Računalna analiza horizontalnog prijenosa gena u spužvama Suberites domuncula i Eunapius subterraneus

Buršić, Luka

Master's thesis / Diplomski rad

2025

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Luka Buršić

Računalna analiza horizontalnog prijenosa gena u spužvama *Suberites domuncula* i *Eunapius subterraneus*

Diplomski rad

Zagreb, 2025.

University of Zagreb
Faculty of Science
Department of Biology

Luka Buršić

Computational analysis of horizontal gene transfer in sponge species *Suberites domuncula* and *Eunapius subterraneus*

Master thesis

Zagreb, 2025.

Ovaj rad je izrađen u Grupi za bioinformatiku na Zavodu za molekularnu biologiju Prirodoslovno-matematičkog fakulteta u Zagrebu, pod mentorstvom prof. dr.sc. Kristiana Vlahovičeka. Rad je predan na ocjenu Biološkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu radi stjecanja zvanja magistar molekularne biologije.

ZAHVALA

Prvo bih htio zahvaliti mentoru prof. dr. sc. Kristianu Vlahovičeku što mi je omogućio izradu diplomskog rada unutar svoje istraživačke grupe.

Veliko, veliko hvala ide mom komentoru Kristianu Bodulić́u, koji nažalost nije mogao postati i službeno komentor zbog, po mom mišljenju, suludog pravila. Molim sve koji ovo čitaju i mogu utjecati na tu odluku da ubuduće uključe asistente koji pomažu u izradi diplomskih radova, jer ovo nije prvi put, a nažalost nije ni zadnji put da asistenti utroše veliku količinu svog vremena na nas studente i ne dobiju ništa zauzvrat. Bez Kristiana ovaj diplomski ne bi postojao. Hvala ti što si čitao sve splačine koje sam ti slao, hvala ti na nebrojenim razgovorima u 2 ujutro kad sam ti slao preružne grafove na pregled, na trenutcima kad si me izvlačio iz pakla zvanog *taxize* na koji smo izgubili 2 tjedna rada, na trenutke kad si me zvao da me smiriš jer sam bio na rubu živaca, i naposlijetku, hvala ti na svem vremenu koji si utrošio radeći sa mnom i znanju koje si mi prenio.

Zahvalio bih se i svojoj obitelji koja me školovala i podržavala u mom studiju, hvala im na strpljenju sa mnom i nadam se da ću i dalje uživati njihovo povjerenje u moje postupke.

Zahvalio bih se svojim prijateljima koji su uvijek bili tu za mene kad mi je trebalo.

Ivi, Ani, Emmi – mojim malim ženama – hvala što su bile glas razuma kad bih se izgubio u vlastitim deluzijama, nježni lahor savjeta u gluhoći vlastite tvrdoglavosti i meki oslonac kad bih klonuo.

Vitu, koji me spuštao na zemlju kad je to bilo potrebno i koji me prisjećao tko zapravo jesam i što mi je bitno u životu.

Dorianu, drago mi je da ne mogu reći ... *yeah, I did it myself*... jer to ne bi bila istina – bez njega ne bih bio tu gdje jesam. Dorian je vidio moje najblistavije i najtmurnije trenutke, trpio moje sinusoide raspoloženja i svejedno je bio tu. Čak i kad sam vidio razočaranje u njegovim očima, nije otišao svojim putem, već mi je pomogao da ponovno stanem na noge, makar to značilo zajedničko iščitavanje skripte na glas prolazeći prstom po redovima kao da tek učim čitati.

Hvala svim prijateljima koje sam stekao na faksu – Josipi, Kianu, Miji, Nikolini, Doni, Leu, Marti, Ivi, Luciji – hvala vam na planinarenjima, na viksama, na kavama, izlascima, na vrelim ljetima i sočnim pomidorima, hladnim planinskim proplancima i toplom bureku, zaklonu dok je u Zagrebu bjesnjela oluja.

Hvala curama s 4. kata, pogotovo Viki, na ljubavnim savjetima i diskutabilnim jelima.

Hvala Firiću na svim suprotstavljanjima svemu.

Hvala mojoj Palmi, predivnim ljudima i predivnoj glazbi.

Hvala ljudima iz Tehničkog muzeja i svoj jadnoj djeci koja su trpjela moja vodstva :)

I na kraju, hvala mom momku, koji me smirivao kad je to bilo potrebno.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Diplomski rad

Računalna analiza horizontalnog prijenosa gena u spužvama *Suberites domuncula* i *Eunapius subterraneus*

Luka Buršić

Rooseveltov trg 6, 10000 Zagreb, Hrvatska

Horizontalni transfer gena (HGT) omogućuje prijenos novih svojstava između evolucijski udaljenih skupina organizama. Iako je HGT opsežno istražen kod prokariota, kod rano razdvojenih skupina metazoa poput spužvi i žarnjaka, istraživanja su oskudna. Ovaj rad bavi se otkrivanjem i analizom HGT u dvije vrste spužvi: *Suberites domuncula*, kozmopolitskoj morskoj spužvi, i *Eunapius subterraneus,* hrvatskoj endemskoj špiljskoj spužvi. Unatoč strogim kriterijima, kombiniranjem različitih računalnih analiza genoma spužvi, otkriven je značajan broj horizontalno prenesenih gena (HTgs). Istraživane spužve razlikuju se prema razini prilagodbe HTgs – duljini introna, upotrebi kodona i grupiranju po očuvanosti. Ove razlike ukazuju na davnije dogadjaje horizontalnog prijenosa u pretku *S. domuncula* od onog u *E. subterranus*. Pronađeni HTgs pokazali su funkcijsku povezanost s regulacijom gena koji sudjeluju u metabolizmu DNA, što potvrđuje njihova očuvanost u srodnim spužvama. Nisu utvrđene funkcije u sekundarnom metabolizmu kao što je pokazano u sličnom istraživanju što može biti uzrokovano različitom kvalitetom početnog materijala i strožom analizom. Također, u radu se ističe potreba za boljim referentnim okvirima u istraživanju spužvi i doprinos istraživanja spužvi u evoluciji metazoa.

Ključne riječi: spužve, endosimbionti, arhitektura gena, horizontalni transfer gena, funkcionalna analiza
(53 stranice, 14 slika, 3 tablice, 91 literaturni navod, jezik izvornika: engleski)
Rad je pohranjen u Središnjoj biološkoj knjižnici

Mentor: prof. dr. sc. Kristian Vlahoviček

Ocjenitelji:
prof. dr. sc. Kristian Vlahoviček
izv. prof. dr. sc. Rosa Karlić
prof. dr. sc. Damjan Franjević

Rad prihvaćen: 6. veljače 2025.

BASIC DOCUMENTATION CARD

University of Zagreb
Faculty of Science
Department of Biology
Master thesis

Computational analysis of horizontal gene transfer in sponge species *Suberites domuncula* and *Eunapius subterraneus*

Luka Buršić

Rooseveltov trg 6, 10000 Zagreb, Croatia

Horizontal gene transfer (HGT) enables the transfer of new traits between evolutionarily distant groups of organisms. While HGT has been extensively studied in prokaryotes, research on early-diverging metazoan groups such as sponges and cnidarians remains scarce. This study focuses on the detection and analysis of HGT in two sponge species: *Suberites domuncula*, a cosmopolitan marine sponge, and *Eunapius subterraneus*, a Croatian endemic cave sponge. Despite strict criteria, a significant number of horizontally transferred genes (HTgs) were identified through the integration of various computational genome analyzes of the sponges. The studied sponges differ in terms of adaptation levels of HTgs – intron length, codon usage, and grouping by conservation. These differences indicate earlier events of horizontal transfer in the ancestor of *S. domuncula* compared to *E. subterranus*.The detected HTgs showed functional associations with the regulation of genes involved in DNA metabolism, supported by their conservation in related sponges. However, no functions related to secondary metabolism were identified, as reported in similar studies. This could be due to differences in the quality of initial material and stricter analyzes applied.The study also emphasizes the need for better reference frameworks in sponge research and highlights the contribution of sponge studies to the understanding of metazoan evolution..

Keywords: sponges, endosymbionts, gene architecture, horizontal gene transfer, functional analysis
(53 pages, 14 figures, 3 tables, 91 references, original in: english)
Thesis is deposited in Central Biological Library.

Mentor: Prof. Kristian Vlahoviček, PhD

Reviewers:
Prof. Kristian Vlahoviček, PhD
Assoc. Prof. Rosa Karlić, PhD
Prof. Damjan Franjević, PhD

Thesis accepted: February 6, 2025

CONTENTS

ABBREVIATION

aa – amino acids

BLAST – Basic Local Alignment Search Tool

BLAT – BLAST-Like Alignment Tool

bp – base pair

CDS – coding sequence

DIAMOND – Double Index Alignment of Next generation sequencing Data

DNA – deoxyribonucleic acid

dsDNA – double stranded DNA

GO – gene ontology

HGT – horizontal gene transfer

HTgs – horizontally transferred genes

kb – kilobase

LCA – lowest common ancestor

MEGAN – MEtaGenome Analyzer

non-HTgs – non-horizontally transferred genes

ONT – Oxford Nanopore Technology

RNA – ribonucleic acid

RNAseq – RNA sequencing

ssDNA – single-stranded DNA

1. INTRODUCTION

1.1. Gene transfer

Each daughter cell originates from its mother cell, inheriting cellular components – most notably, its DNA molecules. Mother cells and daughter cells represent different generations. The transfer of genetic information to the next generation is called vertical transfer or inheritance, in the sense of asexual or sexual reproduction. Asexual reproduction requires only one parent, but the mechanisms vary greatly. Thus, offspring originating from asexual reproduction may be completely identical to its parent, or its genotype may be a mosaic of the parental genotypes achieved through recombination. The most extreme form of asexual reproduction is automictic parthenogenesis, which represents the development from a haploid cell originating from meiosis that fuses with another haploid cell of the same origin. In this scenario, the genotype is entirely inherited from the parent, without the addition of foreign DNA. However, due to recombination of the parental genome, offspring differs in phenotype. Notably, these extreme cases are rare (**Engelstädter 2017**). On the other hand, sexual reproduction increases the diversity of offspring compared to asexual reproduction. In addition to recombination, chromosome segregation, random gamete fusion and avoiding closely related partners prevents consequences of the so-called "inbreeding depression" (**Gao et al. 2015**).

Maintaining the balance between inheriting parental traits and variability is important for survival in the context of preserving features acquired through evolution. However, slight continuous change is crucial for survival in varying conditions. Organisms that reproduce both sexually and asexually can directly influence this balance. Therefore, under unfavorable conditions, the rate of sexual reproduction increases, leading to a greater variability between generations and faster adaptation to new conditions (**Arnqvist & Rowe 2005**). However, asexually reproductive species are limited in acquiring new genetic material, which cannot be overcome exclusively with mutations. Horizontal gene transfer is a trait transfer between individuals of the same generation by acquisition of foreign DNA. It increases the population diversity rate required for survival in changing conditions. This predominantly occurs in prokaryotes, where horizontal gene transfer (HGT) has first been discovered, but accumulating evidence indicate an important role of interdomain HGT and HGT among eukaryotes.

1.2. Horizontal gene transfer

1.2.1. Horizontal gene transfer in prokaryotes

Horizontal gene transfer was first discovered through experiments that characterized DNA as the carrier of hereditary information (**Griffith 1928**; **Avery et al. 1944**), and through research in field of bacterial genetics (**Lederberg & Tatum 1946**; **Zinder & Lederberg 1952**). Three main mechanisms of HGT are well-studied in prokaryotes: transformation, conjugation, and transduction.

In 1928, Griffith showed that a suspension of dead pathogenic S (smooth) strain of the bacterium *Streptococcus pneumoniae*, which forms smooth bacterial colonies on a nutrient medium, could transform a non-pathogenic R (rough) strain into an S strain. Using the dead S strain ensured that the cause was not the overgrowth of the R strain by the S strain, but rather transformation (**Griffith 1928**). Avery continued the experiments in 1944, treating the suspension of the dead S strain bacteria with enzymes and confirmed that the transforming effect was absent when deoxyribonucleases (DNases) were applied, thus proving that DNA molecules hold hereditary information (**Avery et al. 1944**). This was the first evidence of transformation – HGT in which an external free DNA molecule originating from the environment alters the phenotype of the organism. Bacterial transformation requires competence – a physiological state in which a group of genes encoding proteins involved in recognition, capture, processing, internalization, and integration of foreign DNA are activated. A special type of pili non-specifically binds foreign double-stranded linear DNA (dsDNA). This triggers degradation of one strand of DNA and insertion of single-stranded DNA (ssDNA) into the bacterium through protein complexes, known as secretion systems. Afterwards, proteins that associate with ssDNA and the evolutionarily conserved RecA protein initiate the process of homologous recombination, leading to the integration of the DNA (**Johnston et al. 2014**).

