component of tissue regeneration in echinoderms. A study showed that Wnt signaling regulates cell proliferation, dedifferentiation, and apoptosis–all essential processes for visceral regeneration in sea cucumbers [145, 140, 112, 92, 141]. The expression knockdown of several genes in this pathway can inhibit intestinal growth in holothurians [142, 143, 144, 83, 113, 183]. For example, *Wnt6*, *Wnt7* (*Wnt* gene family), *Fz7* (*Frizzled* gene family), and *Dvl* (*Dishevelled* gene family) are four positively selected genes that are all significantly upregulated during the early stages of regeneration and contribute to the early activation of Wnt signaling that initiates intestinal regeneration in sea

cucumbers including, *A. japonicus* [111, 101]. Similarly, in *H. glaberrima*, *Wnt9* is upregulated in early intestinal primordium [112]. Expression knockdown of *Wnt7* and *Dvl* genes significantly inhibit intestinal extension, implying that this pathway is essential for intestinal regeneration [101].

More recently, Mashanov et al. [58] demonstrated that Notch is involved in regulating arm regeneration in the brittle star *Ophioderma brevispinum*. In *O. brevispinum*, Notch acted via a diverse array of downstream genes responsible for the control of cell division, cell death, cell migration, extracellular matrix remodeling, and innate immune response [58].

Not only are these pathways significant, but their interaction is also crucial for the development of several organs, which is covered in detail below.

4.1.0.1 Interaction of Notch and Wnt genes

Studies indicate that the Notch and Wnt pathways are necessary to maintain the proliferative activity in the crypt regions of the developing small intestine [184]. This action of proliferation of crypt cells depend on the integration of Notch and *Wnt/Tcf4* [185, 184]. Blocking Wnt signals through the knockout of *Tcf* gene abolishes cell proliferation in the intestine that results in lethality at birth or the complete absence of proliferative cells in crypt regions, confirming that Notch requires *Wnt/Tcf4* to restore proliferation in the developing intestine [186].

Another study emphasizes on the interaction of Notch and Wnt pathway, important for head regeneration, head organization, and tentacle patterning in hydra [187]. Inhibition of Notch signaling in hydra can lead to the formation of irregular head structures characterized by excess tentacle tissues, which further impairs the formation of *Wnt/beta-catenin* dependent head organizer and, thus, affects its function [188].

Further, the role of *APC* (*adenomatous polyposis coli*) and Notch in colorectal cancers (CRC) is established. *Wnt/beta-catenin* signaling is essential for intestinal

homeostasis and aberrantly activated in most CRCs due to the mutation in *APC*, a tumor suppressor gene [189, 190]. Mutations in *APC* region result in a loss of multiple *beta-catenin* binding sites, leading to the disruption of cellular homeostasis, including cell proliferation and differentiation, apoptosis, and inflammation-associated cancer [191]. Surprisingly, Notch a strong promoter of tumor initiation [184] plays a suppressive role in the expression of *Wnt/beta-catenin* target genes in established tumors and its activation converts high-grade adenoma into low-grade adenoma, thus suppressing intestinal tumors [192].

These features make Notch and Wnt-related genes candidates to demonstrate the utility of the genome of *S. briareus*. In this chapter (Chapter 4), I detailed a case study in which I assessed the genomic representation of the main components of the Notch and Wnt signaling pathways as an indicator of the usefulness of the draft genome of *S. briareus* in regenerate studies. Our assumption was that the presence of well-categorized Wnt and Notch-related genes is essential for any genomic resource to be useful in studying regeneration in echinoderms.

4.2 Materials and Methods

To assess the practical utility of the newly assembled *S. briareus* draft genome, I searched the assembled contigs for the components of Notch and Wnt signaling pathways listed in Tables. 4.1 and 4.2, respectively.

Gene annotation used a combination of homology-based strategies as described below. Reference databases for homology-based annotation were built from amino acid sequences retrieved from different databases, including UniProt (`https://www.uniprot.org/`) [7], NCBI (`https://ncbi.nlm.nih.gov/`) [5], Echinobase (`www.echinobase.org`) [14], and EchinoDB (`https://echinodb.uncc.edu/`) [3, 2]. Reference databases were processed to remove short (< 31 aa), low-quality ($> 5\%$ ambiguity), and duplicated sequences using a custom Python script.

Different programs were used to annotate the target genes in the genomic scaffolds

of *S. briareus*, including BLAST v2.11.0+ [5, 33], exonerate v2.4.0 [15], CD-HIT v4.8.1 [16], and CDD v3.20 [193, 194, 17, 195, 196], as described below are summarized in figure 4.1.

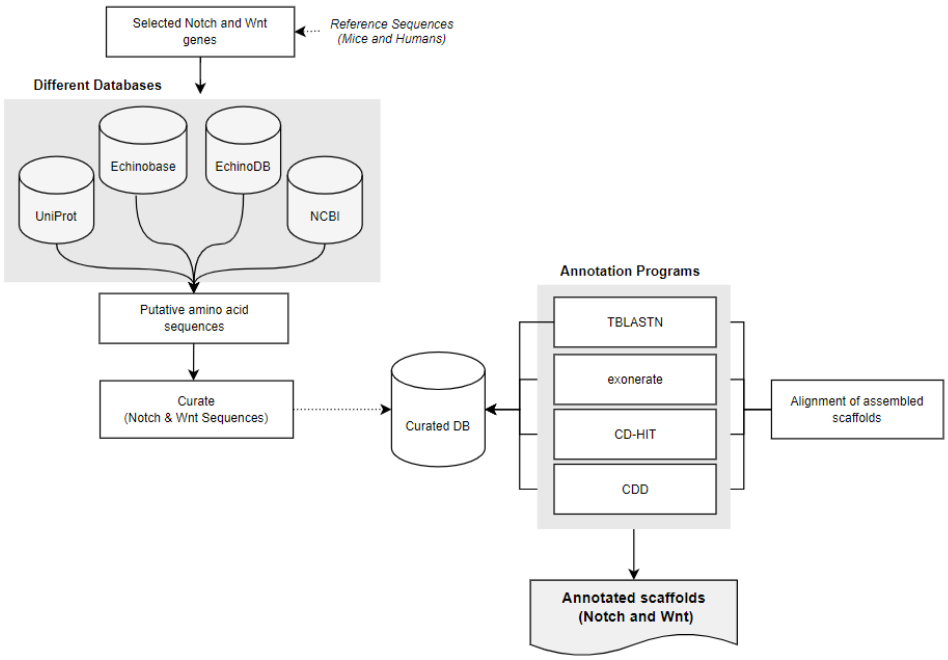

Figure 4.1: Annotation of selected genes from Notch and canonical Wnt signaling pathways in the draft genome of *S. briareus*. The software and public databases used for annotation purposes are indicated by gray boxes. Source: Mittal et al., in preparation

The amino acid sequences from the reference repository were aligned to target exons from BRAKER annotation [173, 32] using TBLASTN [5] with an E-value, bit score, and percentage identity cutoff thresholds of 1.0E-5, 30.0, and 23%, respectively. Simultaneously, I aligned the amino acid query sequences to all assembled contigs using exonerate v2.4.0 [15] with default settings. I also used CD-HIT v4.8.1 [16] to cluster assembled scaffolds (based on the similarity) and compared them with the amino acid query sequences using a sequence identity threshold of 0.5.

Genes identified using the strategies listed above were selected for analyses in NCBI's Conserved Domain Search (`www.ncbi.nlm.nih.gov/Structure/cdd`)[193, 194,

17, 195, 196]. The goal of this analysis was not to eliminate candidate sequences that had protein domains not found in the reference database of the expected gene associated with them. Instead, the objective was to retain genes processing the anticipated diagnostic domains and discard sequences containing diagnostic domains of genes not present in the reference database. The conserved domains were searched against CDD v3.20 database, with an E-value threshold of 1e-05 and composition-based statistics adjustment. The domains from query sequences and resulting sequences for Notch and Wnt pathways were manually inspected for the presence or absence of curated functional protein domains.

4.3 Results

To illustrate the usefulness of the genome of *S. briareus*, I queried the assembled scaffolds with 29 components of the Notch signaling pathway (Figure 4.2, Table 4.1) and 25 canonical genes of the Wnt signaling pathway (Figure 4.3, Table 4.2). Except for *Mastermind*, *Mesp2*, *NCoR*, and *Presenilin 1*, all genes related to Notch pathway were successfully retrieved using BLAST v2.11.0+ search [33]. With the exception of *Mesp2*, I retrieved the same putative coding regions from exonerate v2.4.0 alignments [15] for all selected Notch genes. However, I was able to recover only 22 of 29 Notch genes with CD-HIT v4.8.1 [16]. The genes that did not yield any results, include: *CIR1*, *Furin*, *Mesp2*, *MINT*, *NCoR*, *Notchless* (*NLE*), and *POFUT1*.

Lastly, I screened amino acids reference sequences and the resulting sequences from BLAST, exonerate, and CD-HIT for the presence of expected conserved domain footprints (using CDD v3.20 [17]), resulting in 20 of 29 genes in the draft genome of *S. briareus*. The results are summarized in table 4.1 and the genes are shown in figure 4.2.

Similarly, for Wnt signaling pathway, I was able to identify all selected Wnt genes

Table 4.1: Select components of the Notch signaling pathway identified in the draft genome of *S. briareus* using reference sequences from UniProt [13], Echinobase [14], EchinoDB [2], and NCBI [4]. For each gene, we list its name, the known function in the pathway, and whether or not the gene was recovered from the draft genome with independent BLAST [5] exonerate [15] and CD-HIT [16]. In addition, we also indicate if we could identify conserved protein domains [17] in the predicted protein sequences. Source: Mittal et al., in preparation

Name	Role in the Pathway	BLAST	Exonerate	CD-HIT	Conserved domains	References
ACBD3	Acyl-CoA-Binding Domain-Containing Protein 3 Activator of Numb	Yes	Yes	Yes	No	[197, 198]
ADAM 10/17	A disintegrin and metalloproteinase with thrombospondin motifs	Yes	Yes	Yes	Yes	[138]
CIR1	Co-repressor interacting with RBP-J	Yes	Yes	No	Yes	[138, 180]
Delta/Serrate (Jagged)	Ligand of the notch receptor	Yes	Yes	Yes	No	[138, 180]
Deltex	E3 ubiquitin-protein ligase/ DTX1. Context-dependent positive or negative regulator. Antagonizes Nedd4	Yes	Yes	Yes	Yes	[96, 199]
Fringe	β−1,3-N-acetylglucosaminyltransferase radical fringe/ Lfng(lunatic) or Rfng(Radical). Post-translational maturation of Notch	Yes	Yes	Yes	Yes	[138, 180]
Furin	Paired basic amino acid cleaving enzyme. Receptor proteolysis	Yes	Yes	No	Yes	[200]
HDAC1	Histone deacetylase 1. Co-repressor of RBP-J	Yes	Yes	Yes	Yes	[96]
HES	HES-4-like. Canonical target gene	Yes	Yes	Yes	Yes	[197, 201, 202]
HEY1	Hairy/enhancer-of-split related with YRPW motif protein 1. Canonical target gene	Yes	Yes	Yes	Yes	[197, 201, 202]
LNX2	Ligand of Numb Protein 2. Negative regulator of Numb	Yes	Yes	Yes	Yes	[203]
Mastermind	Mastermind-like protein. Co-activator of RBP-J	No	Yes	Yes	No	[138, 180, 204]
Mesp2	Mesoderm posterior bHLH transcription factor 2. Activates Fringe, induces degradation of Mastermind	No	No	No	No	[204]
Mindbomb	E3 ubiquitin-protein ligase	Yes	Yes	Yes	Yes	[205]
MINT	SHARP/ spen family transcriptional repressor/ Mint/Sharp/SPEN, NCoR/SMRT, KyoT2. Co-repressor of RBP-J	Yes	Yes	No	Yes	[206]
NACK	Notch Activation Complex Kinase. Co-activator of RBP-J	Yes	Yes	Yes	Yes	[207, 208]
NAK	Numb-associated kinase. Positive regulator of the Notch pathway	Yes	Yes	Yes	No	[197]
NCoR	Nuclear receptor corepressor. Co-repressor of RBP-J	No	Yes	No	No	[206]
Nedd4	Neuronal precursor cell-Expressed. Targets Notch and Deltex for degradation	Yes	Yes	Yes	Yes	[199]
Neuralized	Ubiquitination of Jagged	Yes	Yes	Yes	Yes	[58, 209]
Nicastrin	Receptor proteolysis	Yes	Yes	Yes	Yes	[138, 197]
Notch	Neurogenic locus notch	Yes	Yes	Yes	No	[138, 204]
Notchless	Context-dependent positive or negative regulator	Yes	Yes	No	Yes	[210]
Numb	Negative regulator of the Notch pathway	Yes	Yes	Yes	Yes	[197]
p300	CREB-binding protein. Co-activator of RBP-J	Yes	Yes	Yes	Yes	[207, 211]
POFUT1	Protein O-fucosyltransferase 1. Post-translational maturation of Notch	Yes	Yes	No	Yes	[212]
Poglut	Protein O-glucosyltransferase. Post-translational maturation of Notch	Yes	Yes	Yes	No	[213]
Presenilin 1	Receptor proteolysis	No	Yes	Yes	No	[138, 197]
RBP-J	CBF1/ Recombination signal binding protein for immunoglobulin kappa J region. Transcription factor activated by Notch	Yes	Yes	Yes	Yes	[138, 180]

but *R-spondin* in the assembled genome using BLAST 2.11.0+ search [33]. I was able to recover every Wnt gene using exonerate v2.4.0 [15]. CD-HIT [16] returned only 20 of the 25 selected Wnt genes. The genes missing from the analyses were *CK1*, *GSK3*, *LRP*, *Rnf43*, and *Wntless/Evi* (*Wis*). Finally, I used CDD search [17] and identified 21 of 25 genes that contains expected conserved sequence patterns. The results from the different annotation programs for the selected Wnt genes are included in table 4.2