In a brief paper from 1946, Lederberg and Tatum presented the following observation: by combining pairs of different triple auxotrophic mutant bacteria *Escherichia coli*, complex mutation combinations were obtained more frequently than when mutants were cultured separately or treated with sterile filtrates (**Lederberg & Tatum 1946**). In this manner, revertant mutations and transformation with suspension DNA were excluded from consideration. Further research showed that the change is mediated through direct contact. This newly discovered phenomenon was initially thought to represent sexual reproduction in bacteria, and the existence of a zygote was hypothesized (**Lederberg & Tatum 1946**), which is now known to be incorrect. This was the first evidence of conjugation – horizontal gene transfer from one organism mediated by a complex protein structure, known as the conjugation bridge. Conjugation requires direct contact and mutual a plasmid that encodes all necessary proteins for conjugation – from recognition to repellence of non-compatible cells. After successful transition, foreign DNA can integrate or remain episome, although this requires an appropriate origin of replication (**Bergstrom et al. 2000**).

An experiment similar to Lederberg and Tatum's 1946 study, but on *Salmonella* mutants, discovered HGT which was independent of contact. That implied a filtrable agent was causing HGT (**Zinder & Lederberg 1952**). A series of experiments provided evidence for transduction – HGT from one organism mediated by a viral vector. Transduction requires a virus to insert DNA into its host. There are two main types of viral cycles: lytic and lysogenic. The lytic cycle is not a long-term state as it leads to depletion of cellular resources and cell lysis (**Zhang et al. 2022**). In the lysogenic cycle, the virus can remain integrated in the host genome through several generations in the form of a prophage. However, reactivation of the prophage can trigger the lytic cycle, most often under conditions unfavorable for the host. In this process, along with the viral genome, specific or non-specific parts of the bacterial genome may be found in viral particles, which are then delivered to another bacterium during the next round of infection (**Lee et al. 2017**). The inserted linear or circular dsDNA does not represent the viral genome, it cannot initiate the viral cycle, but can be preserved if integrated into the genome of the new host.

1.2.2. Horizontal gene transfer in eukaryotes

In theory, processes analogous to prokaryotes could be found in eukaryotes – acceptance of free DNA from the environment (transformation), plasmid transfer mediated by a conjugation bridge (conjugation), and transfer mediated by a virus (transduction).

The most studied form of eukaryotic HGT is plasmid transfer between eukaryotes and closely associated bacteria. The evidence of this has been found in genera *Agrobacterium*, *Escherichia* and *Bartonella*. All described transfers are mediated by type IV secretion system (T4SS), which suggests that the process is similar to bacterial conjugation. The most studied process is the insertion of T-DNA from *Agrobacterium* bacteria into plants. Under laboratory conditions, the range of recipients is broader and includes not only plant cells, but also fungi and even human cells (**Lacroix et al. 2006**). Another observed conjugation-like plasmid transfer process involves the transfer between *Escherichia coli* and various eukaryotes (**Lacroix & Citovsky 2016**).

Conjugation between prokaryotes and eukaryotes is not limited to the conjugation bridge and can be achieved via T4SS during endosymbiotic interaction. The most significant and extensive gene transfer from endosymbionts to eukaryotes occurred through the development of mitochondria and plastids. This process is now defined separately from HGT as endosymbiotic gene transfer

(ETG) (**Archibald 2015**). A process similar to ETG is observed between intracellular parasite *Bartonella henselae* and its host.

While the stated examples demonstrate conjugation-like plasmid transfer to various hosts, all of them encompassed genetically engineered plasmids with the role of gene replication in the recipient; no natural shuttle plasmid has yet been found. That does not exclude conjugation as a possible mechanism of HGT to eukaryotes since this problem can be solved by integrating the transferred DNA into the host genome.

Transformation and transduction have been widely used as laboratory techniques for gene insertion into eukaryotic cells, called transfections. However, spontaneous transformation has so far been confirmed only in yeast under laboratory conditions similar to natural conditions, but the mechanism is unknown (**Mitrikeski 2013**). Furthermore, transformation after phagocytosis is another potential mechanism for HGT. It is mostly related to HGT in phagocytic amoebas due to the possibility of endosomal escape after phagocytosis (**Eichinger et al. 2005**).

For transduction from prokaryotes to eukaryotes, an interdomain virus would be required as most viruses are highly specific and have a limited host range. Although the existence of such a virus has not been directly shown, research on HGT between the intracellular symbiotic bacterium *Wolbachia* and the mosquito species *Aedes aegypti* found bacteriophage-derived sequences near genes transferred from the bacterium (**Klasson et al. 2009**). This could potentially be evidence of bacteriophages as a common interdomain vector. However, the possibility that a prophage sequence from the bacterium integrated along with other parts of the bacterial genome has not been excluded (**Hotopp et al. 2007**).

1.2.3. Intracellular immunity – hindering horizontal gene transfer

Horizontal gene transfer involves the transfer of genetic material between organisms of different species, unlike vertical transfer, which occurs through generations within the same species. Although HGT can be beneficial to the recipient cell, any foreign DNA poses a potential threat. In many cases, HGT can have a negative impact, directly causing recipient cell death or decreasing its fitness. As a consequence, those genes are consequently lost after several generations (**Brigulla & Wackernagel 2010**). For this reason, organisms develop defense mechanisms against foreign genetic material.

Prokaryotes developed various mechanisms for foreign DNA detection. Their intercellular immunity spans from non-specific to sequence-specific systems. These systems are well studied as they were fundamental for the development of genetic engineering techniques. One of the most well-known mechanisms in prokaryotes is the restriction-modification system, which provides non-specific immunity against foreign DNA. Cellular DNA is methylated on specific positions and restriction endonucleases degrade foreign DNA, which is typically not modified in the same way (**Arias et al. 2022**). The CRISPR-Cas system, unlike the restriction-modification system, is sequence-specific. Fragments of foreign DNA (spacers) are stored in a CRISPR locus that is transcribed upon foreign DNA detection. Transcribed RNA complementary binds to foreign DNA upon subsequent encounters and targets it for degradation with the Cas endonuclease. Considering that the number of stored spacers is evolutionarily limited, the importance of CRISPR-Cas in preventing HGT is unknown. However, CRISPR-Cas has been shown to significantly inhibit the transfer of plasmids carrying antibiotic resistance in bacteria, particularly in agriculture (**Upreti et al. 2024**).

Intercellular immunity has been mostly studied in mammalian cells as they are crucial for genetic engineering in eukaryotic systems, tumor immunotherapies, gene therapies, nucleic acid-based vaccines, and similar fields (**Kong et al. 2023**). Despite the fact that non-model eukaryotes are often not included in research on signaling pathways triggered by foreign DNA, studies in animal cells can highlight the difficulties foreign DNA faces when entering eukaryotic cells.

The intercellular immunity systems in mammal cells consists of non-specific cytosolic DNases like DNase III/TREX, lysosomal DNase II and cytosolic detector proteins such as GMP-AMP synthase (cGAS) and gamma-interferon-inducible protein (IFI16). Cytosolic DNA detectors trigger signal pathways, autophagy and degradation of DNA in lysosomes (**Anindya 2022**). In an intact eukaryotic cell, DNA is predominantly located in the organelles. As such, the presence of any DNA in the cytoplasm suggests infection or aberrations in the cell cycle and genome integrity. Cytosolic detector proteins bind to dsDNA independently of sequence, but dependent on length. Longer DNA fragments stabilize a nucleoprotein liquid-phase condensate with detector proteins required for efficiently triggering signal pathways. In eukaryotic cells, DNases and sensors that trigger signaling pathways must be in balance to prevent the accumulation of DNA or the negative consequences of an excessive immune response to foreign DNA. While compartmentalization has reduced the need to distinguish foreign from host DNA, methylation patterns and GC content also play an important role in intercellular immunity (**Kong et al. 2023**).

1.2.4. Integration – a way to safety for horizontally transferred genes

Integration offers protection to foreign DNA from intercellular immunity. It can occur through homologous recombination (HR) or non-homologous, illegitimate recombination (IR). Homologous recombination is a process that requires a long homologous region between the integrated DNA molecule and integration site (**Kowalczykowski et al. 1994**). Circular DNA, such as inserted plasmids after conjugation, can be integrated after a single HR event. However, integration of a linear DNA molecule, also a common product of horizontal transfer, requires two HR events with recipient DNA at two locations. Although guided by homology and not causing mutations on integration borders, integration of linear DNA via HR can replace the region between the borders. In prokaryotes, HR is the dominant method of integration.

Illegitimate recombination can be performed on regions with microhomology or completely non-homologous region, as is the case in non-homologous end-joining (NHEJ). In eukaryotes, NHEJ is the dominant method of integration, although this depends on experimental conditions, cell type and cell cycle. For example, in the S/G2 cell cycle phase, different variations of HR become the dominant method of integration as they are connected with DNA repair during DNA replication (**Kumari et al. 2024**). During the other cell cycle phases, NHEJ merges DNA molecules with short indels of variable length at the junction site, as exonucleases and polymerases process DNA ends (**Kumari et al. 2024**). The joining of free DNA ends with NHEJ is non-specific for nucleotide sequences. Consequently, NHEJ can merge any DNA unprotected with a telomeric complex, and thus circularize and concatenate foreign DNA prior to genome integration (**Würtele et al. 2003**). Short indels and described integration combinations are a hint for NHEJ integration.

In addition to prokaryotic and eukaryotic cellular integration mechanisms, viruses and transposons embedded in the genome encode enzymes for their own integration, replication, or assembly. Upon their activation, depending on the specificity of integration for each element, these enzymes can nonspecifically integrate foreign DNA (**Brigulla & Wackernagel 2010**).

Despite the protection from intracellular immunity, integration can hinder the retention of horizontally transferred genes (HTgs). Firstly, integration can occur in an essential functional gene or may replace an essential gene, resulting in a lethal phenotype (**Brigulla & Wackernagel 2010**). Furthermore, many integrated genes can be toxic to the recipient, such as genes resulting in overly abundant proteins (**Sorek et al. 2007**). Additionally, changes in the inserted DNA caused by the integration process can affect its function. Finally, the epigenetic system of the recipient can silence inserted genes due to different characteristics of the foreign gene, such as methylation patterns and GC base pair content (**Brigulla & Wackernagel 2010**).

1.3. Successful horizontal gene transfer – methods and proofs

To infer horizontal gene transfer (HGT) events, which may not necessarily lead to phenotypic changes, most modern approaches rely on the analysis of genomic sequence data. These methods can be categorized into two primary groups: parametric and phylogenetic approaches (**Ravenhall et al. 2015**).

Phylogenetic methods assess the evolutionary histories of genes to detect inconsistencies in phylogenies, which may indicate HGT. Phylogenetic approaches are further classified into explicit methods that reconstruct and compare phylogenetic trees and implicit methods that use surrogate measures instead of complete tree reconstruction (**Dessimoz et al. 2008**). Phylogenetic approaches leverage differences between gene evolution and species tree evolution, which result from HGT events. Explicit phylogenetic methods reconstruct gene trees and identify discrepancies that may indicate HGT, while implicit phylogenetic approaches analyze differences in pairwise distances between genes and their corresponding species without explicit tree construction.

Phylogenetic methods have benefited from the recent availability of a large number of sequenced genomes. Since these methods leverage comparative genomics, they provide a more detailed characterization of HGT events, including the identification of donor species and the estimated time of transfer. Despite these advantages, phylogenetic approaches have limitations. Conflicting phylogenies can arise due to factors such as unrecognized gene duplications followed by gene losses, which are not always accounted for in evolutionary models. Additionally, many phylogenetic methods depend on a reference species tree, which can be difficult to construct accurately. The computational cost of reconstructing multiple gene and species trees also presents a significant challenge, as large-scale phylogenetic analyses require substantial computational resources. Furthermore, because phylogenetic methods generally focus on genes or protein sequences as evolutionary units, they are limited in their ability to detect HGT events occurring in non-genic regions or across gene boundaries (**Ravenhall et al. 2015**).

Parametric methods identify genomic regions with significant deviations from the genomic norm, such as variations in guanine-cytosine (GC) content or codon usage patterns. Conceptually, parametric methods infer HGT by calculating a statistical metric, such as GC content, across sliding genomic windows and comparing it to the typical range for the entire genome. Genomic regions with atypical values are inferred as potential HGT candidates.

A key advantage of parametric methods is their ability to function without requiring comparative genomic data. This was particularly beneficial in the early days of genome sequencing when few closely related genomes were available. However, these methods rely on the assumption that the host genome exhibits a uniform compositional signature. This limitation can lead to overpredictions if natural intragenomic variability is mistaken for HGT (**Guindon & Perriere 2001**). Additionally, for parametric methods to accurately detect HGT, the transferred genetic material must retain the compositional signature of the donor and be distinguishable from the recipient. Over time, however, transferred sequences undergo mutational changes similar to the rest of the host genome, a process known as amelioration, which slowly eliminates compositional differences and reduces the detectability of ancient HGT events (**Zhu et al. 2014**).

Given the strengths and weaknesses of phylogenetic and parametric methods, combining both approaches can provide a more comprehensive set of HGT candidate genes. Integrating different parametric methods significantly improves the quality of HGT predictions. Moreover, in the absence of a definitive set of known horizontally transferred genes, discrepancies between different HGT detection methods could be reconciled by using an integrative approach that combines parametric and phylogenetic evidence. However, while combining multiple methods can enhance sensitivity, it also increases the risk of false positives so careful validation and methodological refinement are necessary to achieve reliable HGT detection.

Computational methods have so far identified many HGT events. However, donors are rarely unambiguously defined. The most renowned example is the transformation of plant with T-DNA from *A. tumefaciens* (**Lacroix & Citovsky 2016**)*.* Among unicellular eukaryotes, *Galdieria sulphuraria*, a red alga thriving in extreme environments, has incorporated up to 5% of its protein-coding genes from bacterial sources. These genes, encoding functions like arsenic resistance, likely facilitated adaptation to its harsh habitat (**Schönknecht et al. 2013**). Likewise, *Dictyostelium discoideum* and *Blastocystis* species harbor bacterial genes that confer ecological and functional advantages, demonstrating the widespread impact of HGT on eukaryotic adaptation (**Eme et al. 2017**). In animals, evidence of HGT is rarer but includes transfers from endosymbiotic bacteria. For instance, the genome of the coffee berry borer beetle (*Hypothenemus hampei*) contains a mannanase gene from *Bacillus*, enabling it to digest coffee beans effectively (**Acuña et al. 2012**). Moreover, HGT from *Wolbachia*, an endosymbiotic bacterium, has contributed substantial genetic material to arthropod and nematode hosts, underscoring the ecological and evolutionary relevance of these events (**Hotopp et al. 2007**).

Several studies have been analyzing eukaryotes on the brink of multicellularity. Main models for studying the development of multicellularity in Metazoa are choanoflagellates, unicellular organisms that can live in colonies (**Yue et al. 2013**). On the other side, sponges have been rising interest as one of the first animals to branch from the last common animal ancestor. Unfortunately, only one comprehensive research, on sponge *Amphimedon queenslandica*, has been done on sponge genome regarding HGT.

1.4. Sponges

Porifera, commonly known as sponges, represent one of the oldest animal clades in the evolution of Metazoa (multicellular animals). Molecular studies indicate that sponges share a close evolutionary relationship with choanoflagellates, a group of unicellular organisms considered the closest living relatives of animals (**Brunet and King 2017**). Evolution of multicellular organisms is only one of the many reasons why sponges are interesting to explore.

1.4.1. Sponge morphology

Sponges possess highly specialized cells and mechanisms that enable them to survive in various conditions. However, sponge cells are not organized into tissues, which results in a simple morphology. They exhibit a variety of body forms, all sharing a basic structure characterized by a porous, hollow, asymmetrical body. The main feature of their morphology is the presence of a central cavity (spongocoel), numerous tiny pores (ostia) that connect it with its surroundings and one larger opening (oscula). Sponges often reproduce asexually through budding and remain attached to one another. In some cases, the newly formed individual buds remain physically connected to the parent sponge, forming colonial structures. These colonies may share a common spongocoel and have coordinated biological processes, such as water filtration and feeding. The number of each individual can be distinguished by the number of oscula (**Matoničkin et al. 1998**).

The main function of sponges' body plan is water filtration, which determines various physiological processes like feeding, respiration, and excretion, while inferring ecological interactions. Constant flow of water is actively kept, so that water is taken in through numerous ostia, while the osculum allows water to exit (**Matoničkin et al. 1998**).

Sponges are organized into three distinct layers of cells. The innermost layer is composed of flagellated cells called choanocytes, responsible for the previously described water flow. Choanocytes use their flagella to create water current and capture food particles, which are then absorbed by the cells (**Matoničkin et al. 1998**). Although their main function is water filtration and feeding, they can also differentiate into sperm cells. In many species, they undergo a transformation where the flagellum is lost, and the cell becomes a spermatozoid (**Bergquist 2001**). The outermost sponge layer consists of flattened cells known as pinacocytes. Pinacocytes form the pinacoderm which serves as a protective barrier (**Matoničkin et al. 1998**). Between pinacoderm and choanocytes lies the mesohyl, a gelatinous matrix that contains various cell types and skeletal elements such as spicules or spongin fibers. The sponge skeleton provides structural support and can be composed of a variety of materials, including siliceous spicules, calcareous spicules, and spongin fibers. The skeletal elements are essential for maintaining the shape of the sponge and vary greatly among species. The most prominent cell types found in mesohyl are amoebocytes that carry a great variety of tasks like aiding in nutrient digestion and transport, immunity, interaction with symbionts, asexual and sexual reproduction etc. Amoebocytes have the ability to differentiate into various types of cells, enabling them to take part in regeneration and repair of damaged body parts. Furthermore, amoebocytes can dedifferentiate and start the process of budding or gemmulogenesis for asexual reproduction. During sexual reproduction in some sponges, amoebocytes are the primary precursors for the formation of oocytes, while in others they can act as a reserve source for gamete production (**Ereskovsky et al. 2024**). Since amoebocytes are involved both in reproduction and symbiont interaction, there is an obvious implication that this condition can increase the possibility of HGT.

1.4.2. Sponge evolution

The evolution of sponges is a complex and debated topic, with sponges being one of the earliest diverging metazoan lineages. Molecular and morphological data suggest that sponges are among the first animals to have diverged from a common ancestor of all metazoans, making them a critical group for understanding the origin of multicellularity and the evolution of animal life (**Erwin 2015**). The phylogenetic tree of sponges is divided into four classes, which form two monophyletic clades: Calcarea + Homoscleromorpha, and Hexactinellida + Demospongiae. Calcarea, characterized by the presence of calcium carbonate spicules, are considered to have diverged early in sponge evolution due to their simple body structure and the absence of specialized cells. Homoscleromorpha, in contrast, are characterized by a more tissue-like cell organization since their choanoderm and pinacoderm develop adherent junctions and basement membrane. Hexactinellida, characterized by their silica-based spicules, form a lattice-like structure – a distinctive morphology that makes them important for understanding sponge evolution and the development of skeletal elements. Hexactinellida are generally found at greater depths, which could explain their unique and intricate skeletal structures. Demospongiae is the largest and most diverse class of sponges, including species with siliceous spicules and a spongin protein skeleton. They represent the dominant group of sponges in most marine and freshwater ecosystems and have evolved a variety of forms and functions, including both sexual and asexual reproduction (**Ereskovsky et al. 2024**).

1.4.3. Sponge impact on ecology

Porous structure of the sponge body creates complex habitats that support biodiversity by providing shelter for various marine and freshwater organisms, including small fish, crustaceans, and other invertebrates (**Matoničkin et al. 1998**). As already mentioned, amoebocytes are responsible for interaction with mostly unicellular symbionts. Furthermore, sponges maintain water quality by filtering large volumes of water to feed on organic particles, bacteria, and plankton, thus ensuring that water remains clear, which is essential for photosynthetic organisms and the stability and health of ecosystems. In doing so, they reduce turbidity and promote nutrient and biochemical cycling, such as carbon and nitrogen cycling (**Bell 2008**).

1.4.4. Sponge species

Suberites domuncula is a marine sponge species belonging to the class *Demospongiae*. It is typically found in temperate and subtropical marine environments, at various depths, often forming dense mats or lobed structures. *S. domuncula* can reproduce both sexually, through spawning, and asexually, through budding or the formation of gemmules, which are resistant to harsh conditions. This species is also known for producing bioactive compounds, which have been studied for potential medicinal properties, such as antimicrobial and anticancer activity (**Souiba et al. 2023**). Due to its relatively simple structure and ease of cultivation in laboratory settings, *S. domuncula* is often used as a model organism in marine biology research (**Müller et al. 1999**).

Eunapius subterraneus is the only known subterranean freshwater sponge in the world. It belongs to the class Demospongiae and it is endemic to cave systems of the karst region near Ogulin, Croatia. Sponges grow on rocks, typically on sheltered spots away from strong water currents, in shallow karst areas where groundwater is rich in nutrients. It lives attached to rocks in complete darkness, although it can tolerate small amounts of light. Its lack of pigmentation, slow metabolism, altered cell physiology, and reproductive changes distinguish it from its surface relatives. This species can give rise to two forms of the habitus: plate-like and egg-like habitus (**Bedek et al. 2008**). Genomic and phylogenetic studies based on mitochondrial DNA have shown that, contrary to expectations, *E. subterraneus* is more closely related to other freshwater species than other species of the *Eunapius* genus (**Harcet et al. 2010**), rendering the exact classification unresolved. Additional morphological features, such as gemmule structure and ostia construction, differ significantly from other *Eunapius* species, which could indicate an adaptation to the stygobiont lifestyle or a misclassification of the species within the genus (**Bilandžija et al. 2007**).

1.5. Thesis objective

The objectives of this thesis will be to detect HGT candidates in the genomes of *Suberites domuncula* and *Eunapius subterraneus* using a wide range of computational biology methods. Additionally, genes identified as HGT candidates will be further analyzed in order to determine their potential function and assess their utility in the analyzed sponge genome. Insights into the function and benefits of these genes will be used to explain their horizontal transfer in the context of sponge evolution.

2. MATERIALS

2.1. Genomes

I used genomes of two sponge species: *Suberites domuncula*, and *Eunapius subterraneus*. The genomes were assembled at the Bioinformatics group at the Faculty of Science, University of Zagreb. The scaffold length distribution is shown in **Figure 1.**

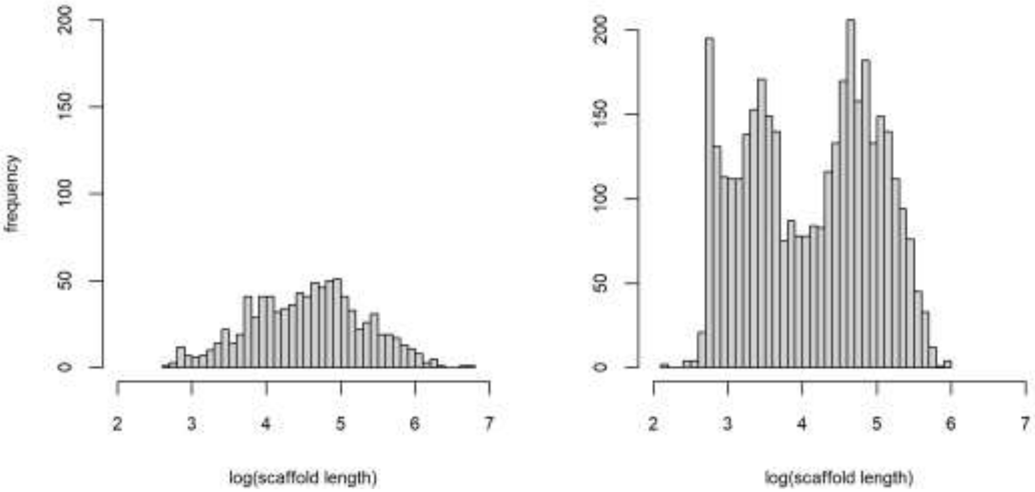


Figure 1. Scaffold length distribution for sponge species used in this thesis. *S. domuncula* (left, N = 900), *E. subterraneus* (right, N = 3,694).

2.2. Prokaryotic database

The prokaryotic database used in this thesis was provided by **Mukherjee et al. (2017)**. The database consists of 1003 genomes and proteomes that emerged as a result of the Genomic Encyclopedia of Bacteria and Archaea initiative (GEBA-I), selected to increase coverage of a wider phylogenetic space. GenBank accession numbers corresponding to all 1003 GEBA-I genomes were provided in Supplementary Table 1. in **Mukherjee et al. (2017)**. I constructed a bash script to retrieve NCBI Genomes accession number for each genome and its coding sequence (CDS) and proteomic sequence. The prokaryotic database contained 3,390,814 proteins, 1,082,131,349 amino acids in total, with a mean protein length of 328 amino acids. I used the prokaryotic database as a positive control in detecting HGT.

2.3. Eukaryotic database

Since eukaryotes generally contain longer proteins and larger proteomes, I constructed the eukaryotic database consisting of 10 organisms from 10 different metazoan classes that span over six most studied phyla (**Table 1**). The eukaryotic database consisted of 381,127 proteins, 196,524,360 amino acids in total, with a mean protein length of 516 amino acids. Lower metazoan phyla, such as *Cnidaria*, were not included to avoid the inclusion of potential HTgs. These organisms were chosen for having the most complete genome assemblies in their corresponding phyla. Coding sequences (CDS) and proteomic sequences were obtained from the Ensembl database (**Dyer et al. 2025**). Since broad connections between horizontal gene transfer (HGT) and chosen metazoan phyla haven't been established, I used the eukaryotic database as a negative control in detecting HGT.