Table 4.2: Select components of the canonical Wnt signaling pathway identified in the draft genome of *S. briareus* using reference sequences from UniProt [13], Echinobase [14], EchinoDB [2], and NCBI [4]. For each gene, we list its name, the known function in the pathway, and whether or not the gene was recovered from the draft genome with independent BLAST [5], exonerate [15], and CD-HIT [16] alignments. In addition, we also indicate if we could identify conserved protein domains [17] in the predicted protein sequences. Source: Mittal et al., in preparation

Name	Role in the Pathway	BLAST	Exonerate	CD-HIT	Conserved domains	References
APC	Negative regulator. Part of the β-catenin destruction complex	Yes	Yes	Yes	Yes	[139]
Axin	Negative regulator. Part of the β-catenin destruction complex	Yes	Yes	Yes	Yes	[139]
β-catenin	Main modulator of the pathway	Yes	Yes	Yes	Yes	[139]
β-TrCP	Ubiquitinates the phosphorylate β-catenin thus targeting it for proteosomal destruction	Yes	Yes	Yes	Yes	[139]
CK1	Phosphorylates β-catenin and the cytoplasmic tail of LRP. Part of the β-catenin destruction complex	Yes	Yes	No	Yes	[139]
Dickkopf	Negative regulator. Binds to LRP	Yes	Yes	Yes	Yes	[139]
Dishevelled	Mediates the recruitment of Axin to the plasmalemma in the ON state of the pathway	Yes	Yes	Yes	Yes	[139]
Frizzled	Wnt receptors	Yes	Yes	Yes	Yes	[139, 100]
Groucho	Negative regulator. Transcriptional co-repressor. Binds to TCF in the OFF state of the pathway	Yes	Yes	Yes	Yes	[139]
GSK3	Phosphorylates β-catenin and the cytoplasmic tail of LRP. Part of the β-catenin destruction complex	Yes	Yes	No	Yes	[139]
Kremen	Dickkopf receptor. Mediates repression of the Wnt pathway	Yes	Yes	Yes	No	[139, 214]
Lgr5	Pathway enhancer. Receptor for R-spondin	Yes	Yes	Yes	Yes	[139]
LRP	Wnt co-receptor	Yes	Yes	No	Yes	[139]
Norrin	Alternative ligand for the Wnt receptors	Yes	Yes	Yes	Yes	[139]
Notum (Wingful)	Negative regulator. Inactivates Wnt in the extracellular space through enzymatic action	Yes	Yes	Yes	Yes	[139, 215]
Porcupine	Palmitoyl transferase, attaches palmitoleic acid to Wnt	Yes	Yes	Yes	No	[139]
Rnf43	Negative regulator. Wnt target gene	Yes	Yes	No	Yes	[139]
R-spondin	Pathway enhancer	No	Yes	Yes	No	[139]
Sclerostin	Negative regulator. Binds to LRP	Yes	Yes	Yes	Yes	[139]
sFRPs	Negative regulators. Sequester Wnts in the extracellular space	Yes	Yes	Yes	Yes	[100]
TCF/Lef	Transcriptional factors regulated by the Wnt pathway. Repress the target genes in the OFF state. Acitvate transcription of the same genes in the ON state	Yes	Yes	Yes	Yes	[139]
Tspan12	Norrin-specific co-receptor	Yes	Yes	Yes	Yes	[139]
Wnt	Paracrine/juxtacrine signaling molecules	Yes	Yes	Yes	Yes	[139, 100]
Wntless/Evi (Wls)	Specific intracellular transporter of Wnts	No	Yes	No	No	[139]
Znrf3	Negative regulator. Wnt target gene	Yes	Yes	Yes	Yes	[139]

and illustrated in figure 4.3.

4.4 Discussion

The important functions attributed to the Notch and Wnt pathways explain, to a certain degree why they are conserved among animals including echinoderms. Accurate coordination of these pathways signals is essential for normal development. For two main reasons, I decided to do this case study to find genes related to the Notch and Wnt pathways in the genome of hairy sea cucumber:

a. I am studying a non-model organism (*S. briareus*), known for its exceptional regeneration abilities so I believe, this is a biologically relevant example as these two pathways (Notch and Wnt) are essential for regeneration in echinoderms.

b. Low BUSCO score of the genome assembly that gives a hint that the assembly could be lacking for practical purposes, like aiding in regenerative studies. Low BUSCO score immediately raises several questions: Is it due to the lack of diverse echinoderm sequences in the BUSCO database, a failure to annotate non-model genomes, or degree of fragmentation in the assembly? Considering that *S. briareus* is a non-model organism and the assembly has a low BUSCO score, I questioned if the assembly would be helpful for regeneration and other biomedical studies? To answer this question, I created our own database of genes that are representative of the regeneration apparatus of the Notch and Wnt pathways. This allowed to demonstrate the usefulness of the genome assembly and the findings (discussed in Results section) revealed that the genome can be used to study the mechanistic role of the signaling pathways involved in the regeneration process in sea cucumbers and other echinoderms.

4.5 Conclusion

Numerous studies have shown how important it is for the Notch signaling system to interact with other signaling pathways throughout the process of tissue regeneration in vertebrates, including echinoderms [216, 49, 217, 218]. For example, the crosstalk between Notch, Bmp, and Wnt pathways in the vertebrate heart development and repair [216], signaling of Notch, TGF-beta, and Wnt in the molecular aspects of regeneration in holothurians [49], interaction between Notch and Wnt signaling pathways in *Drosophila* development and human cancers [217], and the communication of Notch with Wnt, BMP and Yap/Taz pathways in regulation of neuronal stem cells [218]. Therefore, the genome of *S. briareus* provides an opportunity to investigate key genes involved in several pathways, including Notch and Wnt, essential for regeneration. Even though the draft genome is fragmented but based on the results (discussed in the previous section), I can conclude that the genome in its current state is still valuable for a variety of regeneration studies.

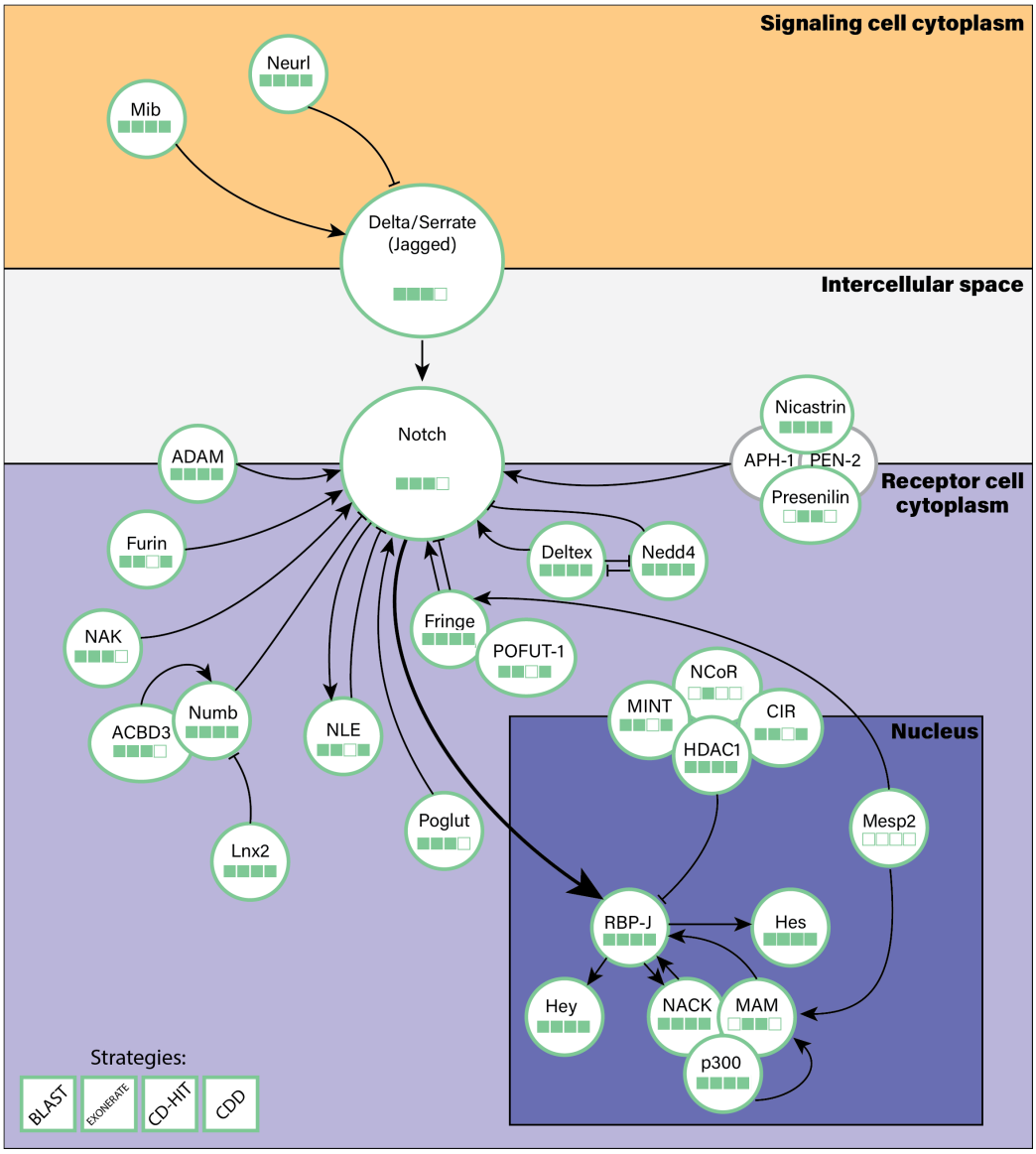

Figure 4.2: Simplified diagram of the Notch signaling pathway (*S. briareus*). The *Delta/Serrate* (*Jagged*) ligands and *Notch* receptors are transmembrane proteins embedded into the plasma membrane of the signaling and receptor cells, respectively. Ligand-receptor interaction triggers conformational changes in the Notch protein that allows for proteolytic cleavage of the receptor by the *ADAM* metalloprotease and the multiprotein γ-secretase complex. The latter includes the catalytic component *presenilin*, as well as regulatory/stabilizing subunits *nicastrin*, *Aph-1*, and *Pen* (*presenilin enhancer*)*-2*. This proteolytic cleavage releases the Notch intercellular domain that translocates into the nucleus and activates the transcription factor *RBP-J* by inducing the release of co-repressors (e.g., *NCOR*, *CIR*, *MINT*, and *HDAC*) and recruitment of co-activators, such as *Mastermind* (*MAM*), *p300*, and *NACK*. The activated transcription factor complex (*Hes* and *Hey*) initiates transcription of the direct targets of the pathway. Adapted from: Mashanov et al., 2022 for Mittal et al., in preparation

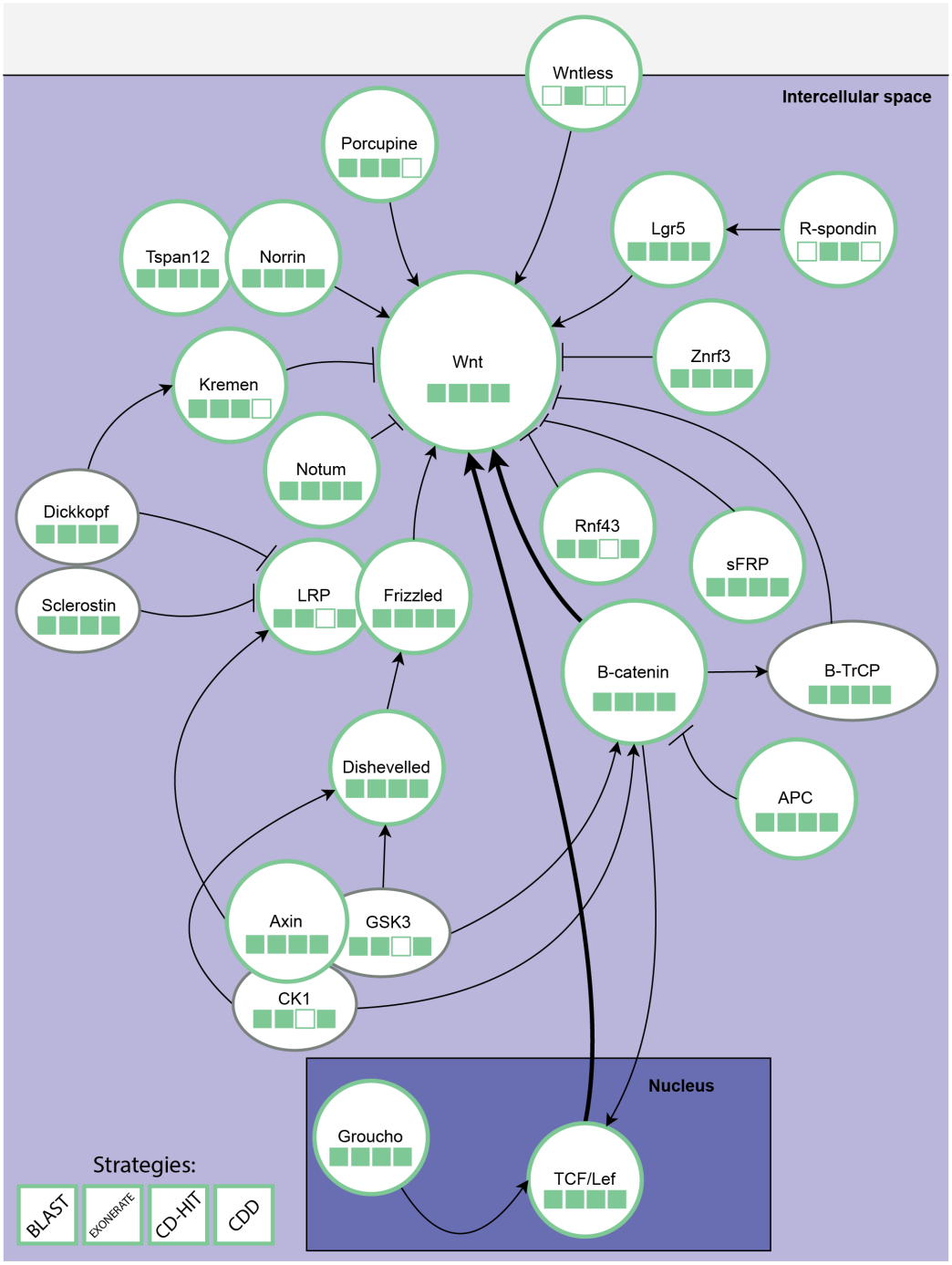

Figure 4.3: Simplified diagram of the Wnt signaling pathway (*S. briareus*). *β-catenin* is a key mediator of the Wnt signaling pathway. In the Wnt OFF state, the transcription factor *TCF* interacts with *Groucho* that mediates transcriptional repression of the target genes of the pathway. If the Wnt signaling is ON, *Axin* is recruited and binds to the cytoplasmic tail of *LRP* which leads to the inactivation of the *β-catenin* complex and converts *TCF* into a transcriptional activator of the same genes repressed by *TCF* in the OFF state. Adapted from: Mashanov et al., 2022 for Mittal et al., in preparation