Table 1. List of organisms and their corresponding phyla and classes that that were used for the eukaryotic database. CDS and protein sequence of denoted versions were obtained from the Ensemble database (**Dyer et al. 2025**).

organism	version	phyla	classes
Schmidtea mediterranea	WBPS19	Platyhelminthes	Turbellaria
Caenorhabditis elegans	WBcel235	Nematoda	Chromadorea
Biomphalaria glabrata	BglaB1	Mollusca	Gastropoda
Daphnia pulicaria	SC_F0_13B	Arthropoda	Branchiopoda
Drosophila melanogaster	BDGP6.46	Arthropoda	Insecta
Strongylocentrotus purpuratus	Spur_5.0	Echinodermata	Echinoidea
Ciona intestinalis	KH	Chordata	Ascidiacea
Danio rerio	GRCz11	Chordata	Actinopterygii
Xenopus tropicalis	UCB_Xtro_10.0	Chordata	Amphibia
Mus musculus	GRCm39	Chordata	Mammalia

2.4. RNA libraries

In this study, I used RNA libraries isolated from *S. domuncula* and *E. subterraneus* samples. RNA isolates were obtained from *E. subterraneus* samples at two different time points: on the first and tenth days of sponge primorph growth. RNeasy Mini Kit (Qiagen) was used to isolate RNA from all samples. RNA was sequenced using Nanopore MinION and Illumina technology for *S. domuncula* and *E. subterraneus*, respectively. Nanopore sequencing was performed directly on RNA, without complementary DNA (cDNA) intermediate. The dorado program (v09; **Oxford Nanopore Technology 2025**) was used for basecalling. Non-stranded paired-end sequencing was performed on Illumina platforms. This sequencing approach reads both ends of a DNA fragment without preserving strand-specific information.

3. METHODS

I utilized a high-performance computing cluster to run all bioinformatics software (operating system: SUSE Linux Enterprise High Performance Computing, v15-SP4). I analyzed the datasets generated by these tools using R Statistical Software (v4.4.2; **R Core Team, 2024**) with the following R packages: *data.table* (v1.16.2; **Barrett et al. 2024**), *taxize* (v0.9.102; **Chamberlain & Szocs 2013**; **Chamberlain et al. 2020**), *tidyr* (v1.3.1; **Wickham et al. 2024**), *IRanges* (v2.40.1; **Lawrence et al. 2013**), *GenomicRanges* (v1.58.0; **Lawrance et al. 2013**), *Biostrings* (v2.74.1; **Pagès et al. 2024**), *ggplot2* (v3.5.1; **Wickham 2016**), *coRdon* (v1.24.0; **Elek et al. 2024**), *pheatmap* (v1.0.12; **Kolde 2019**) and *ape* (v 5.8.1; **Paradis & Schliep 2019**).

3.1. Filtering scaffolds for contamination.

The first step in this analysis was the removal of contaminating scaffolds. During sequencing, various viral, prokaryotic and other contaminations may be present in the sample. Given the high prevalence of endosymbionts in sponges, some contamination is inevitable. There are several methods for removing contamination from NGS reads or completed assemblies. I used the method described by **Bağcı et al. (2021)**. Briefly, I used Double Index Alignment of Next-generation sequencing Data (DIAMOND v2.1.8) to align sponge genomic scaffolds against a database of proteins of known origin, enabling taxonomic classification of scaffolds with MEtaGenome Analyzer (MEGAN v6.21.7; **Huson et al. 2007**). I flagged scaffolds classified as prokaryotic or viral as contamination and excluded them from further analysis.

DIAMOND is a tool for heuristic alignment of nucleotide or protein sequences, searching the query sequences for similarities with target sequences in a specific database. DIAMOND splits query and target sequences into smaller fixed-size sequences called seeds and creates a sorted list of all possible seed positions in the database (indexing). Since indexing is performed on both query and target sequences, the process is referred to as double indexing. Alignments with target sequences do not necessarily use all seed positions; hence, DIAMOND uses the so-called spaced seeds to avoid exclusively considering exact matches. The dynamic programming-based local alignment algorithm then extends sequence hits in both directions until the alignment score begins to decrease (**Buchfink et al. 2021**). For each hit, parameters such as alignment coordinates, alignment length, percentage identity, and expected value (E-value) are calculated. The E-value, a commonly used statistical parameter in sequence similarity search, represents the number of hits that appear randomly when searching a query sequence of a given length in a database of a specific size (**Altschul et al. 1990**). DIAMOND is essentially a faster version of the Basic Local Alignment Search Tool (BLAST). By employing larger erroneous spaced seeds and double indexing, DIAMOND achieves sensitivity similar to BLAST but with significantly increased speed (**Buchfink et al. 2021**). Furthermore, DIAMOND can combine protein and nucleotide sequences, with nucleotide sequences translated into all three reading frames and both orientations (6 possibilities in total).

In the context of contamination filtering, I used DIAMOND to align sponge genomic scaffolds against the non-redundant protein database (nr-NCBI) (**Sayers et al. 2025**; **Goldfarb et al. 2025**) using parameters described by **Bağcı et al. (2021)**. I set a frame-shift penalty to 15 to strike a good balance between producing long alignments without excessive switching of frames. Only alignments whose bit-score lies within 10% of the best score of competing alignments were reported. Range culling was set as the reporting mode to avoid reporting only alignments that cover a small, highly conserved region of the query, which often happens when DIAMOND is applied to longer sequences. I supplied the results of DIAMOND alignments to MEGAN for taxonomic analysis.

MEGAN is a program for taxonomic classification that uses the lowest common ancestor (LCA) algorithm. The taxonomic category of a scaffold is defined as the smallest common taxonomic category among hits with the lowest E-value found on that scaffold (**Huson et al. 2007**). DIAMOND results must be processed ("meganized") as described by **Bağcı et al. (2021)** to be subsequently analyzed by MEGAN. I set the LCA algorithm to long reads. To ensure high statistical significance and filter out weak matches, I set a minimum score of 75 and a maximum expected E-value of 10^{-10}. Additionally, I set a minimum percentage identity of 70% to ensure that only sequences with reasonable similarity to the reference are considered. To refine the assignment process, I included alignments within 5% of the best score in the analysis, but only for taxonomic nodes supported by at least one read. I put no restriction on sequence length, to allow all scaffolds to be analyzed. To reduce the number of false-positive classifications, 75% of the scaffold must align to the reference sequence, and the coverage of the reference sequence had to be at least 70%. I filtered prokaryotic and viral scaffolds from sponge genome assemblies using the *Biostrings* R package. I compared scaffold length difference between genome without contaminations and contaminants with Mann-Whitney U test. Genome assemblies filtered from contamination were used in further analysis.

3.2. BRAKER3 annotation

Coordinates of genes, coding sequences (CDS) and protein sequences pf the analyzed sponge genomes were predicted by the BRAKER3 (v3.0.0; **Gabriel et al. 2024**) annotation program for protein-coding genes in eukaryotic genomes. BRAKER3 is a program built upon *ab initio* annotation programs GeneMark-ETP and AUGUSTUS, which use statistical models and genome sequence alone to predict gene regions. Notably, BRAKER3 also relies on experimental data. Previous versions of the algorithm were limited to use either RNA-seq short reads (BRAKER1) or homologous proteins from large protein databases (BRAKER2). With the development of TSEBRA combiner, these data could be combined and incorporated in the *ab initio* annotation pipelines that iteratively train statistical models on a target genome. In this way, ab-initio-only annotation and annotation based exclusively on experimental data complement and improve accuracy within BRAKER3 pipelines.

I used BRAKER3 on the contamination-free sponge genomes with the supplied protein evidence in form of the OrthodB11 metazoan protein database (v11; **Kuznetsov et al. 2023**), supplemented with predicted protein sequences of the reference sponge species *Amphimedon queenslandica*. In case of *Eunapius subterraneus*, we also provided two short-read RNAseq libraries sequenced from *E. subterraneus* primorph cultures as RNA-seq-based evidence. Genome annotations, CDS and amino acid sequences of proteins predicted by BRAKER3 were used in the later steps of the analysis.

3.3. Horizontal gene transfer detection

3.3.1. Protein-based horizontal gene transfer detection

The Alien Index (AI) is a primary method for detecting HGT using protein sequences. This method is based on calculating the logarithmic value of the ratio between the E-value of the best eukaryotic alignment hit and the best prokaryotic alignment hit for each query protein. Proteins with AI values above a certain threshold are classified as HGT candidates (**Conaco et al. 2016**). To determine HGT candidates for sponge proteins, three DIAMOND alignments were performed against the nr-NCBI database: (1) prokaryotic database proteins, (2) eukaryotic database proteins, (3) sponge proteins. All alignments with bit-scores within 80% of the best score of competing alignments were reported. Due to the size of the prokaryotic database, I used 10% of randomly selected protein sequences.

I determined the taxonomic domain for every hit using the *taxize* R package and extracted only prokaryotic and eukaryotic hits. I retained only the best eukaryotic and prokaryotic hit (lowest E-value) for each protein. In case of identical E-values, the best hit was chosen by a longer alignment length, E-values of 0 were set to the minimal non-zero value found in the dataset. I used the DIAMOND results from prokaryotic and eukaryotic database proteins as positive and negative control, respectively. Accordingly, I determined the threshold AI value by using grid search and optimizing the F-beta score for HGT classification. Proteins having only prokaryotic or only eukaryotic hits were given maximal and minimal AI values in the dataset, respectively. The F-beta score is the weighted harmonic mean of precision (proportion of true positive results among all positive predictions) and recall (proportion of true positive results among all actual positives). Optimizing the F-beta score ensures that the model balances well between precision and recall (**Evidently AI Team 2024**). I set the β-factor to 0.5 to give precision a larger weight. I classified sponge proteins with AI above the selected threshold and alignment length longer than 150 aa as HGT candidates.

3.3.2. Nucleotide-based horizontal gene transfer detection

Horizontal gene transfer can also be detected using nucleotide-based pipelines. To do so, two BLAST-Like Alignment Tool (BLAT; v36; **Kent 2002**) similarity searches were performed against the contamination-free sponge genomic sequence: (1) CDS sequences from the prokaryotic database, and (2) CDS sequences from the eukaryotic database. BLAT is a sequence alignment tool similar to BLAST. These tools differ in efficiency and are generally used for different purposes. BLAST is highly sensitive, uses a slower, more thorough algorithm suitable for finding distant homologs. On the other hand, BLAT is optimized for speed and works best with closely related sequences for large data. In conclusion, BLAT is more suitable than BLAST for high-throughput alignments with high identity, such as CDS versus genome alignments (**Bhagwat et al. 2012**).

For this analysis, I used the default BLAT settings. After performing BLAT alignments, I filtered the resulting alignments for the maximal E-value of 10^{-30}, and merged alignments regions of the same CDS whose distance was lower than 10 kb to encompass plausible breaks caused by introns. Afterwards, I merged all overlapping alignment regions per scaffold. I processed alignments from the prokaryotic and the eukaryotic database separately, overlapped them, and then excluded prokaryotic alignment regions that overlapped with eukaryotic alignments. Finally, I merged all prokaryotic regions whose distance was lower than 1 kb and classified them as HGT candidate regions.

To ensure BRAKER3 did not miss prokaryotic-like genes that could be acquired by horizontal transfer, I annotated HGT candidate regions using AUGUSTUS (v3.4.0), an *ab initio* annotation program for eukaryotic organisms that can be trained on a prokaryotic protein dataset. I added 100 bp to each end of HGT candidate regions to ensure the inclusion of the whole protein sequence. The combination of AUGUSTUS and a prokaryotic dataset could help recognize horizontally acquired prokaryotic genes that adopted eukaryotic genomic features. I used the CDS from the prokaryotic database to train AUGUSTUS. The minimal intron length was set to zero. I set the species parameter to ‘amphimedon’, as *A. queenslandica*, the reference sponge genome.

To get a set of genes predicted by BRAKER3 that can be found in HGT candidate regions, I overlapped the whole-genome BRAKER3 annotations with HGT candidate regions. Then, I analyzed both BRAKER3 and AUGUSTUS subregions within the HGT candidate regions. I overlapped the subregions predicted by both programs and divided the subregions into two categories: AUGUSTUS-derived and BAKER3-derived subregions. The annotation for a subregion was derived from AUGUSTUS if the subregion was AUGUSTUS-specific (no overlaps with BRAKER3) or if BRAKER3 subregions cumulatively overlapped with less than 30% of the AUGUSTUS subregion. The annotation for a subregion was derived from BRAKER3 if the subregion was BRAKER3-specific (no overlaps with AUGUSTUS) or if BRAKER3 subregions cumulatively overlapped with 30% or more of an AUGUSTUS region.

3.3.3. Filtering horizontal gene transfer candidates based on genomic location

From the previous analysis, I obtained the IDs of gene candidates determined through protein-based (BRAKER3 annotations) and nucleotide-based (BRAKER3 and AUGUSTUS annotations) HGT detection pipelines. I counted the number of genes annotated by AUGUSTUS and BRAKER3 and merge them into a gene set representing HGT candidates. I counted the number of HGT candidates on each genomic scaffold. HGT candidates on scaffolds smaller than 50 kb or scaffolds where more than 50% of genes were classified as HGT candidates were excluded from further analysis. These scaffolds could represent potential genomic contamination undetected with the DIAMOND+MEGAN pipeline. BRAKER3 and AUGUSTUS annotated HGT candidates on remaining scaffolds were regarded as HTgs.

3.4. General features of detected horizontally transferred genes

After protein-based and nucleotide-based detection of HGT and scaffold filtering, I examined the general features of detected HTgs and compared them with non-HTgs. Gene regions for both HTgs and non-HTgs were determined by joining all exon regions and intronic regions of the longest transcripts annotated by BRAKER3 or AUGUSTUS per gene. Firstly, I compared the number of exons per gene for HTgs and non-HTgs using the Mann-Whitney U test. Secondly, I compared the distribution of exon, intron and gene length of HTgs and non-HTgs with the Mann-Whitney U test, Furthermore, I analyzed the GC content of HTgs and non-HTgs for both sponge species. I compared the distribution of GC content of HTgs and non-HTgs to the mean GC content of the whole genome with the Welch’s t-test.