CHAPTER 5: Viruses and Evolution of *RAG-Like* genes

5.1 Overview

Viruses infect all domains of life including echinoderms [219]. Echinoderms infected with known viruses include sea cucumber (*A. japonicus*), starfish (*A. forbesi*, *A. rubens*, *Henricia sp.*), and sea urchins (*C. atratus*, *T. gratilla*, *E. mathaei*) [220]. These viruses were discovered by RNA-Seq viromics data analysis in echinoderms. A study in 2014 that utilized publicly available sea star transcriptomes, reported that millions of sea stars have died on the west coast of North America due to sea star wasting disease (SSWD) [221]. The study provided evidence that connected the mass mortality to a densovirus (*Parvoviridae*) which rapidly degraded sea stars, resulting in limb autotomy, behavioral abnormalities, and death [221]. Another study (again uses transcriptomic data) revealed a viral infection among sea cucumbers (*A. japonicus*) which may resulted in skin ulcers on the dorsal and ventral side [222]. Reports indicated that an aspherical virus (80–100 nm in diameter) and a spherical virus (120–250 nm in diameter) were observed in affected *A. japonicus* [223] which may have caused skin ulceration, however the relationship between the viruses and the disease is still unclear. Thus, the transcriptomic data of *S. briareus* provide an opportunity to explore highly divergent RNA viruses in sea cucumbers and gain insights into the evolution of immune system in echinoderms.

Echinoderms possess an innate immune system which is characterized as unexpectedly complex, robust, and flexible for detecting and responding to a wide range of potential pathogens in the marine environment [105]. The array of pattern recognition receptors in the genome of *S. purpuratus* suggests that they have a sophisticated system of pathogen detection, mediated by a pair of genes with striking similarities

to vertebrate *RAG* or recombination activating genes [106]. These genes are called *SpRAG1L* (*SpRAG1-Like*) and *SpRAG2L* (*SpRAG2-Like*). The similarities are identified on the basis of locus structure, deduced amino acid sequence and domain structure of these proteins [224]. These genes are essential for the immune system's defense mechanism against foreign substances, including bacteria and viruses.

Recombination activating genes (*RAG*) are characterized by the presence of antigen receptor genes which helps to recognize antigens, including chemicals, bacteria, viruses, or pollen (`https://bioconda.github.io/recipes/titanomics/README.html`). These receptors are the products of the germline recombination of gene segments named V (variable), D (diversity) and J (joining) within the precursors of T and B cells [225, 226, 227]. The recombination of V, D, and J segments is important as it provides greater recognition of foreign invaders and the capacity to fight infections effectively. During V(D)J recombination, the RAG complex binds to a section of DNA called recombination signal sequence (RSS), which lies next to a V, D, or J segment [106, 226, 228, 225]. The RAG complex then cuts the DNA between the segment and the RSS to detach the segment and move it to a different location. This process of DNA rearrangement within T and B cells is repeated multiple times so that the variety of proteins can be produced throughout life to combat potential pathogens in marine environment [228, 226]. The variability in antigen receptors helps them to recognize and respond to specific peptide antigens by producing antibodies against it [227], thus increasing the level of antiviral immunity in an organism. Therefore, the goal of this study was to examine RNA-Seq data of *S. briareus* to discover single-stranded viruses within echinoderms and elucidate the evolution of *RAG-Like* genes in hairy sea cucumbers in relation to other echinoderms, including purple sea urchin, *S. purpuratus*.

5.2 Materials and Methods

Viral signatures within *S. briareus*

Computational viral discovery requires decontamination of host sequences, followed by enrichment of putative viral reads, with *de novo* assembly of the enriched reads into contigs, then concludes with taxonomic/functional annotation of assembled contigs.

Raw reads from *S. briareus* located at the National Center of Biotechnology Information (NCBI) were downloaded using SRA Toolkit v.2.11.3 [171, 1]. These short paired-end Illumina reads were then cleaned using Trimmomatic v0.39 [153] and decontaminated for host sequences using STAR alignment tool v2.7.9a [172]. The specifics of cleaning RNA-Seq reads and aligning these reads against the draft genome of *S. briareus*, are discussed in "Materials and Methods" section of Chapter 3 in detail. Next, I used SAMtools [229] to capture the unmapped reads for viral sequences. The unmapped reads were then *de novo* assembled using SPAdes v3.13.1 [230, 231]. Finally, the key statistics of the assembled putative viral contigs were calculated using Quast v5.0.2 [232].

Viral contigs $>$ 2500 bp in length were searched against RNA-dependent RNA polymerase (RdRP) sequences within Palmscan [233, 234], geNomad [18], and VirSorter [235]. These programs employ a Hidden Markov Model (HMM) approach to identify and annotate divergent viruses in the sequencing data. Contigs positive for RdRp (classified contigs) were queried against viral RefSeq [136] database using Lambda BLASTX [236] for annotation. Contigs negative for RdRp (unclassified contigs) were searched against Non-redundant protein sequences (nr) [4]. Owing to the lack of viral hits, I searched every metaspade contig against the nr database and used NCBI's efetch tool [237] to fetch their taxonomy and description. The tools used for the viral study are summarized in figure 5.1.

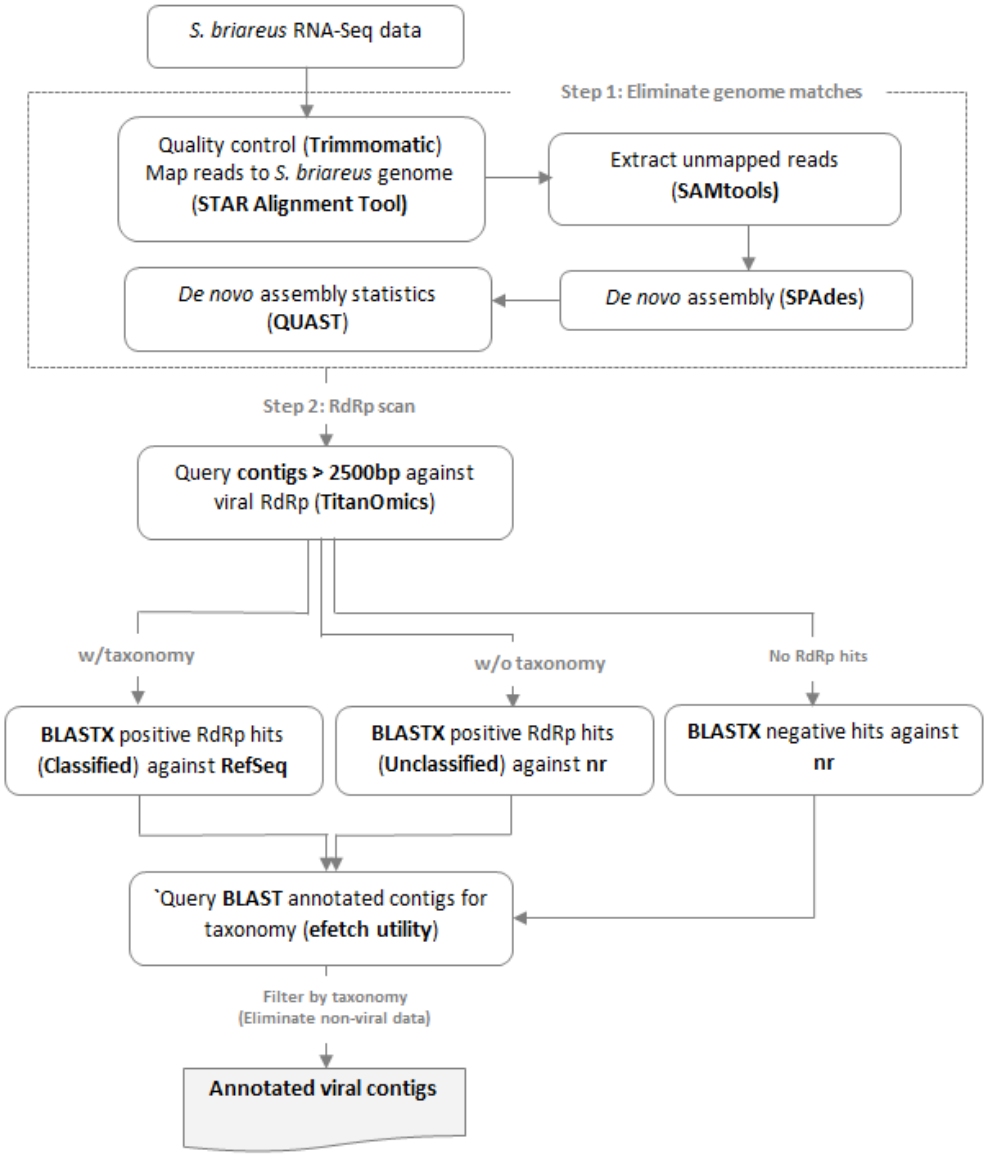

Figure 5.1: Schematic workflow of the procedures used to study divergent viral sequences in the genome of *S. briareus*. Source: Varnika Mittal

RAG-Like sequences within *S. briareus*

Echinoderms have homologues of the recombination activating genes (*RAG*) that protects them against possible pathogens in the marine environment [105]. Those genes are called *SpRAG1-Like* and *SpRAG2-Like* genes. Hence, the protein sequences of *RAG-Like* gene were identified for *S. purpuratus* and other echinoderms at NCBI. These sequences were used as a bait for HMM searching [238] in MAFFT v7.273woe

local alignment [239, 240, 241] with 1000 iterations. The alignment was produced using default settings of 1.53 gap open penalty and 0.123 gap extension penalty. A profile HMM was created for *RAG-Like* protein sequences using hmmer v3.3.2 [242, 243, 19] to profile the similar sequences in the draft genome of *S. briareus*.

Next, I computed a maximum likelihood (ML) tree of the aligned genes with IQ-TREE version 2.2 [244] searching for the best model and partition scheme using ModelFinder as implemented within IQ-TREE [245]. The tree search included 100 searches for the best ML tree inference, 1,000 ultrafast bootstrap pseudo-replicates, and 1,000 replicates for the SH-like approximate likelihood ratio test. Finally, I used SeaView to visualize the computed phylogenetic tree [246].

5.3 Results

Statistics of viral signatures within *S. briareus*

An independent *de novo* assembly (SPAdes v3.13.1 [230, 231]) of extracted unmapped reads resulted in 24,092 contigs with the total assembly length of 35,982,977 bp. (see table 5.1 for more information on the *de novo* assembly). The GC% and N50 of the DNA segments assembled this way were 45.24 and 1,908 bp, respectively. These sequences were then used to aid in the identification of divergent viruses in the draft genome of hairy sea cucumber, *S. briareus*.

Table 5.1: Summary metrics of the *de novo* assembly of the unmapped reads to the draft genome of *S. briareus*. Source: Varnika Mittal

Metrics	**Quantification**
Number of reads	17,242,089
Number of contigs	24,092
Total assembly length (bp)	35,982,977
Largest contig	12,352
Total length (>= 50000 bp)	0
N50 contig length	1,908
L50	5,789
N75 contig length	1,097
L75	12,019
Assembly GC content	45.24%

I employed only contigs longer than 2,500 bp to detect and annotate divergent RNA viruses in the genome which resulted in 2,887 sequences. These sequences were then fed to TitanOmics (`https://pypi.org/project/TitanOmics/`, `https://bioconda.github.io/recipes/titanomics/README.html`) that includes Palmscan [233, 234], geNomad [18], and VirSorter [235] to check for the presence of viral RdRp-like sequences in the genome. The scan for viral RdRp-like sequences uses a profile HMM approach to reveal the viral sequences, even when the sequences have low levels of sequence identity to known viruses. Despite using the HMM-based program, I got only 21 hits (produced by geNomad [18]) in which 6 of them remained "unclassified" that is, their taxonomy is missing. The classified hits were either characterized as a "phage" or a "virus". The taxonomic classification of most of the hits pronounced them as "bacteriophages" with a virus score, ranging between 0.81 and 0.97. I got 2 hits with a virus score of 0.93 that may belong to a phylum of viruses. These results are included in table 5.2.

Table 5.2: Results of the geNomad's [18] analysis. The first column "Sequence" indicate the identifier of the sequence in the input FASTA file. The column "Accession" represent the unique identifier for some of the functionally annotated records. These are assigned based on the entries in Pfam [19], TIGRFAM, COG, and KEGG databases. Column "Length" and "Number of Genes" indicate the length of the sequence and the total number genes encoded in the sequence, respectively. The column "Virus Score" gives you an information of how confident geNomad is that the sequence is a virus and lastly, the column "Taxonomy" represents the taxonomic assignment of the virus genome. Source: Jose Figueroa

Sequence	Accession	Length	Number of Genes	Virus Score	Taxonomy
NODE_35		7111	1	0.972	Viruses;Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes
NODE_2257		2756	1	0.972	"
NODE_2106		2828	1	0.967	"
NODE_2853	PIK57805.1	2510	1	0.965	Unclassified
NODE_1770		3020	1	0.943	Viruses;Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes
NODE_1138		3499	1	0.937	Viruses;Varidnaviria;Bamfordvirae;Nucleocytoviricota;Megaviricetes
NODE_264		5053	1	0.937	Viruses;Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes
NODE_1971		2898	1	0.934	Viruses;Varidnaviria;Bamfordvirae;Nucleocytoviricota;Megaviricetes
NODE_1915		2931	3	0.923	Viruses;Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes
NODE_2724		2561	1	0.918	"
NODE_825		3868	3	0.862	"
NODE_2187	PIK55509.1	2788	1	0.85	Unclassified
NODE_2846	XP_022101867.1	2514	1	0.843	"
NODE_213		5344	1	0.834	Viruses;Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes
NODE_622		4178	4	0.83	"
NODE_300		4925	4	0.824	"
NODE_1885	PIK62152.1	2947	1	0.824	Unclassified
NODE_1897	PIK62575.1	2941	1	0.824	"
NODE_1055		3588	3	0.818	Viruses;Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes
NODE_278		4991	1	0.806	"
NODE_1943	PIK49722.1	2915	3	0.8	Unclassified

Next, I used BLAST [33] to annotate the classified and unclassified hits. The classified hits were searched against viral RefSeq database [136], however we used nr database for the unclassified hits [4]. I also used nr database to query *de novo* assembled contigs longer than 2500 bp (2,887 sequences) to find any viral sequences. Then, I used NCBI's efetch utility (`https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi`) [237] to retrieve the taxonomy and description of the annotated BLAST results. The BLAST search yielded more than 400 results for "bacteria", which is rather common because the hits were classed as Caudoviricetes, a phage that is frequently found in bacterial genomes.

Statistics of *RAG-Like* sequences within *S. briareus*

To understand the immune system of marine echinoderms, I took advantage of the protein sequences deposited at NCBI [136] for *RAG-Like* genes from *S. purpuratus* (Pacific purple sea urchin) and other echinoderms, such as *A. planci* (Crown-of-thorns starfish), *L. variegatus* (Green sea urchin), *P. miniata* (Bat star), and *A. rubens* (Common starfish). I got 32 *RAG1-Like* and 2 *RAG2-Like* protein sequences from NCBI. These protein sequences were then used to generate a local alignment with 1000 iterations (MAFFT v7.273woe [239, 240, 241]), yielding an alignment file that was utilized to construct a profile HMM (HMMER v3.3.2 [242, 243, 19]). In order to find *RAG-Like* immune genes in the draft genome of *S. briareus*, I screened these 34 protein sequences against the profile HMM which resulted into 3 protein sequences that fit the profile of *RAG* genes (see table 5.3 for results).

Lastly, the objective was to align and illustrate the evolution of *RAG-Like* sequences in echinoderms, including the sequences resulted from profile HMM [242, 243, 19] for *S. briareus*. Therefore, I used IQ-Tree 2 [244] and SeaView [246] to infer phylogeny based on conserved positions to determine the genetic relationship of *RAG-Like* immune genes in echinoderms (see figure 5.2).

Table 5.3: HMMScan results for *RAG-Like* sequences found in the draft genome of hairy sea cucumber, *S. briareus*. The first column "Query" indicate the identifier of the sequence in the input FASTA file containing the peptide sequences of hairy sea cucumber and the second column 'E-value' represents the number of matches expected to be found by chance when searching *RAG-Like* sequences against the peptide sequences of hairy sea cucumber, *S. briareus*. Length indicates the total length of the query sequence aligns with the RAG-Like sequence. Source: Varnika Mittal

Query	**Description**	**E-value**	**Length**
transcript_7420	recombination activating protein-like 1 [*Sclerodactyla briareus*]	0.0052	471
transcript_12353	recombination activating protein-like 2 [*Sclerodactyla briareus*]	1e-23	242
transcript_19889	recombination activating protein-like 3 [*Sclerodactyla briareus*]	1.4e-26	224

5.4 Discussion

I examined the transcriptome data of *S. briareus* to mine for RNA-Seq viruses using whole-genome mapping, extraction, *de novo* assembly, and annotation. Finding and describing new viral sequences in the RNA-Seq data of *S. briareus* was the aim. With the exception of three matches to synthetic sequences, I received majority of results for either bacteria or eukaryotes. Based on the BLAST hits to bacteria, it appears that they are either prophages within bacterial genomes or standard microbes that are likely part of the food of the organism.

I also looked at the immune responses and the molecular mechanisms of immune cells in echinoderms by studying *RAG-Like* genes that enable the identification and exclusion of invading microbes in marine environment. *RAG* evolution within echinoderms is critical to our understanding of the development of the adaptive immune system supported by the variability of antigen receptors. Therefore, *RAG* genes are important to study and loss-of-function of *RAG* genes may result in the disruption of V(D)J recombination which may further impair the immune system and can cause genetic disorders or death [247, 248]. The results of the *RAG* study suggest 3 matches in the draft genome of *S. briareus* that correspond with the HMM profile of *RAG-Like* sequences from *S. purpuratus* and other echinoderms (please refer table 5.3 for results). Further, to understand the evolution of *RAG* genes in echinoderms, a gene tree was

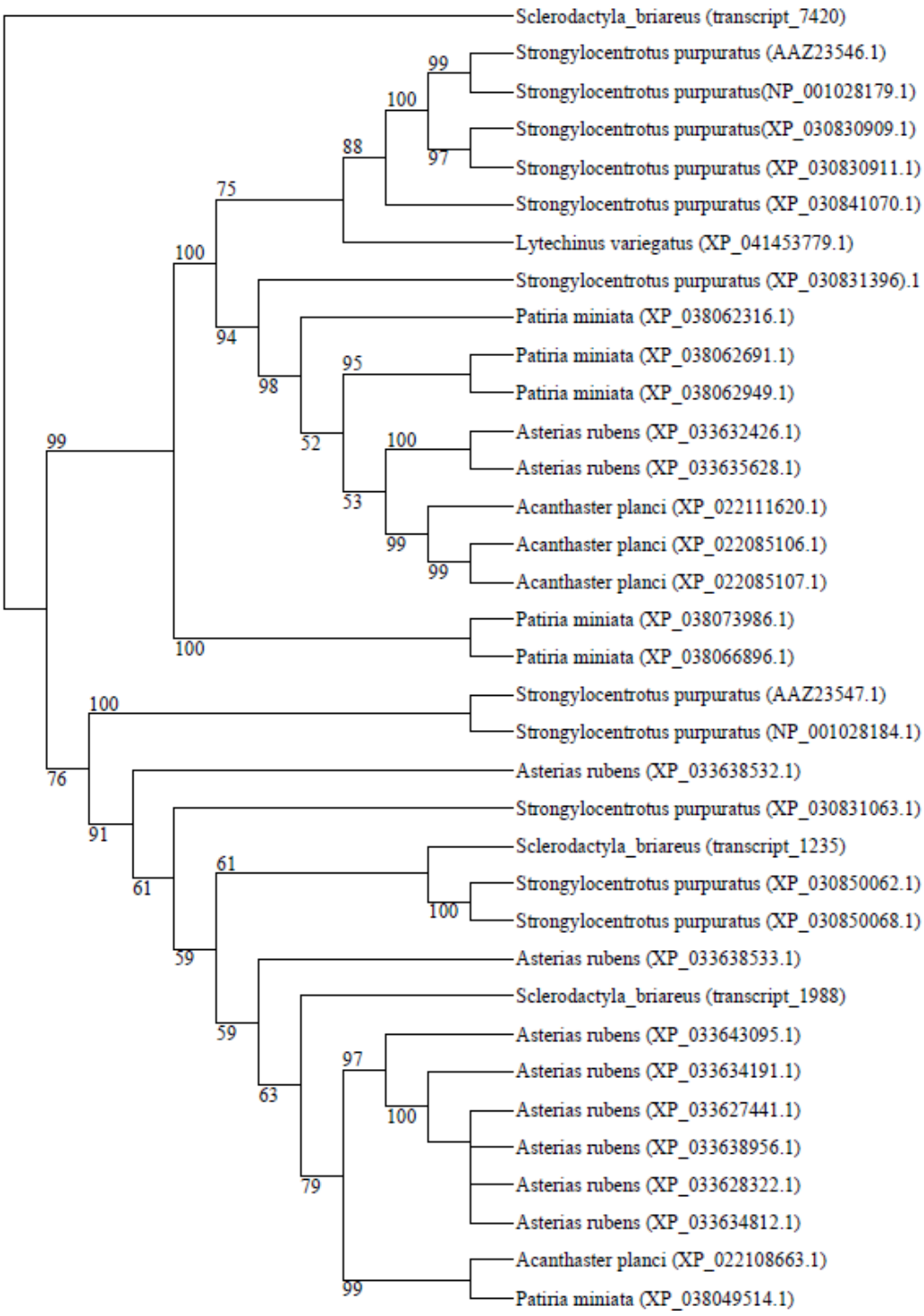

Figure 5.2: A gene phylogeny of *RAG-Like* genes among *S. briareus* and other echinoderms. Source: Varnika Mittal

constructed. The tree shows that, in comparison to the two other sequences, *Sclerodactyla briareus* (transcript_1235) and *Sclerodactyla briareus* (transcript_1988),

Sclerodactyla briareus (transcript_7420) has had more amino acid substitutions. The two sequences appear to be more phylogenetically close to *Strongylocentrotus purpuratus* and *Asterias rubens* as compared to other echinoderms, including *Acanthaster planci*, *Patiria miniata*, and *Lytechinus variegatus* (see figure 3.1 for details).

5.5 Conclusions

The absence of viral findings and lack of *RAG-Like* sequences in the draft genome of *S. briareus* points to:

1. Fragmentation of the assembly and
2. Lack of metagenomics information in the sequenced data

To overcome the limitation, more samples (DNA and RNA) need to be sequenced. We should also use third-generation sequencing technology, such as CRISPR or Oxford Nanopore to produce long DNA fragments to improve the assembly. The improved genome combined with increased data can aid in the detection of divergent single-stranded viruses and phages. Further, the improved genome may facilitate the increased identification of homologous sequences of *RAG-Like* genes, which will help to learn more about the immune system of sea cucumbers in marine atmosphere.

CHAPTER 6: Future Work

Generating error-free assemblies is an ultimate goal of any research project. Despite the ever-advancing improvements in data generation and assembly algorithms, the genome assembly of *S.briareus* is lacking representations of chromosomal units. Therefore, the future work requires bridging of this shortcoming by working on improved datasets to significantly reduce the gaps in the assembly. One way to address this problem is to use third-generation sequencing technology such as, Nanopore sequencing to produce long DNA fragments to fill the gaps in the assembly. Longer reads will provide increased ability to read through highly repetitive and homologous regions and effectively increases the proportion of the genome thus, increasing the confidence of the genomic assembly. The improved genome will better support gene predictions that are crucial for regeneration and other biomedical studies on echinoderms. Additionally, the whole genome will help identify different viruses and phages as well as the defense mechanisms employed by echinoderms to fend off infections in the marine environment.

The future work will also involve improving and maintaining EchinoDB, a publicly accessible, web-facing application and the backend server by keep adding data to the current echinoderm database and maintaining the application for the users who would like to access annotations and infer relationships between taxa or gene content of their sequence data.

REFERENCES

[1] B. University, "*Sclerodactyla briareus de novo* ovary transcriptome paired end reads," 2015. Accessed 1 June 2022. Available from: `https://www.ncbi.nlm.nih.gov/sra/?term=SRR1139189`.

[2] V. Mittal, R. W. Reid, D. J. Machado, V. Mashanov, and D. A. Janies, "Echinodb: An update to the web-based application for genomic and transcriptomic data on echinoderms.," *bioRxiv*, 2022.

[3] D. A. Janies, Z. Witter, G. V. Linchangco, D. W. Foltz, A. K. Miller, A. M. Kerr, J. Jay, R. W. Reid, and G. A. Wray, "Echinodb, an application for comparative transcriptomics of deeply-sampled clades of echinoderms," *BMC bioinformatics*, vol. 17, no. 1, pp. 1–6, 2016.

[4] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D501–D504, 2005.

[5] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden, "Ncbi blast: a better web interface," *Nucleic acids research*, vol. 36, no. suppl_2, pp. W5–W9, 2008.

[6] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "Genbank," *Nucleic acids research*, vol. 41, no. D1, pp. D36–D42, 2012.

[7] U. Consortium, "Uniprot: a hub for protein information," *Nucleic acids research*, vol. 43, no. D1, pp. D204–D212, 2015.

[8] A. Bateman, "Uniprot: a universal hub of protein knowledge," in *Protein Science*, vol. 28, pp. 32–32, WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2019.

[9] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, *et al.*, "Trinity: reconstructing a full-length transcriptome without a genome from rna-seq data," *Nature biotechnology*, vol. 29, no. 7, p. 644, 2011.

[10] R. Henschel, M. Lieber, L.-S. Wu, P. M. Nista, B. J. Haas, and R. D. LeDuc, "Trinity rna-seq assembler performance optimization," in *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond*, pp. 1–8, 2012.

[11] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, "Busco: assessing genome assembly and annotation completeness with single-copy orthologs," *Bioinformatics*, vol. 31, no. 19, pp. 3210–3212, 2015.

[12] R. M. Waterhouse, M. Seppey, F. A. Simão, M. Manni, P. Ioannidis, G. Klioutchnikov, E. V. Kriventseva, and E. M. Zdobnov, "Busco applications from quality assessments to gene prediction and phylogenomics," *Molecular biology and evolution*, vol. 35, no. 3, pp. 543–548, 2018.

[13] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, *et al.*, "The universal protein resource (uniprot)," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D154–D159, 2005.

[14] G. A. Cary, R. A. Cameron, and V. F. Hinman, "Echinobase: tools for echinoderm genome analyses," *Eukaryotic Genomic Databases: Methods and Protocols*, pp. 349–369, 2018.

[15] G. S. C. Slater and E. Birney, "Automated generation of heuristics for biological sequence comparison," *BMC bioinformatics*, vol. 6, no. 1, pp. 1–11, 2005.

[16] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[17] A. Marchler-Bauer, M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz, *et al.*, "Cdd: Ncbi's conserved domain database," *Nucleic acids research*, vol. 43, no. D1, pp. D222–D226, 2015.

[18] A. P. Camargo, S. Roux, F. Schulz, M. Babinski, Y. Xu, B. Hu, P. S. Chain, S. Nayfach, and N. C. Kyrpides, "Identification of mobile genetic elements with genomad," *Nature Biotechnology*, pp. 1–10, 2023.