Lastly, I analyzed the CDS of each HTgs and non-HTgs to determine codon usage bias. Amino acid coding system is degenerated, as several codons code for the same amino acid. I counted the amount of each codon and the number of codons that code a specific amino acid. I expressed the codon usage as the proportion of a codon count within all codons that code for the same amino acid. I divided the codon usage of HTgs with the codon usage calculated for non-HTgs and calculated a base-10 logarithmic value of the ratio. Thus, positive values denote higher codon usage in HTgs, while negative values correspond to lower codon usage in HTgs compared to non-HTgs. Additionally, I plotted codon usage on a biplot using *cordon* package to examine whether HTgs and non-HTgs show different clustering. Due to a substantially higher number of non-HTgs, non-HTG data points were randomly sampled to match the number of HTgs.

3.5. Expression analysis

To analyze the expression level of HTgs, I calculated the relative expression from one *S. domuncula* and two *E. subterraneus* RNA libraries. For *S. domuncula*, adapter removal was performed with longQC (v1.2.1; **Fukasawa et al. 2020**). Nanopore reads were mapped on the *S. domuncula* genome with minimap2 program (v2.24; **Li 2018**) using 14 bp k-mers, splice-aware mode without masking repetitive genomic regions. For *E. subterraneus*, adapter removal and quality filtering were processed using the Bbduk program, part of the BBTools package (v36.20; **Bushnell 2014**). Bbduk program removed bases at the sequence ends with a quality Phred score below 8 and excluded sequences with an average quality Phred Score below 16. Illumina reads were mapped to the *E. subterraneus* genome using the BBmap program, part of the BBTools package with the parameters "ambiguous=random secondary=f maxindel=100000." FeatureCounts program (v2.0.1; **Liao et al. 2014**) with default parameters was used to count mapped reads for both sponges. For *E. subterraneus,* an option customized for counting paired-end reads was used. Reads mapped to multiple locations were ignored, and only one of the paired reads was counted. To compare relative expression, Reads Per Kilobase Million (RPKM) and Fragments Per Kilobase Million (FPKM) values were calculated from the number of reads mapped to *S. domuncula* and *E. subterraneus* genomes, respectively. Both RPKM and FPKM values enable reliable comparison of gene expression levels within and between samples and are defined as the number of sequenced fragments mapped per kilobase of gene length for every million reads mapped. In other words, they represent normalized units for gene length (in kilobases) and the number of mapped sequences (in millions) in the library for single-end and paired-end reads, respectively. I used the Mann-Whitney U test to compare the relative expression of HTgs and non-HTgs in each of the analyzed libraries. I also compared the relative expression of genes stratified by HGT status in day 1 and day 10 *E. subterraneus* libraries using the Wilcoxon signed-rank test.

3.6. Functional annotation

To functionally annotate HTgs I performed a DIAMOND alignment of all sponge proteins against the UniProtKB/Swiss-Prot database (**The UniProt Consortium 2025**) with default parameters. DIAMOND results were mapped to GO terms using the Blast2GO tool (v6.0, **Götz et al. 2008**). GO terms represent a controlled vocabulary of gene and gene product attributes that are divided into 3 domains: cellular component, molecular function, and biological process. Moreover, the GO vocabulary is designed to be species-neutral and inclusive towards prokaryotic and eukaryotic clades.

I used the Fisher's exact test to analyze the potential enrichment of GO terms in HTgs, I quantified the association between GO terms and HGT candidates using the odds ratio based on a 2x2 contingency table analysis. I manually examined 20 molecular function and biological process GO terns exhibiting the strongest enrichment within HTgs. I analyzed the genomic location of HTgs enriched with the selected GO terms.

3.7. Detecting homologues of horizontally transferred genes

Finally, I tried to detect homologous HTgs in related sponge species, I used BLAT with default parameters to perform a similarity search between CDS sequences of HTgs and a collection of publicly available and in-house sponge genomes. For both *S. domuncula* and *E. subterraneus* I built a database of seven publicly available sponge genomes and added the genome of either *E. subterraneus* or *S. domuncula*, respectively. The publicly available genomes included the following species: *Amphimedon queenslandica*, *Ephydatia muelleri*, *Geodia barretti*, *Halisarca caerula*, *Oopsacas minuta*, *Tethya californiana* and *Xestospongia bergquistia*. Genomes were downloaded from the NCBI database (**Sayers et al. 2025**; **Goldfarb et al. 2025**). I filtered the alignments with E-values larger than 10^{-10}. I calculated the length percentage of each HTgs that aligns to the specific genome and visualized the results with a hierarchical clustering heatmap for both HTgs and sponge species. Clustering was performed using complete clustering and Euclidean method for calculating clustering distance. To contextualize data interpretation, I downloaded a simple phylogenetic tree of utilized sponge species from the NCBI Taxonomy Browser (**Schoch et al. 2020; Sayers et al. 2019**). I compared the hierarchical clustering results with the cladogram derived from NCBI. Afterwards I determined sponges most related to *S. domuncula* and *E. subterraneus* using the obtained heatmaps and cladograms. I specifically analyzed HTgs that aligned with more than one sponge genome, with at least one alignment covering more than 50% of the HTgs, excluding the most related sponge. I represented the length percentage of HTgs that align with available sponge genomes for *S. domuncula* and *E. subterraneus* using a heatmap.

4. RESULTS

4.1. Filtering scaffolds for contamination

I filtered the analyzed sponge genomes for contamination, which included the removal of prokaryotic and viral scaffolds, as classified by MEGAN. Descriptive statistics of whole genomes, genomes without contamination and contaminant scaffolds are shown in **Table 2.** The distributions of scaffold length for whole genomes, genomes without contamination and contaminant scaffolds are shown in **Figure 2.** In genomes of both species, prokaryotic contaminants were significantly more abundant than viral contaminants.

Table 2. Descriptive statistics of whole genomes, genomes without contamination and contaminant scaffolds after contamination filtering. no. – number of scaffolds, length – total genome length, N50, L50, N90, L90 – measures of genome assembly quality.

	scaffolds	length	no.	N50	L50	N90	L90
Suberites domuncula	all	123,813,400	900	509,668	60	68,394	330
	uncontaminated	119,914,452	856	530,070	56	71,635	303
	contamination	3,898,948	44	123,850	8	44,466	29
	prokaryotes	3,818,929	42	123,850	8	49,456	27
	virus	80,019	2	44,466	1	35,553	2

	scaffolds	length	no.	N50	L50	N90	L90
Eunapius Subterraneus	all	202,844,449	3,694	161,204	368	39,134	1,355
	uncontaminated	190,478,823	3,549	165,058	339	38,838	1,265
	contamination	12,365,626	145	118,777	31	48,956	91
	prokaryotes	12,000,178	140	118,980	30	48,956	87
	virus	365,448	5	100,802	2	39,080	5

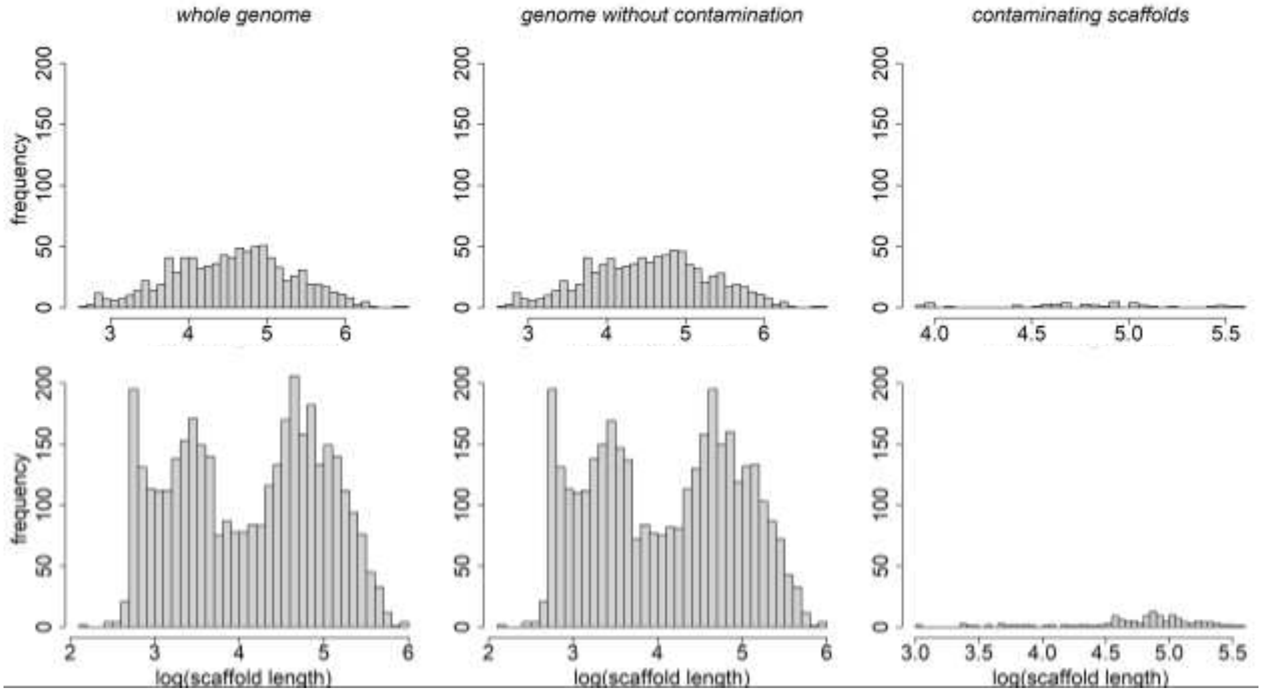


Figure 2. Scaffold log-length distributions of whole genome, genome without contamination, and contaminating scaffolds for *S. domuncula* (up, N = 900; 856; 44, respectively) and *E. subterraneus* (down, N = 3,694; 3,549; 145, respectively).

The log-lengths of *S. domuncula* scaffolds were fairly normally distributed. The genome of *S. domuncula* was well assembled with 900 scaffolds comprising a genome length of 120 Mb. However, after filtering for contamination, 5% of the assembled genome in scaffold number was classified as contaminants, corresponding to 3% of the assembled genome length. Contaminant scaffolds (y) were significantly longer than non-contaminant scaffolds (x) (Mann-Whitney U test, $\tilde{x}$ = 37,323.5, $\tilde{y}$ = 60,535, p = 0.043).

The distributions of whole genome and non-contaminant scaffold log-lengths of *E. subterraneus* is bimodal. The genome of *E. subterraneus* contains 3,694 scaffolds comprising a genome length of 200 Mb. However, after filtering for contamination, 4% of the assembled genome in scaffold number was classified as contaminants, corresponding to 6% of the assembled genome length. Contaminant scaffolds (y) were significantly longer than non-contaminant scaffolds (x) (Mann-Whitney U test, $\tilde{x}$ = 13,915, $\tilde{y}$ = 71,188, p < 0.001).

4.2. BRAKER3 annotation

I annotated the contamination-free scaffolds with the BRAKER3 tool. This resulted in 37,100 annotated protein-coding genes coding for 15,882,566 amino acids in total, with a mean protein length of 428 amino acids in *S. domuncula*, and 47,188 protein-coding genes, coding for 18,408,909 amino acids in total, with a mean protein length of 390 amino acids in *E. subterraneus*.

4.3. Horizontal gene transfer detection

4.3.1. Protein-based horizontal gene transfer detection

I calculated AI for overall sponge proteins annotated with BRAKER3. To calculate the AI threshold for optimal F-beta value, proteins from prokaryotic and eukaryotic databases were used as positive and negative HGT controls, respectively. **Figure 3.** depicts the distribution of prokaryotic and eukaryotic AI values calculated in the prokaryotic and eukaryotic datasets, F-beta values for different AI thresholds and the distribution of protein AI values for both sponge species according to HGT status.

Distribution of AI values showed a clear polarity between the eukaryotic and prokaryotic database. Prokaryotic proteins predominantly exhibited positive AI values, while eukaryotic proteins showed negative AI values, with a small overlap around the value of zero. The highest F-0.5 value of 0.994 corresponded to the AI threshold of 1.9. The graph showing the F-0.5 values for different AI exhibits a positive slope that peaked and shifted to a negative slope around an AI value of zero.

AI values for both sponges showed a roughly normal distribution with high peaks for AI values around zero. The distribution of *S. domuncula* AI values was fairly symmetrical, while distribution of *E. subterraneus* AI values was notably left-skewed. HGT candidates exhibited a narrower AI value range among *E. subterraneus* proteins compared to *S. domuncula* proteins. In total, the protein-based HGT detection method classified 9,174 (25%) *S. domuncula* proteins as candidates for HGT. On the other hand, 3,483 (7%) proteins were classified as candidates for HGT in *E. subterraneus*.

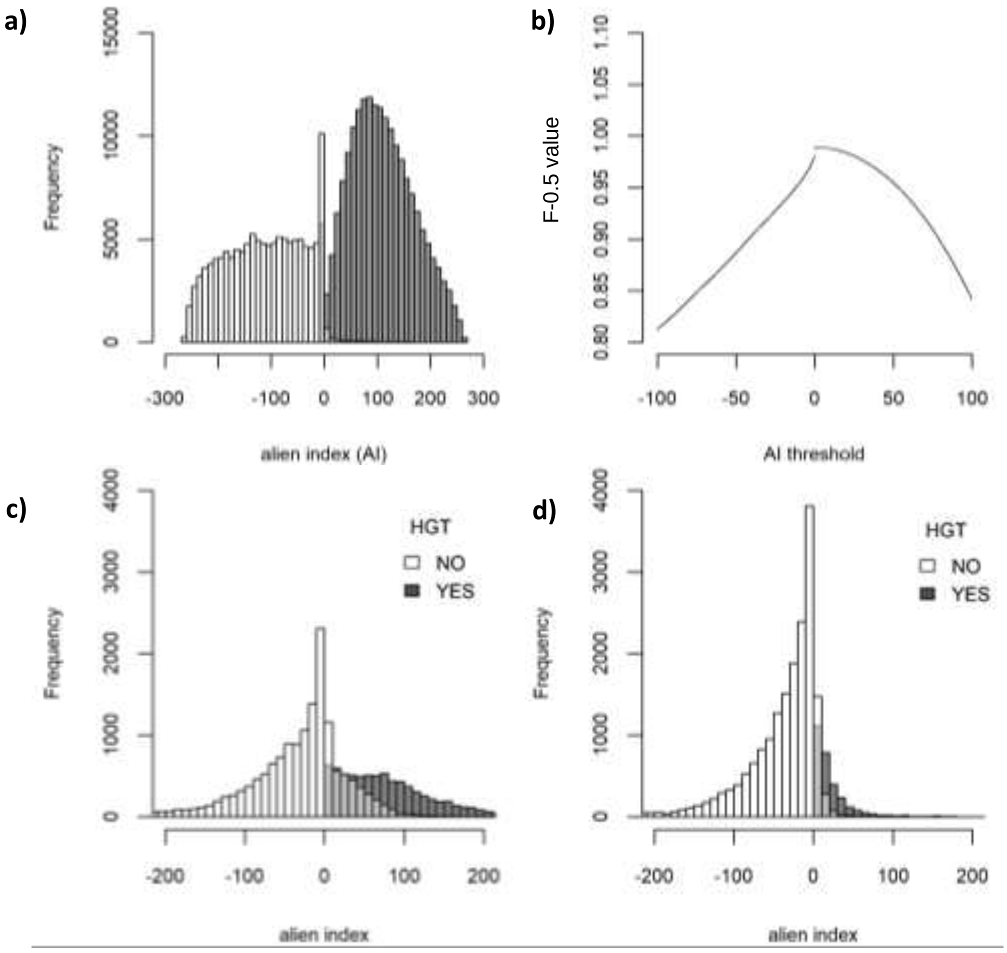


Figure 3. Protein-based horizontal gene transfer detection. a) Distribution of prokaryotic (dark) and eukaryotic (light) alien index (AI) values; **b)** F-0.5 values for the tested AI thresholds; **c,d)** distribution of alien index values by HGT status in *S. domuncula* and *E. subterraneus*, respectively. All depicted distributions show proteins with both prokaryotic and eukaryotic DIAMOND hits.

4.3.2. Nucleotide-based horizontal gene transfer detection

I overlapped BLAT alignment hits of sponge contamination-free scaffolds against prokaryotic and eukaryotic databases to characterize HGT regions. Additionally, I overlapped AUGUSTUS and BRAKER3 annotations of HGT regions and defined consensus candidates for HGT based on AUGUSTUS-derived annotation and BRAKER3-derived annotation (**Table 3**).

Species *E. subterraneus* exhibited a notably larger number of hits with the eukaryotic database, compared to *S. domuncula*. However, after overlapping hits with eukaryotic and prokaryotic databases, *the* number of defined HGT regions was higher in *S. domuncula*. Similarly, the number of BRAKER3 and AUGUSTUS HGT candidates was proportional to the number of HGT regions and was significantly higher in *S. domuncula.*

Table 3. Protein-based horizontal gene transfer detection. pro-hits, eu-hits – number of prokaryotic and eukaryotic BLAT alignment hits against sponge contamination-free scaffolds, respectively; HGT regions – number of defined HGT regions; AUGUSTUS, BRAKER3 – consensus candidates for HGT with AUGUSTUS-derived annotation and BRAKER3-derived annotation, respectively.

	pro-hits	eu-hits	HGT regions	BRAKER3	AUGUSTUS
S. domuncula	314,240	4,638,149	2,995	5,924	150
E. subterraneus	108,094	26,318,728	373	537	17

4.3.3. Filtering horizontally transferred genes based on genomic position

After defining HGT candidates with both protein-based and nucleotide-based methods, I performed the last filtering which eliminated scaffolds predominantly containing HTgs (>50%) as well as relatively short scaffolds (<50 kb). The results of filtering are shown on **Figure 4.** Scaffolds without HGT candidates are not depicted.

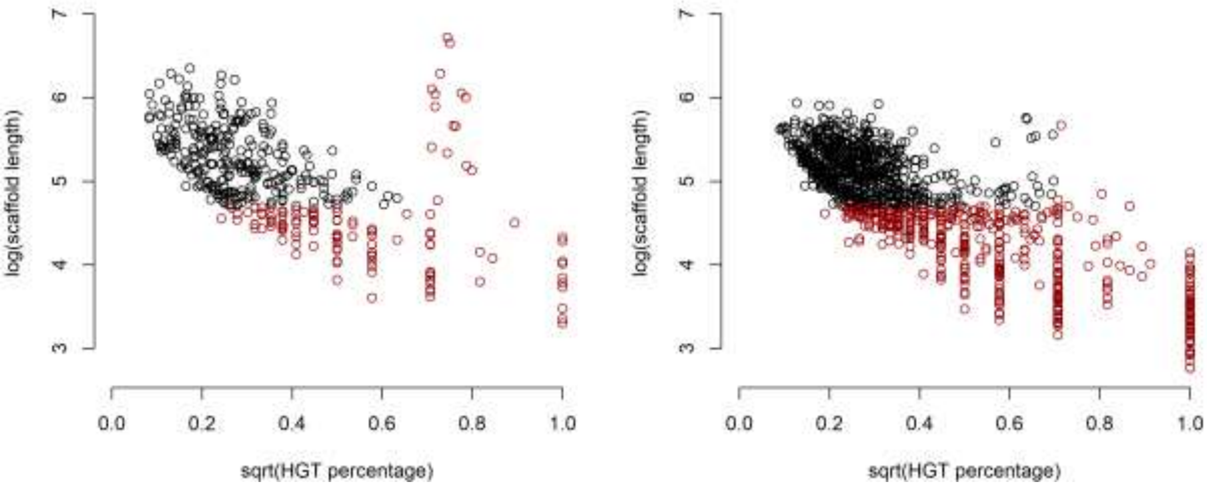


Figure 4. **Log-length of sponge scaffolds over logarithmic values of the proportion of regions of HTgs in *S. domuncula* (left), and *E. subterraneus* (right).** Scaffolds with suspiciously many (>50%) or no candidates for HGT and scaffolds that were too short (<50 kb) are colored red and were expelled from further analysis. Scaffolds with no HTgs are not shown.

As previously shown, the genome of species *E. subterraneus* is divided into more scaffolds compared to *S. domuncula*. However, after filtering, both sponge species retained 30% of their contamination-free scaffolds. Thus, on 2,174 remaining scaffolds 2,550 genes were classified as HTgs in *E. subterraneus*, and on 552 remaining scaffolds 1,118 genes were classified as HTgs in *S. domuncula*.

4.4. General features of detected horizontally transferred genes

Following the detection of HTgs, I analyzed general features of the detected HTgs. This included the distribution of number of exons per gene, the distribution of gene, intron and exon lengths, the distribution of GC content, and codon usage bias.

The distribution of per-gene exon and intron numbers in HTgs and non-HTgs is shown in **Figure 5.** For both *S. domuncula* and *E. subterraneus*, HTgs more commonly contained less than six and seven exons per gene, respectively. Higher numbers of exons per gene are more frequent in non-HTgs. The highest number of exons per gene in *S. domuncula* were 82 for HTgs and 120 for non-HTgs. Furthermore, the highest number of exons per gene in *E. subterraneus* were 33 for HTgs and 86 for non-HTgs. The per-gene number of exons was significantly lower in HTgs (x) when compared to non-HTgs (y) in *S. domuncula* (Mann-Whitney U test, $\bar{x}$ = 3.7, $\bar{y}$ = 4.6, p < 0.001). and *E. subterraneus* (Mann-Whitney U test, $\bar{x}$ = 3.5, $\bar{y}$ = 4.7, p < 0.001).

I compared gene, intron and exon lengths for HTgs and non-HTgs in both sponge species. The distributions of gene, intron and exon lengths are shown in **Figure 6.** HTgs (x) were significantly longer when compared to non-HTgs (y) in both *S. domuncula* (Mann-Whitney U test, $\tilde{x}$ = 1,701 bp, $\tilde{y}$ = 1,184 bp, p < 0.001) and *E. subterraneus* (Mann-Whitney U test, $\tilde{x}$ = 1,463 bp, $\tilde{y}$ = 1,347 bp, p < 0.001). Introns in HTgs (x) were significantly longer when compared to non-HTgs (y) in *S. domuncula* (Mann-Whitney U test, $\tilde{x}$ = 156 bp, $\tilde{y}$ = 149 bp, p = 0.001), but shorter in *E. subterraneus* (Mann-Whitney U test, $\tilde{x}$ = 87 bp, $\tilde{y}$ = 135 bp, p < 0.001). Exons in HTgs (x) were significantly longer when compared to non-HTgs (y) in both *S. domuncula* (Mann-Whitney U test, $\tilde{x}$ = 217 bp, $\tilde{y}$ = 154 bp, p < 0.001) and *E. subterraneus* (Mann-Whitney U test, $\tilde{x}$ = 263 bp, $\tilde{y}$ = 150 bp, p < 0.001).

I compared the distribution of GC content of genes classified as HTgs and non-HTgs. The distributions for both gene classes in both sponge species are depicted in **Figure 7.** In *S. domuncula*, the average GC content of all predicted sponge genes was 0.52. The average GC content for non-HTgs did not differ significantly (Welch's t-test, $\bar{x}$ = 0.52, t = 1.41, p = 0.156), while it was significantly lower for HTgs (Welch's t-test, $\bar{x}$ = 0.51, t = -9.2, p < 0.001). Moreover, the GC content distribution of non-HTgs was symmetrical, while several local peaks at 0.46 and 0.53 can be distinguished in the GC content distribution of HTgs. In *E. subterraneus*, the average GC content of all sponge genes was 0.53. The average GC content for non-HTgs does not differ significantly (Welch's t-test, $\bar{x}$ = 0.53, t = 1.40, p = 0.160), while it was significantly lower for HTgs (Welch's t-test, $\bar{x}$ = 0.52, t = -7.73, p < 0.001). Moreover, both distributions were fairly symmetrical. However, both HTgs and non-HTgs had a broad secondary peak at 0.47, which was more pronounced in HTgs.