[19] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, *et al.*, "The pfam protein families database: towards a more sustainable future," *Nucleic acids research*, vol. 44, no. D1, pp. D279–D285, 2016.

[20] Tree of Life Web Project, "Tree of life web project. 2002. deuterostomia. version 01 january 2002 (temporary). http://tolweb.org/deuterostomia/2466/2002.01.01 in the tree of life web project," 2002. [Online; accessed via tolweb.org on 2022-06-03].

[21] D. R. Maddison, K.-S. Schulz, W. P. Maddison, *et al.*, "The tree of life web project," *Zootaxa*, vol. 1668, no. 1, pp. 19–40, 2007.

[22] L. D. Bromham and B. M. Degnan, "Hemichordates and deuterostome evolution: robust molecular phylogenetic support for a hemichordate+ echinoderm clade," *Evolution & Development*, vol. 1, no. 3, pp. 166–171, 1999.

[23] E. A. Lazo-Wasem, "*Thyone briareus* (ypm iz 008052.ec). digital image: Yale peabody museum; photo by eric a. lazo-wasem 2014" - *Thyone briareus* (lesueur, 1824) collected in united states of america (licensed under http://creativecommons.org/publicdomain/zero/1.0/)," 2014. [Online; accessed via GBIF.org on 2024-02-22].

[24] L. Gall, "Invertebrate zoology division, yale peabody museum. yale university peabody museum. occurrence dataset," 2020.

[25] V. Mashanov and J. García-Arrarás, "Gut regeneration in holothurians: a snapshot of recent developments," *The biological bulletin*, vol. 221, no. 1, pp. 93–109, 2011.

[26] V. Mittal *et al.*, "Link to echinodb application." Accessed 25 Feb 2022. Available from: `https://echinodb.uncc.edu`, 2022.

[27] V. Mittal *et al.*, "Link to *Ophioderma brevispinum* data in echinodb application." Accessed 25 Feb 2022. Available from: `https://echinodb.uncc.edu/BStarApp/`, 2022.

[28] V. Mittal *et al.*, "Link to *Lytechinus variegatus* data in echinodb application." Accessed 25 Feb 2022. Available from: `https://echinodb.uncc.edu/SUrchinApp/`, 2022.

[29] S. Andrews, "Fastqc: a quality control tool for high throughput sequence data," 2021.

[30] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern, "Augustus: ab initio prediction of alternative transcripts," *Nucleic acids research*, vol. 34, no. suppl_2, pp. W435–W439, 2006.

[31] T. Brůna, A. Lomsadze, and M. Borodovsky, "Genemark-ep+: eukaryotic gene prediction with self-training in the space of genes and proteins," *NAR genomics and bioinformatics*, vol. 2, no. 2, p. lqaa026, 2020.

[32] T. Brůna, K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky, "Braker2: automatic eukaryotic genome annotation with genemark-ep+ and augustus supported by a protein database," *NAR genomics and bioinformatics*, vol. 3, no. 1, p. lqaa108, 2021.

[33] T. Madden, "The blast sequence analysis tool," in *The NCBI Handbook [Internet]. 2nd edition*, National Center for Biotechnology Information (US), 2013.

[34] H. B. Fell, *Phylum Echinodermata*, pp. 776–837. London: Macmillan Education UK, 1972.

[35] V. S. Mashanov, O. R. Zueva, and J. E. García-Arrarás, "Transcriptomic changes during regeneration of the central nervous system in an echinoderm," *BMC genomics*, vol. 15, no. 1, pp. 1–21, 2014.

[36] G. A. Wray, "Echinodermata. spiny-skinned animals: sea urchins, starfish, and their allies," *Tree of Life*, 1999.

[37] A. E. Bely and K. G. Nyberg, "Evolution of animal regeneration: re-emergence of a field," *Trends in ecology & evolution*, vol. 25, no. 3, pp. 161–170, 2010.

[38] S. Dupont and M. Thorndyke, "Bridging the regeneration gap: insights from echinoderm models," *Nature Reviews Genetics*, vol. 8, no. 4, pp. 320–320, 2007.

[39] Y. Ben Khadra, M. Sugni, C. Ferrario, F. Bonasoro, P. Oliveri, P. Martinez, and M. D. Candia Carnevali, *Regeneration in Stellate Echinoderms: Crinoidea, Asteroidea and Ophiuroidea*, pp. 285–320. Cham: Springer International Publishing, 2018.

[40] M. D. Candia Carnevali and F. Bonasoro, "Introduction to the biology of regeneration in echinoderms," 2001.

[41] M. C. Carnevali, "Regeneration in echinoderms: repair, regrowth, cloning," *Invertebrate Survival Journal*, vol. 3, no. 1, pp. 64–76, 2006.

[42] B. J. Drager, M. A. Harkey, M. Iwata, and A. H. Whiteley, "The expression of embryonic primary mesenchyme genes of the sea urchin, *Strongylocentrotus purpuratus*, in the adult skeletogenic tissues of this and other species of echinoderms," *Developmental biology*, vol. 133, no. 1, pp. 14–23, 1989.

[43] P. Dubois and L. Ameye, "Regeneration of spines and pedicellariae in echinoderms: a review," *Microscopy research and technique*, vol. 55, no. 6, pp. 427–437, 2001.

[44] B. M. Heatfield and D. F. Travis, "Ultrastructural studies of regenerating spines of the sea urchin *Strongylocentrotus purpuratus* i. cell types without spherules," *Journal of Morphology*, vol. 145, no. 1, pp. 13–49, 1975.

[45] K. A. Long, C. W. Nossa, M. A. Sewell, N. H. Putnam, and J. F. Ryan, "Low coverage sequencing of three echinoderm genomes: the brittle star *Ophionereis fasciata*, the sea star *Patiriella regularis*, and the sea cucumber *Australostichopus mollis*," *Gigascience*, vol. 5, no. 1, pp. s13742–016, 2016.

[46] X. Zhang, L. Sun, J. Yuan, Y. Sun, Y. Gao, L. Zhang, S. Li, H. Dai, J.-F. Hamel, C. Liu, *et al.*, "The sea cucumber genome provides insights into morphological evolution and visceral regeneration," *PLoS Biology*, vol. 15, no. 10, p. e2003790, 2017.

[47] C. Rojas-Cartagena, P. Ortíz-Pineda, F. Ramírez-Gómez, E. C. Suárez-Castillo, V. Matos-Cruz, C. Rodríguez, H. Ortíz-Zuazaga, and J. E. García-Arrarás, "Distinct profiles of expressed sequence tags during intestinal regeneration in the sea cucumber *Holothuria glaberrima*," *Physiological genomics*, vol. 31, no. 2, pp. 203–215, 2007.

[48] L. Sun, M. Chen, H. Yang, T. Wang, B. Liu, C. Shu, and D. M. Gardiner, "Large scale gene expression profiling during intestine and body wall regeneration in the sea cucumber *Apostichopus japonicus*," *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, vol. 6, no. 2, pp. 195–205, 2011.

[49] J. G. Medina-Feliciano and J. E. García-Arrarás, "Regeneration in echinoderms: Molecular advancements," *Frontiers in Cell and Developmental Biology*, vol. 9, p. 768641, 2021.

[50] A. G. Bodnar and J. A. Coffman, "Maintenance of somatic tissue regeneration with age in short-and long-lived species of sea urchins," *Aging cell*, vol. 15, no. 4, pp. 778–787, 2016.

[51] V. S. Mashanov, O. R. Zueva, and J. E. García-Arrarás, "Myc regulates programmed cell death and radial glia dedifferentiation after neural injury in an echinoderm," *BMC developmental biology*, vol. 15, no. 1, pp. 1–9, 2015.

[52] L. H. Hyman, "The invertebrates: echinodermata: the coelomate bilateria," *(No Title)*, vol. 4, 1955.

[53] M. D. Candia Carnevali and F. Bonasoro, "Microscopic overview of crinoid regeneration," *Microscopy research and technique*, vol. 55, no. 6, pp. 403–426, 2001.

[54] E. Sodergren, G. M. Weinstock, E. H. Davidson, R. A. Cameron, R. A. Gibbs, R. C. Angerer, L. M. Angerer, M. I. Arnone, D. R. Burgess, R. D. Burke, *et al.*, "The genome of the sea urchin *Strongylocentrotus purpuratus*," *Science*, vol. 314, no. 5801, pp. 941–952, 2006.

[55] Y. Li, R. Wang, X. Xun, J. Wang, L. Bao, R. Thimmappa, J. Ding, J. Jiang, L. Zhang, T. Li, *et al.*, "Sea cucumber genome provides insights into saponin biosynthesis and aestivation regulation," *Cell Discovery*, vol. 4, no. 1, p. 29, 2018.

[56] L. Sun, C. Jiang, F. Su, W. Cui, and H. Yang, "Chromosome-level genome assembly of the sea cucumber *Apostichopus japonicus*," *Scientific Data*, vol. 10, no. 1, p. 454, 2023.

[57] S. C. Materna, K. Berney, and R. A. Cameron, "The *S. purpuratus* genome: a comparative perspective," *Developmental biology*, vol. 300, no. 1, pp. 485–495, 2006.

[58] V. Mashanov, J. Akiona, M. Khoury, J. Ferrier, R. Reid, D. J. Machado, O. Zueva, and D. Janies, "Active notch signaling is required for arm regeneration in a brittle star," *PloS one*, vol. 15, no. 5, p. e0232981, 2020.

[59] D. Janies and R. Mooi, "*Xyloplax* is an asteroid," *Echinoderm research*, pp. 311–316, 1998.

[60] D. A. Janies, J. R. Voight, and M. Daly, "Echinoderm phylogeny including *Xyloplax*, a progenetic asteroid," *Systematic biology*, vol. 60, no. 4, pp. 420–438, 2011.

[61] G. V. Linchangco Jr, D. W. Foltz, R. Reid, J. Williams, C. Nodzak, A. M. Kerr, A. K. Miller, R. Hunter, N. G. Wilson, W. J. Nielsen, *et al.*, "The phylogeny of extant starfish (asteroidea: Echinodermata) including *Xyloplax*, based on comparative transcriptomics," *Molecular phylogenetics and evolution*, vol. 115, pp. 161–170, 2017.

[62] P. L. Davidson, H. Guo, L. Wang, A. Berrio, H. Zhang, Y. Chang, A. L. Soborowski, D. R. McClay, G. Fan, and G. A. Wray, "Chromosomal-level genome assembly of the sea urchin *Lytechinus variegatus* substantially improves functional genomic analyses," *Genome biology and evolution*, vol. 12, no. 7, pp. 1080–1086, 2020.

[63] S. Zamora, I. A. Rahman, and A. B. Smith, "Plated cambrian bilaterians reveal the earliest stages of echinoderm evolution," *PLoS One*, vol. 7, no. 6, p. e38296, 2012.

[64] G. B. Gillard, D. J. Garama, and C. M. Brown, "The transcriptome of the nz endemic sea urchin kina (*Evechinus chloroticus*)," *BMC genomics*, vol. 15, no. 1, pp. 1–15, 2014.

[65] M. R. Hall, K. M. Kocot, K. W. Baughman, S. L. Fernandez-Valverde, M. E. Gauthier, W. L. Hatleberg, A. Krishnan, C. McDougall, C. A. Motti, E. Shoguchi, *et al.*, "The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef pest," *Nature*, vol. 544, no. 7649, pp. 231–234, 2017.

[66] J. G. Medina-Feliciano, S. Pirro, J. E. García-Arrarás, V. Mashanov, and J. F. Ryan, "Draft genome of the sea cucumber *Holothuria glaberrima*, a model for the study of regeneration," *Frontiers in Marine Science*, vol. 8, p. 321, 2021.

[67] L. Zhang, J. He, P. Tan, Z. Gong, S. Qian, Y. Miao, H.-Y. Zhang, G. Tu, Q. Chen, Q. Zhong, *et al.*, "The genome of an apodid holothuroid (*Chiridota heheva*) provides insights into its adaptation to a deep-sea reducing environment," *Communications Biology*, vol. 5, no. 1, pp. 1–11, 2022.

[68] S. Kinjo, M. Kiyomoto, T. Yamamoto, K. Ikeo, and S. Yaguchi, "Hpbase: A genome database of a sea urchin, *Hemicentrotus pulcherrimus*," *Development, Growth & Differentiation*, vol. 60, no. 3, pp. 174–182, 2018.

[69] M. Elphick, D. Semmens, L. Blowes, J. Levine, C. Lowe, M. Arnone, *et al.*, "Reconstructing salmfamide neuropeptide precursor evolution in the phylum echinodermata: ophiuroid and crinoid sequence data provide new insights. front endocrinol (lausanne). 2015; 6: 1–10," 2015.

[70] V. Mashanov, D. J. Machado, R. Reid, C. Brouwer, J. Kofsky, and D. A. Janies, "Twinkle twinkle brittle star: the draft genome of *Ophioderma brevispinum* (echinodermata: Ophiuroidea) as a resource for regeneration research," *BMC genomics*, vol. 23, no. 1, pp. 1–17, 2022.

[71] P. V. Sergiev, A. A. Artemov, E. B. Prokhortchouk, O. A. Dontsova, and G. V. Berezkin, "Genomes of *Strongylocentrotus franciscanus* and *Lytechinus variegatus*: are there any genomic explanations for the two order of magnitude difference in the lifespan of sea urchins?," *Aging (Albany NY)*, vol. 8, no. 2, p. 260, 2016.

[72] J. E. García-Arrarás, L. Estrada-Rodgers, R. Santiago, I. I. Torres, L. Díaz-Miranda, and I. Torres-Avillán, "Cellular mechanisms of intestine regeneration in the sea cucumber, *Holothuria glaberrima* selenka (holothuroidea: Echinodermata)," *Journal of Experimental Zoology*, vol. 281, no. 4, pp. 288–304, 1998.

[73] J. E. García-Arrarás, M. I. Lázaro-Peña, and C. A. Díaz-Balzac, "Holothurians as a model system to study regeneration," *Marine organisms as model systems in biology and medicine*, pp. 255–283, 2018.