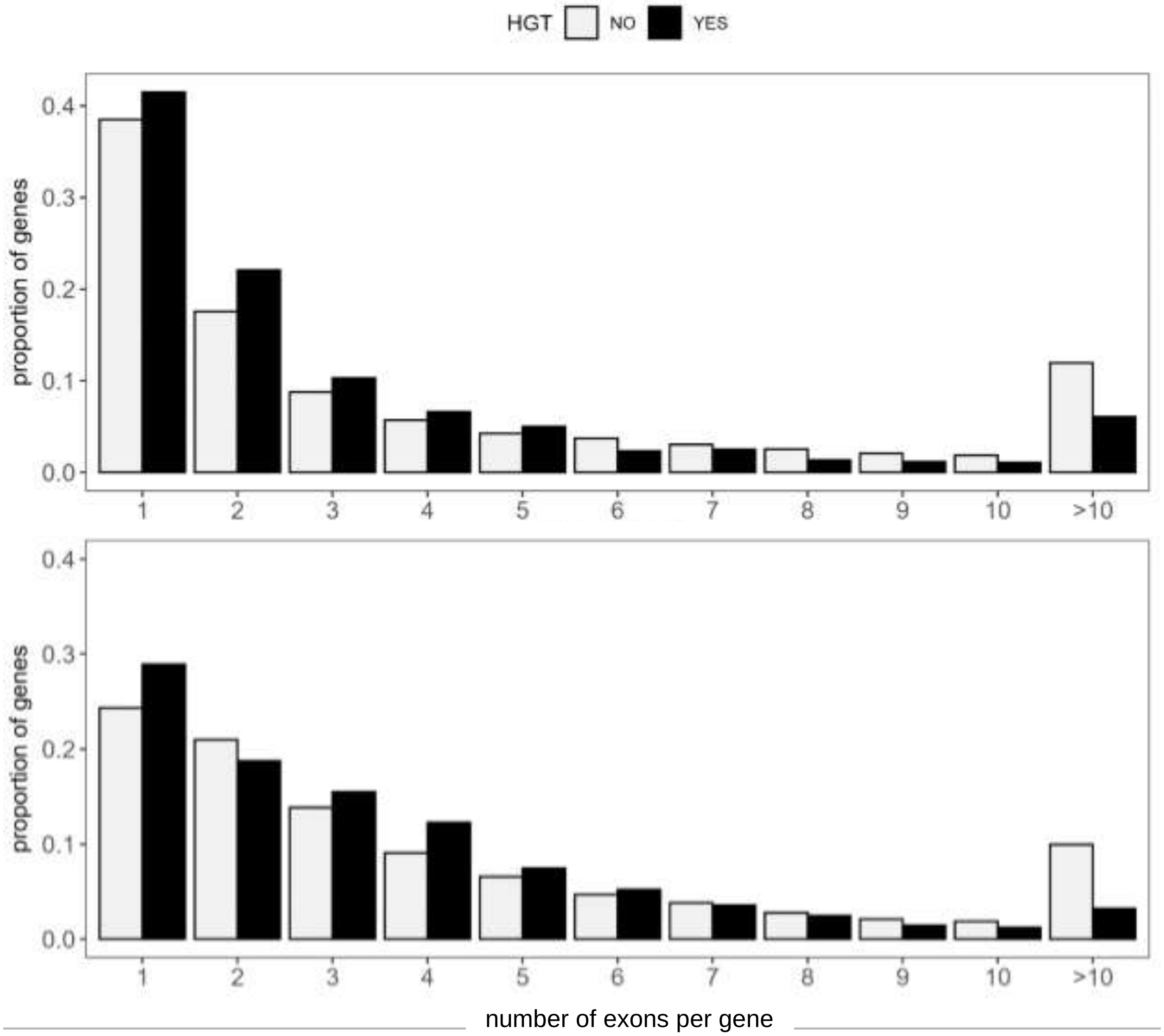


Figure 5. The distribution of the proportion of HTgs and non-HTgs according to the number of exons per gene for *S. domuncula* (up) and *E. subterraneus* (down).

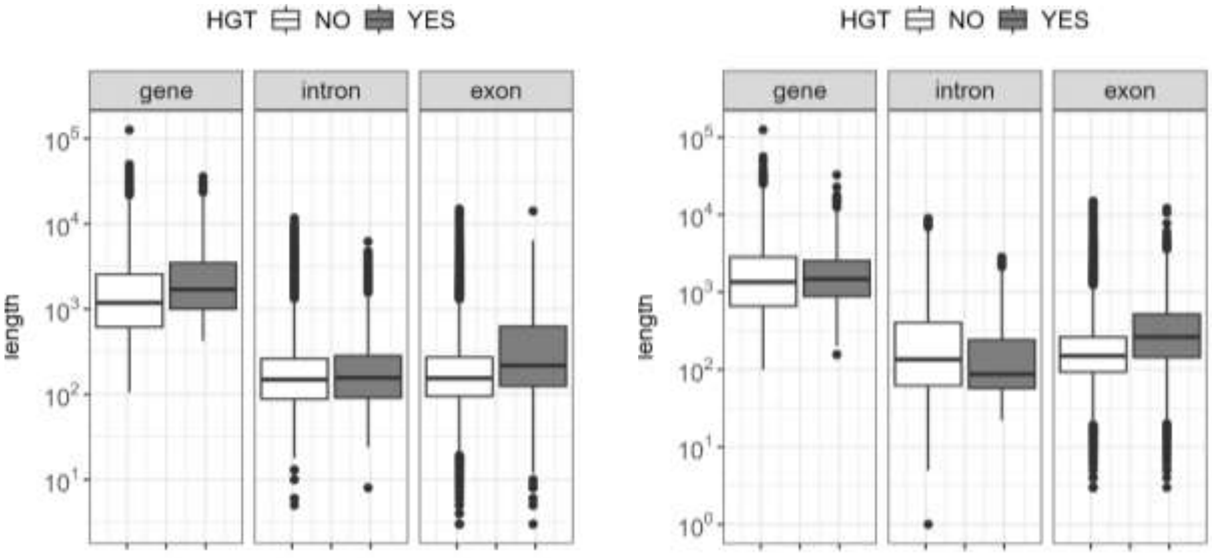


Figure 6. The distributions of gene, intron and exon lengths for *S. domuncula* (left) and *E. subterraneus* (right). The horizontal black line represents the median of the distribution, while the vertical black lines extend from –1.5IQR to 1.5IQR, (IQR – interquartile range). The extreme values are shown by black dots. The lengths are logarithmically scaled.

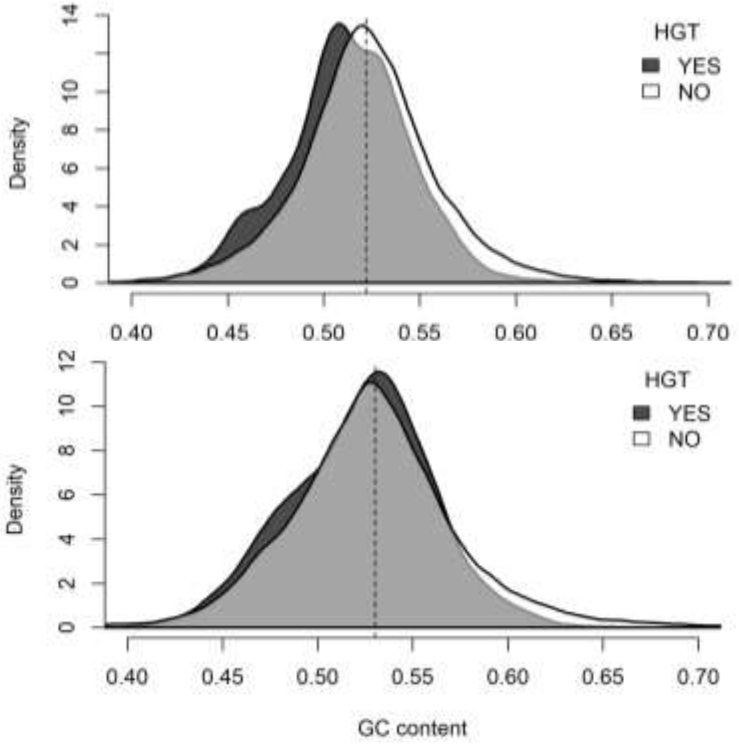


Figure 7. The distribution of GC content for HTgs and non-HTgs in *S. domuncula* (up), and *E. subterraneus* (down). Dashed line shows the average GC content of all genes in the denoted sponge species.

Next, I counted the number of codons corresponding to certain amino acids and compared the results for HTgs and non-HTgs. Synonymous codons code for the same amino acid so codon usage counts were expressed as relative codon usage for each group of synonymous codons. **Figure 8.** shows a visible difference in relative codon usage between HTgs and non-HTgs in both sponge species.

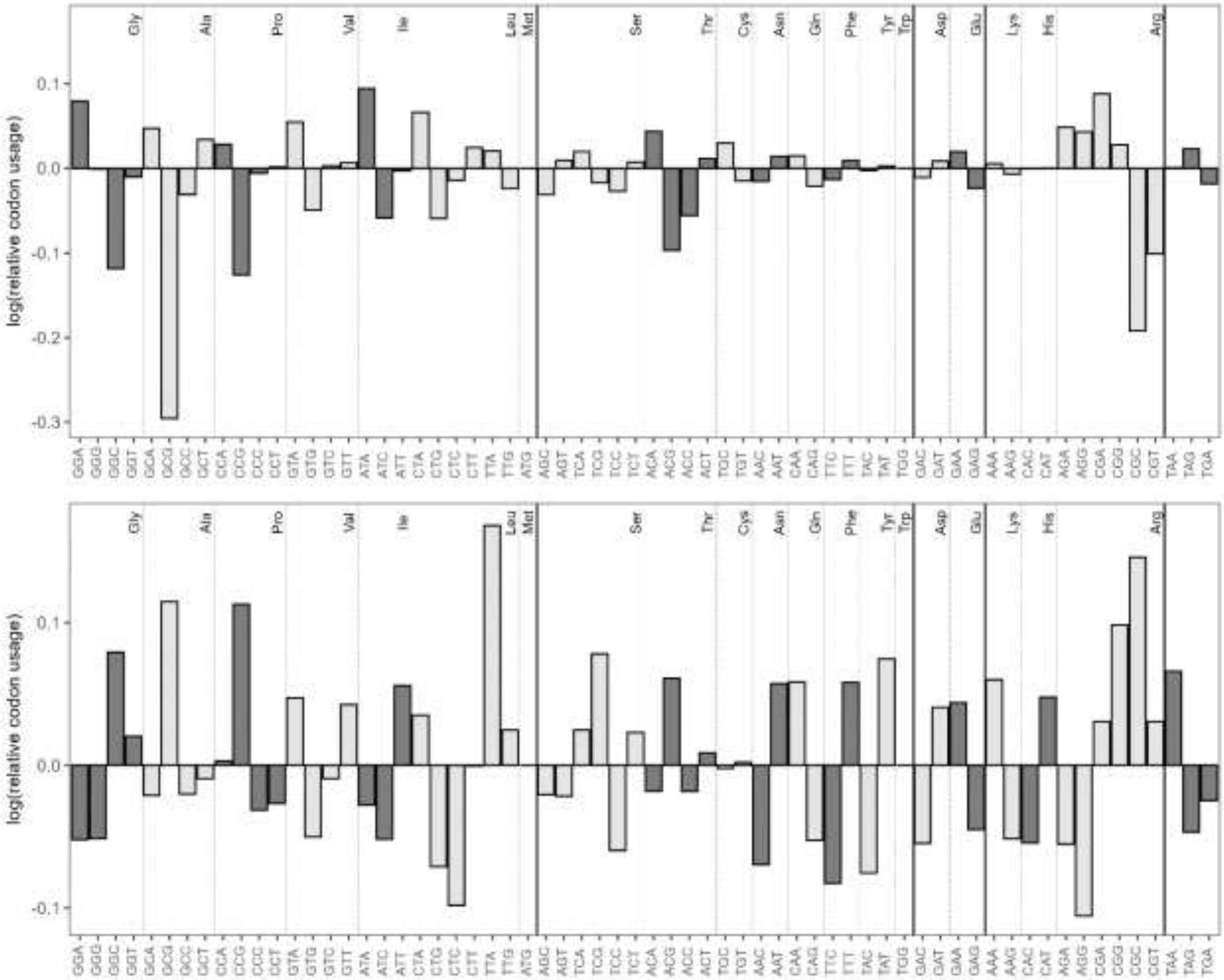


Figure 8. Codon usage comparison between HTgs and non-HTgs in *S. domuncula* (up) and *E. subterraneus* (down). Relative codon usage is expressed as a ratio of the proportion of a codon count within all codons that code for the specific amino acid in HTgs and non-HTgs. Values are logarithmically scaled. Positive values denote higher codon usage in HTgs, while negative lower codon usage in HTgs compared to non-HTgs. Interchange of two shades of grey (light and dark) and dashed lines group codons corresponding to the same amino acid. Amino acids of each group are represented by their three-lettered abbreviation. Continuous vertical lines divide amino acids into four groups: non-polar, polar, acidic, and basic. The last group represents STOP codons.

The difference in relative codon usage was more uniform in *S. domuncula* although several codons show a higher difference in *E. subterraneus.* Trends were observed in nucleotide composition dependent of the nucleotide position in the codon. In *S. domuncula*, HTgs showed a preference towards codons with ending AT base pairs, especially towards adenine on the third position. In *E. subterraneus*, codons with all-AT or all-CG base pairs showed the highest relative codon usage in HTgs. Interestingly, the CG dinucleotide was present in all positions in codons with high relative codon usage in HTgs found in *E. subterraneus.*

Codon usage was additionally compared with a biplot. The plot visualizes results of Discriminant Analysis (DA), specifically Between-Group Analysis (BGA), used to separate HTgs and non-HTgs based on codon usage patterns. The biplot shows strong clustering of HTgs in *E. subterraneus*, while no substantial clustering of HTgs is observed in *S. domuncula* (**Figure 9.**).