[74] H. C. Reinardy, C. E. Emerson, J. M. Manley, and A. G. Bodnar, "Tissue regeneration and biomineralization in sea urchins: role of notch signaling and presence of stem cell markers," *PloS one*, vol. 10, no. 8, p. e0133860, 2015.

[75] R. A. Cameron, M. Samanta, A. Yuan, D. He, and E. Davidson, "Spbase: the sea urchin genome database and web site," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D750–D754, 2009.

[76] P. Kudtarkar and R. A. Cameron, "Echinobase: an expanding resource for echinoderm genomic information," *Database*, vol. 2017, 2017.

[77] C. H. Wu, L.-S. L. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R. S. Ledley, B. E. Suzek, *et al.*, "The protein information resource," *Nucleic acids research*, vol. 31, no. 1, pp. 345–347, 2003.

[78] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[79] V. S. Mashanov, I. Y. Dolmatov, and T. Heinzeller, "Transdifferentiation in holothurian gut regeneration," *The Biological Bulletin*, vol. 209, no. 3, pp. 184–193, 2005.

[80] C. Rojas-Cartagena, P. Ortíz-Pineda, F. Ramírez-Gómez, E. C. Suárez-Castillo, V. Matos-Cruz, C. Rodríguez, H. Ortíz-Zuazaga, and J. E. García-Arrarás, "Distinct profiles of expressed sequence tags during intestinal regeneration in the sea cucumber *Holothuria glaberrima*," *Physiological genomics*, vol. 31, no. 2, pp. 203–215, 2007.

[81] P. A. Ortiz-Pineda, F. Ramírez-Gómez, J. Pérez-Ortiz, S. González-Díaz, S.-D. Jesús, J. Hernández-Pasos, D. Valle-Avila, C. Rojas-Cartagena, E. C. Suárez-Castillo, K. Tossas, *et al.*, "Gene expression profiling of intestinal regeneration in the sea cucumber," *BMC genomics*, vol. 10, no. 1, pp. 1–21, 2009.

[82] V. S. Mashanov, O. R. Zueva, and J. E. García-Arrarás, "Posttraumatic regeneration involves differential expression of long terminal repeat (ltr) retrotransposons," *Developmental Dynamics*, vol. 241, no. 10, pp. 1625–1636, 2012.

[83] S. A. Bello, V. Torres-Gutiérrez, E. J. Rodríguez-Flores, E. J. Toledo-Román, N. Rodríguez, L. M. Díaz-Díaz, L. D. Vázquez-Figueroa, J. M. Cuesta, V. Grillo-Alvarado, A. Amador, *et al.*, "Insights into intestinal regeneration signaling mechanisms," *Developmental biology*, vol. 458, no. 1, pp. 12–31, 2020.

[84] D. J. Quispe-Parra, J. G. Medina-Feliciano, S. Cruz-González, H. Ortiz-Zuazaga, and J. E. García-Arrarás, "Transcriptomic analysis of early stages of intestinal regeneration in *Holothuria glaberrima*," *Scientific reports*, vol. 11, no. 1, pp. 1–14, 2021.

[85] M. Alicea-Delgado and J. E. García-Arrarás, "*Wnt/β-catenin* signaling pathway regulates cell proliferation but not muscle dedifferentiation nor apoptosis during sea cucumber intestinal regeneration," *Developmental Biology*, vol. 480, pp. 105–113, 2021.

[86] J. G. Medina-Feliciano, S. Pirro, J. E. García-Arrarás, V. Mashanov, and J. F. Ryan, "Draft genome of the sea cucumber *Holothuria glaberrima*, a model for the study of regeneration," *Frontiers in Marine Science*, vol. 8, p. 603410, 2021.

[87] D. L. Pawson, *Marine Flora and Fauna of the Northeastern United States: Echinodermata: Holothuroidea*, vol. 405. Department of Commerce, National Oceanic and Atmospheric Administration . . . , 1977.

[88] D. L. Pawson, D. J. Pawson, and R. A. King, "A taxonomic guide to the echinodermata of the south atlantic bight, usa: 1. sea cucumbers (echinodermata: Holothuroidea)," *Zootaxa*, vol. 2449, no. 1, pp. 1–48, 2010.

[89] V. S. Mashanov, O. Zueva, and J. E. García-Arrarás, "Postembryonic organogenesis of the digestive tube: why does it occur in worms and sea cucumbers but fail in humans?," *Current topics in developmental biology*, vol. 108, pp. 185–216, 2014.

[90] V. S. Mashanov, O. R. Zueva, and J. E. García-Arrarás, "Radial glial cells play a key role in echinoderm neural regeneration," *BMC biology*, vol. 11, no. 1, pp. 1–19, 2013.

[91] D. N. Menton and A. Z. Eisen, "Cutaneous wound healing in the sea cucumber, *Thyone briareus*," *Journal of morphology*, vol. 141, no. 2, pp. 185–203, 1973.

[92] S. Miguel-Ruiz, E. José, and J. E. García-Arrarás, "Common cellular events occur during wound healing and organ regeneration in the sea cucumber *Holothuria glaberrima*," *BMC Developmental Biology*, vol. 7, no. 1, pp. 1–19, 2007.

[93] I. Y. Dolmatov and T. T. Ginanova, "Muscle regeneration in holothurians," *Microscopy research and technique*, vol. 55, no. 6, pp. 452–463, 2001.

[94] V. S. Mashanov, O. Zueva, and J. E. García-Arrarás, "Postembryonic organogenesis of the digestive tube: why does it occur in worms and sea cucumbers but fail in humans?," *Current topics in developmental biology*, vol. 108, pp. 185–216, 2014.

[95] A. L. Rychel and B. J. Swalla, *Regeneration in Hemichordates and Echinoderms*, pp. 245–265. Dordrecht: Springer Netherlands, 2009.

[96] E. Gazave, P. Lapébie, G. S. Richards, F. Brunet, A. V. Ereskovsky, B. M. Degnan, C. Borchiellini, M. Vervoort, and E. Renard, "Origin and evolution of the notch signaling pathway: an overview from eukaryotic genomes," *BMC evolutionary biology*, vol. 9, no. 1, pp. 1–27, 2009.

[97] H. Marlow, E. Roettinger, M. Boekhout, and M. Q. Martindale, "Functional roles of notch signaling in the cnidarian *Nematostella vectensis*," *Developmental biology*, vol. 362, no. 2, pp. 295–308, 2012.

[98] M. J. Layden and M. Q. Martindale, "Non-canonical notch signaling represents an ancestral mechanism to regulate neural differentiation," *Evodevo*, vol. 5, no. 1, pp. 1–14, 2014.

[99] M. B. Favarolo and S. L. López, "Notch signaling in the division of germ layers in bilaterian embryos," *Mechanisms of Development*, vol. 154, pp. 122–144, 2018.

[100] J. C. Croce, S.-Y. Wu, C. Byrum, R. Xu, L. Duloquin, A. H. Wikramanayake, C. Gache, and D. R. McClay, "A genome-wide survey of the evolutionarily conserved wnt pathways in the sea urchin *Strongylocentrotus purpuratus*," *Developmental biology*, vol. 300, no. 1, pp. 121–131, 2006.

[101] J. Yuan, Y. Gao, L. Sun, S. Jin, X. Zhang, C. Liu, F. Li, and J. Xiang, "Wnt signaling pathway linked to intestinal regeneration via evolutionary patterns and gene expression in the sea cucumber *Apostichopus japonicus*," *Frontiers in genetics*, vol. 10, p. 112, 2019.

[102] D. O. Mellott, J. Thisdelle, and R. D. Burke, "Notch signaling patterns neurogenic ectoderm and regulates the asymmetric division of neural progenitors in sea urchin embryos," *Development*, vol. 144, no. 19, pp. 3602–3611, 2017.

[103] S. Foster, Y. V. Teo, N. Neretti, N. Oulhen, and G. M. Wessel, "Single cell rna-seq in the sea urchin embryo show marked cell-type specificity in the

Delta/Notch pathway," *Molecular reproduction and development*, vol. 86, no. 8, pp. 931–934, 2019.

[104] J. S. Kauffman and R. A. Raff, "Patterning mechanisms in the evolution of derived developmental life histories: the role of wnt signaling in axis formation of the direct-developing sea urchin *Heliocidaris erythrogramma*," *Development genes and evolution*, vol. 213, no. 12, pp. 612–624, 2003.

[105] L. C. Smith, V. Arizza, M. A. Barela Hudgell, G. Barone, A. G. Bodnar, K. M. Buckley, V. Cunsolo, N. M. Dheilly, N. Franchi, S. D. Fugmann, *et al.*, "Echinodermata: the complex immune system in echinoderms," in *Advances in comparative immunology*, pp. 409–501, Springer, 2018.

[106] M. Gellert, "V(d)j recombination: *RAG* proteins, repair factors, and regulation," *Annual review of biochemistry*, vol. 71, no. 1, pp. 101–132, 2002.

[107] M. S. Elliott and L. M. Elliott, "Developing r shiny web applications for extension education," *Applied Economics Teaching Resources (AETR)*, vol. 2, no. 4, pp. 9–19, 2020.

[108] S. Mathew and J. Varia, "Overview of amazon web services," *Amazon Whitepapers*, vol. 105, pp. 1–22, 2014.

[109] G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. Madsen, "Blast+: architecture and applications," *BMC Bioinformatics*, vol. 10, p. 421, 2008.

[110] A. Priyam, B. J. Woodcroft, V. Rai, I. Moghul, A. Munagala, F. Ter, H. Chowdhary, I. Pieniak, L. J. Maynard, M. A. Gibbins, *et al.*, "Sequenceserver: a modern graphical user interface for custom blast databases," *Molecular biology and evolution*, vol. 36, no. 12, pp. 2922–2924, 2019.

[111] L. Sun, H. Yang, M. Chen, and D. Xu, "Cloning and expression analysis of *Wnt6* and *Hox6* during intestinal regeneration in the sea cucumber *Apostichopus japonicus*," *Genet. Mol. Res*, vol. 12, no. 4, pp. 5321–5334, 2013.

[112] V. S. Mashanov, O. R. Zueva, and J. E. Garcia-Arraras, "Expression of *Wnt9*, *TCTP*, and *Bmp1/Tll* in sea cucumber visceral regeneration," *Gene Expression Patterns*, vol. 12, no. 1-2, pp. 24–35, 2012.

[113] M. Alicea-Delgado and J. E. García-Arrarás, "Wnt/β-catenin signaling pathway regulates cell proliferation but not muscle dedifferentiation nor apoptosis during sea cucumber intestinal regeneration," *Developmental biology*, vol. 480, pp. 105–113, 2021.

[114] V. Mittal *et al.*, "Link to blast sequenceserver in echinodb application." Accessed 25 Feb 2022. Available from: `https://echinodb.uncc.edu/sequenceserver/`, 2022.

[115] M. Pendola, G. Jain, and J. S. Evans, "Skeletal development in the sea urchin relies upon protein families that contain intrinsic disorder, aggregation-prone, and conserved globular interactive domains," *Plos one*, vol. 14, no. 10, p. e0222068, 2019.

[116] B. Livingston, C. Killian, F. Wilt, A. Cameron, M. Landrum, O. Ermolaeva, V. Sapojnikov, D. Maglott, A. Buchanan, and C. Ettensohn, "A genome-wide analysis of biomineralization-related proteins in the sea urchin *Strongylocentrotus purpuratus*," *Developmental biology*, vol. 300, no. 1, pp. 335–348, 2006.

[117] J. S. Evans, "The biomineralization proteome: protein complexity for a complex bioceramic assembly process," *Proteomics*, vol. 19, no. 16, p. 1900036, 2019.

[118] H. A. Lowenstam and S. Weiner, *On biomineralization*. Oxford University Press, USA, 1989.

[119] S. Mann, "Biomineralization: principles and concepts in bioinorganic materials chemistry," 2001.

[120] I. C. Wilkie, M. Sugni, H. Gupta, M. C. Carnevali, and M. Elphick, "The mutable collagenous tissue of echinoderms: From biology to biomedical applications," 2021.

[121] M. Tamori, A. Yamada, N. Nishida, Y. Motobayashi, K. Oiwa, and T. Motokawa, "*Tensilin-like* stiffening protein from *Holothuria leucospilota* does not induce the stiffest state of catch connective tissue," *Journal of experimental biology*, vol. 209, no. 9, pp. 1594–1602, 2006.

[122] J. P. Tipper, G. Lyons-Levy, M. A. Atkinson, and J. A. Trotter, "Purification, characterization and cloning of *tensilin*, the collagen-fibril binding and tissue-stiffening factor from *Cucumaria frondosa* dermis," *Matrix biology*, vol. 21, no. 8, pp. 625–635, 2002.

[123] M. Demeuldre, E. Hennebert, M. Bonneel, B. Lengerer, S. Van Dyck, R. Wattiez, P. Ladurner, and P. Flammang, "Mechanical adaptability of sea cucumber *Cuvierian tubules* involves a mutable collagenous tissue," *Journal of Experimental Biology*, vol. 220, no. 11, pp. 2108–2119, 2017.

[124] M. Inoue, R. Birenheide, O. Koizumi, Y. Kobayakawa, Y. Muneoka, and T. Motokawa, "Localization of the neuropeptide ngiwyamide in the holothurian nervous system and its effects on muscular contraction," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 266, no. 1423, pp. 993–1000, 1999.

[125] R. Birenheide, M. Tamori, T. Motokawa, M. Ohtani, E. Iwakoshi, Y. Muneoka, T. Fujita, H. Minakata, and K. Nomoto, "Peptides controlling stiffness of connective tissue in sea cucumbers," *The Biological Bulletin*, vol. 194, no. 3, pp. 253–259, 1998.

[126] W. Reese, "Nginx: the high-performance web server and reverse proxy," *Linux Journal*, vol. 2008, no. 173, p. 2, 2008.

[127] B. Haas and A. Papanicolaou, "Transdecoder." Available from: `http://transdecoder.github.io/`, 2014. Accessed 13 July 2022.

[128] L. Li, C. J. Stoeckert, and D. S. Roos, "Orthomcl: identification of ortholog groups for eukaryotic genomes," *Genome research*, vol. 13, no. 9, pp. 2178–2189, 2003.

[129] National Center for Biotechnology Information (NCBI), "The ncbi collection of predicted proteins of the sea urchin *Strongylocentrotus purpuratus*," 2021. `ftp://ftp.ncbi.nih.gov/genomes/Strongylocentrotus_purpuratus/protein/`.

[130] B. C. of Medicine, "Strongylocentrotus purpuratus spur_5.0 protein models," 2019. `https://www.ncbi.nlm.nih.gov/assembly/GCF_000002235.5`.

[131] S. Dupont and M. C. Thorndyke, "Growth or differentiation? adaptive regeneration in the brittlestar amphiura filiformis," *Journal of Experimental Biology*, vol. 209, no. 19, pp. 3873–3881, 2006.

[132] O. Bronstein and A. Kroh, "The first mitochondrial genome of the model echinoid *Lytechinus variegatus* and insights into odontophoran phylogenetics," *Genomics*, vol. 111, no. 4, pp. 710–718, 2019.

[133] A. Czarkwiani, D. V. Dylus, and P. Oliveri, "Expression of skeletogenic genes during arm regeneration in the brittle star *Amphiura filiformis*," *Gene Expression Patterns*, vol. 13, no. 8, pp. 464–472, 2013.

[134] A. Czarkwiani, C. Ferrario, D. V. Dylus, M. Sugni, and P. Oliveri, "Skeletal regeneration in the brittle star *Amphiura filiformis*," *Frontiers in zoology*, vol. 13, pp. 1–17, 2016.

[135] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, and A. Bairoch, "Uniprotkb/swiss-prot," in *Plant bioinformatics*, pp. 89–112, Springer, 2007.

[136] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, *et al.*, "Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation," *Nucleic acids research*, vol. 44, no. D1, pp. D733–D745, 2016.

[137] V. Mashanov and D. Janies, "Ncbi's accession details for the data for article - active notch signaling is required for arm regeneration in a brittle star.," 2021. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE142391`.

[138] G. D. Hurlbut, M. W. Kankel, R. J. Lake, and S. Artavanis-Tsakonas, "Crossing paths with *Notch* in the hyper-network," *Current opinion in cell biology*, vol. 19, no. 2, pp. 166–175, 2007.

[139] R. Nusse and H. Clevers, "*Wnt/β-catenin* signaling, disease, and emerging therapeutic modalities," *Cell*, vol. 169, no. 6, pp. 985–999, 2017.

[140] J. E. García-Arrarás and I. Y. Dolmatov, "Echinoderms: potential model systems for studies on muscle regeneration," *Current pharmaceutical design*, vol. 16, no. 8, pp. 942–955, 2010.

[141] A. G. Candelaria, G. Murray, S. K. File, and J. E. García-Arrarás, "Contribution of mesenterial muscle dedifferentiation to intestine regeneration in the sea cucumber *Holothuria glaberrima*," *Cell and tissue research*, vol. 325, no. 1, pp. 55–65, 2006.

[142] I. Y. Dolmatov, "Molecular aspects of regeneration mechanisms in holothurians," *Genes*, vol. 12, no. 2, p. 250, 2021.

[143] I. Y. Dolmatov and T. T. Ginanova, "Muscle regeneration in holothurians," *Microscopy research and technique*, vol. 55, no. 6, pp. 452–463, 2001.

[144] C. Ferrario, M. Sugni, I. M. Somorjai, and L. Ballarin, "Beyond adult stem cells: Dedifferentiation as a unifying mechanism underlying regeneration in invertebrate deuterostomes," *Frontiers in Cell and Developmental Biology*, vol. 8, p. 587320, 2020.

[145] V. S. Mashanov, O. R. Zueva, C. Rojas-Catagena, and J. E. Garcia-Arraras, "Visceral regeneration in a sea cucumber involves extensive expression of *survivin* and *mortalin* homologs in the mesothelium," *BMC Developmental Biology*, vol. 10, pp. 1–24, 2010.

[146] I. Y. Dolmatov, S. V. Afanasyev, and A. V. Boyko, "Molecular mechanisms of fission in echinoderms: Transcriptome analysis," *PLoS One*, vol. 13, no. 4, p. e0195836, 2018.

[147] Y. Wang, M. Tian, Y. Chang, C. Xue, and Z. Li, "Investigation of structural proteins in sea cucumber (*Apostichopus japonicus*) body wall," *Scientific Reports*, vol. 10, no. 1, p. 18744, 2020.

[148] D. C. Hardie, T. R. Gregory, and P. D. Hebert, "From pixels to picograms: a beginners' guide to genome quantification by feulgen image analysis densitometry," *Journal of Histochemistry & Cytochemistry*, vol. 50, no. 6, pp. 735–749, 2002.

[149] E. M. Rasch, C. E. Lee, and G. A. Wyngaard, "Dna–feulgen cytophotometric determination of genome size for the freshwater-invading copepod eurytemora affinis," *Genome*, vol. 47, no. 3, pp. 559–564, 2004.

[150] V. S. Donnenberg, R. J. Landreneau, M. E. Pfeifer, and A. D. Donnenberg, "Flow cytometric determination of stem/progenitor content in epithelial tissues: an example from nonsmall lung cancer and normal lung," *Cytometry Part A*, vol. 83, no. 1, pp. 141–149, 2013.

[151] G. Marçais and C. Kingsford, "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers," *Bioinformatics*, vol. 27, no. 6, pp. 764–770, 2011.

[152] G. W. Vurture, F. J. Sedlazeck, M. Nattestad, C. J. Underwood, H. Fang, J. Gurtowski, and M. C. Schatz, "Genomescope: fast reference-free genome profiling from short reads," *Bioinformatics*, vol. 33, no. 14, pp. 2202–2204, 2017.

[153] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

[154] J. O'Connell, O. Schulz-Trieglaff, E. Carlson, M. M. Hims, N. A. Gormley, and A. J. Cox, "Nxtrim: optimized trimming of illumina mate pair reads," *Bioinformatics*, vol. 31, no. 12, pp. 2035–2037, 2015.

[155] X. Yang, D. Liu, F. Liu, J. Wu, J. Zou, X. Xiao, F. Zhao, and B. Zhu, "Htqc: a fast quality control toolkit for illumina sequencing data," *BMC bioinformatics*, vol. 14, no. 1, pp. 1–4, 2013.

[156] H. Xu, X. Luo, J. Qian, X. Pang, J. Song, G. Qian, J. Chen, and S. Chen, "Fastuniq: a fast *de novo* duplicates removal tool for paired short reads," *PloS one*, vol. 7, no. 12, p. e52249, 2012.

[157] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol, "Abyss: a parallel assembler for short read sequence data," *Genome research*, vol. 19, no. 6, pp. 1117–1123, 2009.

[158] A. Thrash, F. Hoffmann, and A. Perkins, "Toward a more holistic method of genome assembly assessment," *BMC bioinformatics*, vol. 21, no. 4, pp. 1–8, 2020.

[159] B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, *et al.*, "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement," *PloS one*, vol. 9, no. 11, p. e112963, 2014.

[160] J. Ruan and H. Li, "Fast and accurate long-read assembly with wtdbg2," *Nature methods*, vol. 17, no. 2, pp. 155–158, 2020.

[161] M. Chakraborty, J. G. Baldwin-Brown, A. D. Long, and J. Emerson, "Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage," *Nucleic acids research*, vol. 44, no. 19, pp. e147–e147, 2016.

[162] M. Hunt, "assembly-stats." Available from: `https://github.com/sanger-pathogens/assembly-stats`, 2014-2021. Accessed 18 Feb 2019.

[163] U. D. B. Core, "assemblathon2-analysis." Available from: `https://github.com/ucdavis-bioinformatics/assemblathon2-analysis`, 2012. Accessed 24 Mar 2019.

[164] C. Chu, R. Nielsen, and Y. Wu, "Repdenovo: inferring de novo repeat motifs from short sequence reads," *PloS one*, vol. 11, no. 3, p. e0150719, 2016.

[165] J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, and A. F. Smit, "Repeatmodeler2 for automated genomic discovery of transposable element families," *Proceedings of the National Academy of Sciences*, vol. 117, no. 17, pp. 9451–9457, 2020.

[166] J. Jurka, "Repeats in genomic dna: mining and meaning," *Current opinion in structural biology*, vol. 8, no. 3, pp. 333–337, 1998.

[167] J. Jurka, "Repbase update: a database and an electronic journal of repetitive elements," *Trends in genetics*, vol. 16, no. 9, pp. 418–420, 2000.

[168] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz, "Repbase update, a database of eukaryotic repetitive elements," *Cytogenetic and genome research*, vol. 110, no. 1-4, pp. 462–467, 2005.

[169] W. Bao, K. K. Kojima, and O. Kohany, "Repbase update, a database of repetitive elements in eukaryotic genomes," *Mobile Dna*, vol. 6, no. 1, pp. 1–6, 2015.

[170] A. Smit, R. Hubley, and P. Green, "Repeatmasker open-4.0," 2013-2015. Accessed 10 Nov 2020. Available from: `http://www.repeatmasker.org`.

[171] R. Leinonen, H. Sugawara, M. Shumway, and I. N. S. D. Collaboration, "The sequence read archive," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D19–D21, 2010.

[172] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "Star: ultrafast universal rna-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[173] K. J. Hoff, S. Lange, A. Lomsadze, M. Borodovsky, and M. Stanke, "Braker1: unsupervised rna-seq-based genome annotation with genemark-et and augustus," *Bioinformatics*, vol. 32, no. 5, pp. 767–769, 2016.

[174] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, *et al.*, "The ensembl genome database project," *Nucleic acids research*, vol. 30, no. 1, pp. 38–41, 2002.

[175] B. C. of Medicine, "*Strongylocentrotus purpuratus* (spur 01) (spur 5.0)." Accessed 12 Aug 2020. Available from: `https://metazoa.ensembl.org/Strongylocentrotus_purpuratus`.

[176] C. Lesueur, "Descriptions of several new species of holothuria," *Journal of the Academy of Natural Sciences of Philadelphia*, vol. 4, no. Part I, pp. 155–163, 1824.

[177] F. R. Kille, "Regeneration in *Thyone briareus* lesueur following induced autotomy," *The Biological Bulletin*, vol. 69, no. 1, pp. 82–108, 1935.

[178] F. R. Kille, "Regeneration of gonad tubules following extirpation in the sea-cucumber, *Thyone briareus* (lesueur)," *The Biological Bulletin*, vol. 76, no. 1, pp. 70–79, 1939.

[179] K. D. Walton, J. C. Croce, T. D. Glenn, S.-Y. Wu, and D. R. McClay, "Genomics and expression profiles of the *Hedgehog* and notch signaling pathways in sea urchin development," *Developmental biology*, vol. 300, no. 1, pp. 153–164, 2006.

[180] M. Ehebauer, P. Hayward, and A. Martinez-Arias, "Notch signaling pathway," *Science's STKE*, vol. 2006, no. 364, pp. cm7–cm7, 2006.

[181] E. M. Erkenbrack, "Notch -mediated lateral inhibition is an evolutionarily conserved mechanism patterning the ectoderm in echinoids," *Development genes and evolution*, vol. 228, pp. 1–11, 2018.

[182] B. Lloyd-Lewis, P. Mourikis, and S. Fre, "Notch signalling: sensor and instructor of the microenvironment to coordinate cell fate and organ morphogenesis," *Current opinion in cell biology*, vol. 61, pp. 16–23, 2019.

[183] N. A. Auger, J. G. Medina-Feliciano, D. J. Quispe-Parra, S. Colón-Marrero, H. Ortiz-Zuazaga, and J. E. García-Arrarás, "Characterization and expression of holothurian wnt signaling genes during adult intestinal organogenesis," *Genes*, vol. 14, no. 2, p. 309, 2023.

[184] S. Fre, S. Pallavi, M. Huyghe, M. Laé, K.-P. Janssen, S. Robine, S. Artavanis-Tsakonas, and D. Louvard, "Notch and wnt signals cooperatively control cell proliferation and tumorigenesis in the intestine," *Proceedings of the National Academy of Sciences*, vol. 106, no. 15, pp. 6309–6314, 2009.

[185] S. Fre, M. Huyghe, P. Mourikis, S. Robine, D. Louvard, and S. Artavanis-Tsakonas, "Notch signals control the fate of immature progenitor cells in the intestine," *Nature*, vol. 435, no. 7044, pp. 964–968, 2005.

[186] J. H. van Es, M. E. Van Gijn, O. Riccio, M. Van Den Born, M. Vooijs, H. Begthel, M. Cozijnsen, S. Robine, D. J. Winton, F. Radtke, *et al.*, "*Notch/γ-secretase* inhibition turns proliferative cells in intestinal crypts and adenomas into goblet cells," *Nature*, vol. 435, no. 7044, pp. 959–963, 2005.

[187] S. Münder, S. Tischer, M. Grundhuber, N. Büchels, N. Bruckmeier, S. Eckert, C. A. Seefeldt, A. Prexl, T. Käsbauer, and A. Böttger, "Notch -signalling is required for head regeneration and tentacle patterning in *Hydra*," *Developmental biology*, vol. 383, no. 1, pp. 146–157, 2013.

[188] M. Broun, L. Gee, B. Reinhardt, and H. R. Bode, "Formation of the head organizer in *Hydra* involves the canonical wnt pathway," 2005.

[189] T. W. Parker and K. L. Neufeld, "*APC* controls *Wnt*-induced *β-catenin* destruction complex recruitment in human colonocytes," *Scientific reports*, vol. 10, no. 1, p. 2957, 2020.

[190] K. W. Kinzler and B. Vogelstein, "Lessons from hereditary colorectal cancer," *Cell*, vol. 87, no. 2, pp. 159–170, 1996.

[191] O. J. Sansom, K. R. Reed, A. J. Hayes, H. Ireland, H. Brinkmann, I. P. Newton, E. Batlle, P. Simon-Assmann, H. Clevers, I. S. Nathke, *et al.*, "Loss of *Apc* in vivo immediately perturbs wnt signaling, differentiation, and migration," *Genes & development*, vol. 18, no. 12, pp. 1385–1390, 2004.

[192] H.-A. Kim, B.-K. Koo, J.-H. Cho, Y.-Y. Kim, J. Seong, H. J. Chang, Y. M. Oh, D. E. Stange, J.-G. Park, D. Hwang, *et al.*, "*Notch1* counteracts *Wnt/β-catenin* signaling through chromatin modification in colorectal cancer," *The Journal of clinical investigation*, vol. 122, no. 9, pp. 3248–3259, 2012.

[193] A. Marchler-Bauer and S. H. Bryant, "Cd-search: protein domain annotations on the fly," *Nucleic acids research*, vol. 32, no. suppl_2, pp. W327–W331, 2004.

[194] A. Marchler-Bauer, S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, *et al.*, "Cdd: a conserved domain database for the functional annotation of proteins," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D225–D229, 2010.

[195] A. Marchler-Bauer, Y. Bo, L. Han, J. He, C. J. Lanczycki, S. Lu, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, *et al.*, "Cdd/sparcle: functional classification of proteins via subfamily domain architectures," *Nucleic acids research*, vol. 45, no. D1, pp. D200–D203, 2017.

[196] S. Lu, J. Wang, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, G. H. Marchler, J. S. Song, *et al.*, "Cdd/sparcle: the conserved domain database in 2020," *Nucleic acids research*, vol. 48, no. D1, pp. D265–D268, 2020.

[197] R. Kopan and M. X. G. Ilagan, "The canonical notch signaling pathway: unfolding the activation mechanism," *Cell*, vol. 137, no. 2, pp. 216–233, 2009.

[198] Y. Zhou, J. B. Atkins, S. B. Rompani, D. L. Bancescu, P. H. Petersen, H. Tang, K. Zou, S. B. Stewart, and W. Zhong, "The mammalian golgi regulates numb signaling in asymmetric cell division by releasing acbd3 during mitosis," *Cell*, vol. 129, no. 1, pp. 163–178, 2007.

[199] T. Sakata, H. Sakaguchi, L. Tsuda, A. Higashitani, T. Aigaki, K. Matsuno, and S. Hayashi, "*Drosophila Nedd4* regulates endocytosis of notch and suppresses its ligand-independent activation," *Current biology*, vol. 14, no. 24, pp. 2228–2236, 2004.

[200] G. Van Tetering and M. Vooijs, "Proteolytic cleavage of notch :"hit"' and "run"," *Current molecular medicine*, vol. 11, no. 4, pp. 255–269, 2011.

[201] N. Teider, D. K. Scott, A. Neiss, S. D. Weeraratne, V. M. Amani, Y. Wang, V. E. Marquez, Y.-J. Cho, and S. L. Pomeroy, "*Neuralized1* causes apoptosis and downregulates notch target genes in medulloblastoma," *Neuro-oncology*, vol. 12, no. 12, pp. 1244–1256, 2010.

[202] T. Iso, L. Kedes, and Y. Hamamori, "*HES* and *HERP* families: multiple effectors of the notch signaling pathway," *Journal of cellular physiology*, vol. 194, no. 3, pp. 237–255, 2003.

[203] P. W. Young, "*LNX1/LNX2* proteins: Functions in neuronal signalling and beyond," *Neuronal signaling*, vol. 2, no. 2, 2018.

[204] M. Kitagawa, "Notch signalling in the nucleus: roles of *Mastermind*-like (maml) transcriptional coactivators," *The journal of biochemistry*, vol. 159, no. 3, pp. 287–294, 2016.

[205] B.-K. Koo, K.-J. Yoon, K.-W. Yoo, H.-S. Lim, R. Song, J.-H. So, C.-H. Kim, and Y.-Y. Kong, "*Mind bomb-2* is an e3 ligase for notch ligand," *Journal of Biological Chemistry*, vol. 280, no. 23, pp. 22335–22342, 2005.

[206] G. S. Kim, H.-S. Park, and Y. C. Lee, "Opthis identifies the molecular basis of the direct interaction between csl and smrt corepressor," *Molecules and Cells*, vol. 41, no. 9, p. 842, 2018.

[207] K. Jin, W. Zhou, X. Han, Z. Wang, B. Li, S. Jeffries, W. Tao, D. J. Robbins, and A. J. Capobianco, "Acetylation of *Mastermind*-like 1 by *p300* drives the recruitment of *NACK* to initiate notch-dependent *Transcriptionp300* mediates *NACK* recruitment to the notch complex," *Cancer Research*, vol. 77, no. 16, pp. 4228–4237, 2017.

[208] K. L. Weaver, M.-C. Alves-Guerra, K. Jin, Z. Wang, X. Han, P. Ranganathan, X. Zhu, T. DaSilva, W. Liu, F. Ratti, *et al.*, "*NACK* is an integral component of the notch transcriptional activation complex and is critical for development and tumorigenesis*NACK* is a novel coactivator essential for notch function," *Cancer research*, vol. 74, no. 17, pp. 4741–4751, 2014.

[209] E. Koutelou, S. Sato, C. Tomomori-Sato, L. Florens, S. K. Swanson, M. P. Washburn, M. Kokkinaki, R. C. Conaway, J. W. Conaway, and N. K. Moschonas, "*Neuralized-like* 1 (*Neurl1*) targeted to the plasma membrane by n-myristoylation regulates the notch ligand *Jagged1*," *Journal of Biological Chemistry*, vol. 283, no. 7, pp. 3846–3853, 2008.

[210] S. Cormier, S. Le Bras, C. Souilhol, S. Vandormael-Pournin, B. Durand, C. Babinet, P. Baldacci, and M. Cohen-Tannoudji, "The murine ortholog of *Notchless*, a direct regulator of the notch pathway in *Drosophila melanogaster*,

is essential for survival of inner cell mass cells," *Molecular and cellular biology*, vol. 26, no. 9, pp. 3541–3549, 2006.

[211] A. E. Wallberg, K. Pedersen, U. Lendahl, and R. G. Roeder, "*p300* and *PCAF* act cooperatively to mediate transcriptional activation from chromatin templates by notch intracellular domains *in vitro*," *Molecular and cellular biology*, vol. 22, no. 22, pp. 7812–7819, 2002.

[212] S. Shi and P. Stanley, "*Protein O-fucosyltransferase 1* is an essential component of notch signaling pathways," *Proceedings of the National Academy of Sciences*, vol. 100, no. 9, pp. 5234–5239, 2003.

[213] P. Taylor, H. Takeuchi, D. Sheppard, C. Chillakuri, S. M. Lea, R. S. Haltiwanger, and P. A. Handford, "*Fringe*-mediated extension of o-linked fucose in the ligand-binding region of *Notch1* increases binding to mammalian notch ligands," *Proceedings of the National Academy of Sciences*, vol. 111, no. 20, pp. 7290–7295, 2014.

[214] B. Mao, W. Wu, G. Davidson, J. Marhold, M. Li, B. M. Mechler, H. Delius, D. Hoppe, P. Stannek, C. Walter, *et al.*, "*Kremen* proteins are *Dickkopf* receptors that regulate *Wnt/β-catenin* signalling," *Nature*, vol. 417, no. 6889, pp. 664–667, 2002.

[215] O. Gerlitz and K. Basler, "*Wingful*, an extracellular feedback inhibitor of *Wingless*," *Genes & development*, vol. 16, no. 9, pp. 1055–1059, 2002.

[216] D. MacGrogan, J. Münch, and J. L. de la Pompa, "Notch and interacting signalling pathways in cardiac development, disease, and regeneration," *Nature Reviews Cardiology*, vol. 15, no. 11, pp. 685–704, 2018.

[217] P. Hayward, K. Brennan, P. Sanders, T. Balayo, R. DasGupta, N. Perrimon, and A. M. Arias, "Notch modulates wnt signaling by associating with *Armadillo/β-catenin* and regulating its transcriptional activity," 2005.

[218] R. Zhang, A. Engler, and V. Taylor, "Notch : an interactive player in neurogenesis and disease," *Cell and tissue research*, vol. 371, pp. 73–89, 2018.

[219] H. M. Harris and C. Hill, "A place for viruses on the tree of life," *Frontiers in Microbiology*, vol. 11, p. 604048, 2021.

[220] M. Guo and C. Li, "Current progress on identification of virus pathogens and the antiviral effectors in echinoderms," *Developmental & Comparative Immunology*, vol. 116, p. 103912, 2021.

[221] I. Hewson, J. B. Button, B. M. Gudenkauf, B. Miner, A. L. Newton, J. K. Gaydos, J. Wynne, C. L. Groves, G. Hendler, M. Murray, *et al.*, "Densovirus associated with sea-star wasting disease and mass mortality," *Proceedings of the National Academy of Sciences*, vol. 111, no. 48, pp. 17278–17283, 2014.

[222] H. Liu, F. Zheng, X. Sun, X. Hong, S. Dong, B. Wang, X. Tang, and Y. Wang, "Identification of the pathogens associated with skin ulceration and peristome tumescence in cultured sea cucumbers *Apostichopus japonicus* (selenka)," *Journal of Invertebrate Pathology*, vol. 105, no. 3, pp. 236–242, 2010.

[223] P. Wang, Y. Chang, J. Yu, C. Li, and G. Xu, "Acute peristome edema disease in juvenile and adult sea cucumbers *Apostichopus japonicus* (selenka) reared in north china," *Journal of Invertebrate Pathology*, vol. 96, no. 1, pp. 11–17, 2007.

[224] S. D. Fugmann, C. Messier, L. A. Novack, R. A. Cameron, and J. P. Rast, "An ancient evolutionary origin of the *Rag1/2* gene locus," *Proceedings of the National Academy of Sciences*, vol. 103, no. 10, pp. 3728–3733, 2006.

[225] L. M. Carmona and D. G. Schatz, "New insights into the evolutionary origins of the recombination-activating gene proteins and v (d) j recombination," *The FEBS journal*, vol. 284, no. 11, pp. 1590–1605, 2017.

[226] D. G. Schatz, M. A. Oettinger, and D. Baltimore, "The v (d) j recombination activating gene, *RAG-1*," *Cell*, vol. 59, no. 6, pp. 1035–1048, 1989.

[227] D. G. Schatz, "Antigen receptor genes and the evolution of a recombinase," in *Seminars in immunology*, vol. 16, pp. 245–256, Elsevier, 2004.

[228] H. Ru, M. G. Chambers, T.-M. Fu, A. B. Tong, M. Liao, and H. Wu, "Molecular mechanism of v (d) j recombination from synaptic *RAG1-RAG2* complex structures," *Cell*, vol. 163, no. 5, pp. 1138–1152, 2015.

[229] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[230] S. Nurk, A. Bankevich, D. Antipov, A. Gurevich, A. Korobeynikov, A. Lapidus, A. Prjibelsky, A. Pyshkin, A. Sirotkin, Y. Sirotkin, *et al.*, "Assembling genomes and mini-metagenomes from highly chimeric reads," in *Research in Computational Molecular Biology: 17th Annual International Conference, RECOMB 2013, Beijing, China, April 7-10, 2013. Proceedings 17*, pp. 158–170, Springer, 2013.

[231] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "metaspades: a new versatile metagenomic assembler," *Genome research*, vol. 27, no. 5, pp. 824–834, 2017.

[232] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "Quast: quality assessment tool for genome assemblies," *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, 2013.

[233] A. Babaian and R. Edgar, "*Ribovirus* classification by a polymerase barcode sequence," *PeerJ*, vol. 10, p. e14055, 2022.

[234] R. C. Edgar, J. Taylor, V. Lin, T. Altman, P. Barbera, D. Meleshko, D. Lohr, G. Novakovsky, B. Buchfink, B. Al-Shayeb, *et al.*, "Petabase-scale sequence alignment catalyses viral discovery," *Nature*, vol. 602, no. 7895, pp. 142–147, 2022.

[235] S. Roux, F. Enault, B. L. Hurwitz, and M. B. Sullivan, "Virsorter: mining viral signal from microbial genomic data," *PeerJ*, vol. 3, p. e985, 2015.

[236] H. Hauswedell, J. Singer, and K. Reinert, "Lambda: the local aligner for massive biological data," *Bioinformatics*, vol. 30, no. 17, pp. i349–i355, 2014.

[237] E. Sayers, "A general introduction to the e-utilities," *Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US)*, 2010.

[238] R. D. Finn, J. Clements, and S. R. Eddy, "Hmmer web server: interactive sequence similarity searching," *Nucleic acids research*, vol. 39, no. suppl_2, pp. W29–W37, 2011.

[239] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata, "Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform," *Nucleic acids research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[240] K. Katoh, K.-i. Kuma, H. Toh, and T. Miyata, "Mafft version 5: improvement in accuracy of multiple sequence alignment," *Nucleic acids research*, vol. 33, no. 2, pp. 511–518, 2005.

[241] K. Katoh and H. Toh, "Recent developments in the mafft multiple sequence alignment program," *Briefings in bioinformatics*, vol. 9, no. 4, pp. 286–298, 2008.

[242] S. R. Eddy, "Profile hidden markov models.," *Bioinformatics (Oxford, England)*, vol. 14, no. 9, pp. 755–763, 1998.

[243] S. R. Eddy, "Accelerated profile hmm searches," *PLoS computational biology*, vol. 7, no. 10, p. e1002195, 2011.

[244] B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, and R. Lanfear, "Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era," *Molecular biology and evolution*, vol. 37, no. 5, pp. 1530–1534, 2020.

[245] R. Lanfear, B. Calcott, S. Y. Ho, and S. Guindon, "Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses," *Molecular biology and evolution*, vol. 29, no. 6, pp. 1695–1701, 2012.

[246] M. Gouy, S. Guindon, and O. Gascuel, "Seaview version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building," *Molecular biology and evolution*, vol. 27, no. 2, pp. 221–224, 2010.

[247] F. W. Alt, G. Rathbun, E. Oltz, G. Taccioli, and Y. Shinkai, "Function and control of recombination-activating gene activity.," *Annals of the New York Academy of Sciences*, vol. 651, pp. 277–294, 1992.

[248] L. D. Notarangelo, M.-S. Kim, J. E. Walter, and Y. N. Lee, "Human rag mutations: biochemistry and clinical implications," *Nature Reviews Immunology*, vol. 16, no. 4, pp. 234–246, 2016.