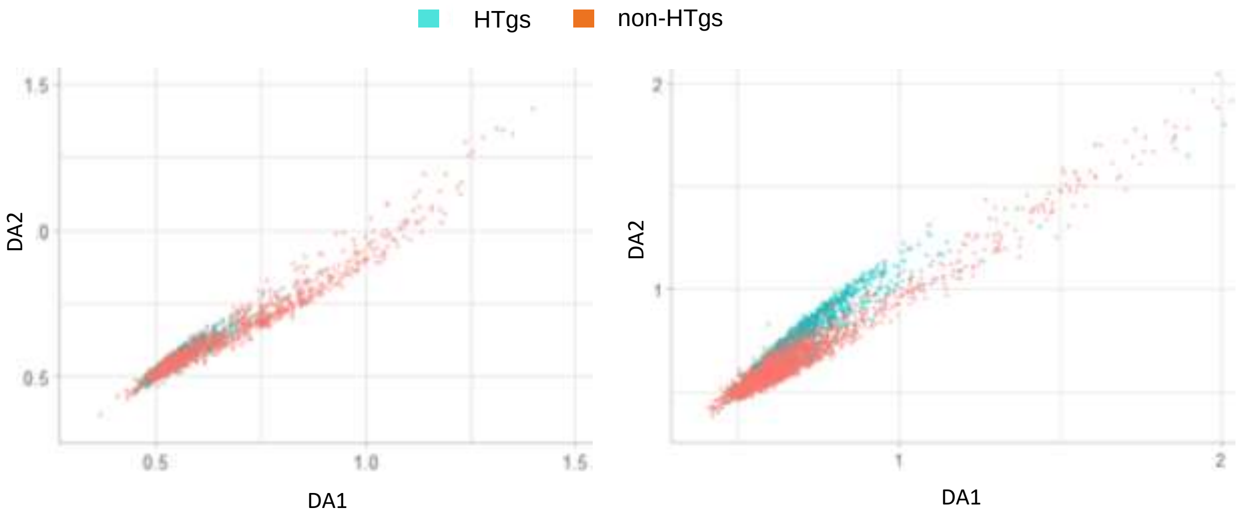


Figure 9. Biplot of the codon usage bias for *S. domuncula* (left) and *E. subterraneus* (right)*.* Due to a higher number of non-HTgs in several levels of magnitude, data for non-HTgs was randomly sampled to the amount of data for HTgs for the graph. DA1, DA2 – discriminant analysis axis.

4.5. Expression analysis

After normalizing the number of RNAseq reads per gene length and library size, I compared the relative expression rate for HTgs and non-HTgs in all analyzed RNAseq libraries. For *S. domuncula* only one RNAseq library was available, while for *E. subterraneus* RNAseq libraries for two temporal points of sponge primorph growth are analyzed. The results are presented in **Figure 10.** HTgs (x) exhibited lower relative expression when compared to non-HTgs (y) in *S. domuncula* (Mann-Whitney U test, $\bar{x}$ = 0.5, $\bar{y}$ = 3.1, p < 0.001). HTgs (x) had lower relative expression when compared to non-HTgs (y) for primorph growth in day 1 in *E. subterraneus* (Mann-Whitney U test, $\tilde{x}$ = 0.0, $\tilde{y}$ = 0.2, $p < 0.001$). HTgs (x) had lower relative expression when compared to non-HTgs (y) for primorph growth in day 10 in *E. subterraneus* (Mann-Whitney U test, $\tilde{x}$ = 0.0, $\tilde{y}$ = 0.7, $p < 0.001$). Additionally, 68%, 36%, and 37% of HTgs exhibited no noticeable expression in *S. domuncula* and *E. subterraneus* in samples of day 1 and day 10, respectively. Within HTgs with any detected expression, only 10 (3%) HTgs had relative expression above 10 in *S. domuncula.* In HTgs with any detected expression, only 78 (5%) and 81 (5%) HTgs had relative expression above 10 in *E. subterraneus* in samples of day 1 and day 10, respectively*.* Comparing relative expression in day 1 and day 10 of *E. subterraneus* primordial growth revealed that HTgs showed a significant higher relative expression in day 10 (Wilcox Signed Rank test, $\bar{x}$ = 3.7, $\bar{y}$ = 3.6, $p < 0.001$), while HTgs exhibited a significantly lower expression in day 10 of primordial growth (Wilcox Signed Rank test, $\tilde{x}$ = 0.2, $\tilde{y}$ = 0.3, $p < 0.001$).

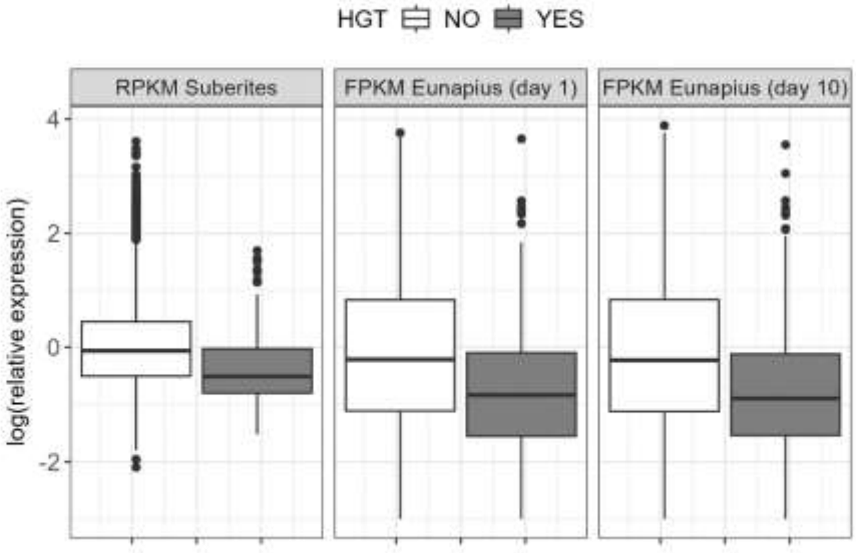
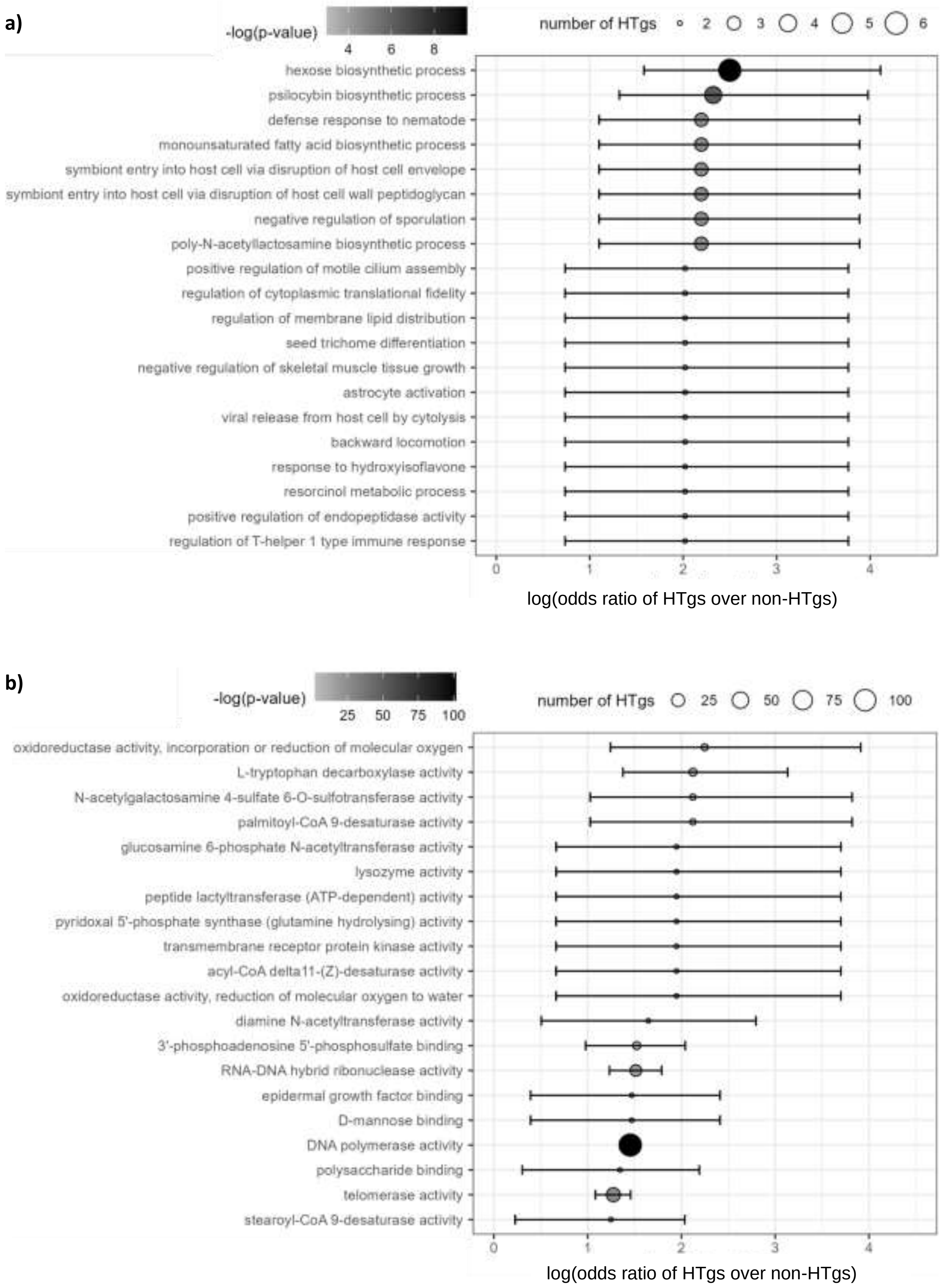


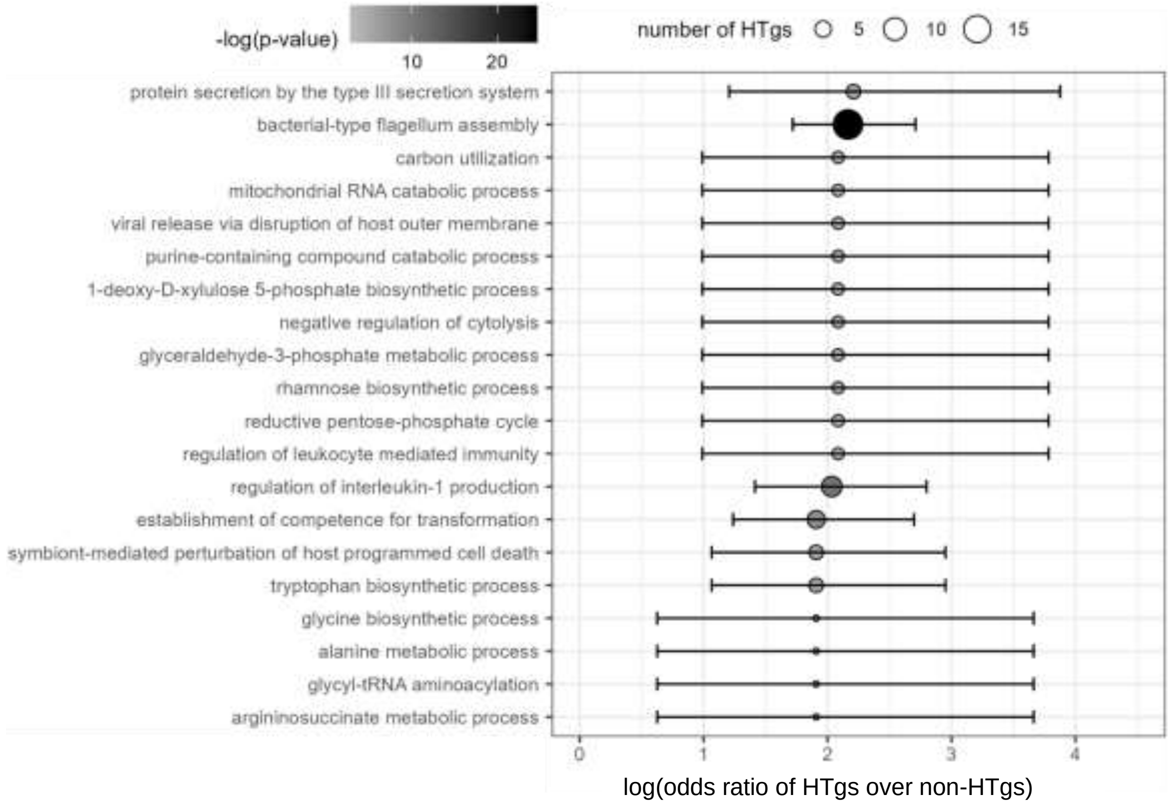
Figure 10. The distribution of relative expression of HTgs and non-HTgs. The values are logarithmically scaled. For *S. domuncula* only one RNAseq library was available, while for *E. subterraneus* RNAseq libraries from day 1 and day 10 of sponge primorph growth were analyzed.

4.6. Functional analysis

I also analyzed the potential functions of determined HTgs by performing an enrichment analysis of GO term annotations. The manually examined molecular function and biological process GO terns exhibiting the strongest association with HTgs are shown in **Figure 11**.


a)
-log(p-value)
4
6
8
number of HTgs
2
3
4
5
6
hexose biosynthetic process
psilocybin biosynthetic process
defense response to nematode
monounsaturated fatty acid biosynthetic process
symbiont entry into host cell via disruption of host cell envelope
symbiont entry into host cell via disruption of host cell wall peptidoglycan
negative regulation of sporulation
poly-N-acetyllactosamine biosynthetic process
positive regulation of motile cilium assembly
regulation of cytoplasmic translational fidelity
regulation of membrane lipid distribution
seed trichome differentiation
negative regulation of skeletal muscle tissue growth
astrocyte activation
viral release from host cell by cytolysis
backward locomotion
response to hydroxyisoflavone
resorcinol metabolic process
positive regulation of endopeptidase activity
regulation of T-helper 1 type immune response
0
1
2
3
4
log(odds ratio of HTgs over non-HTgs)
b)
-log(p-value)
25
50
75
100
number of HTgs
25
50
75
100
oxidoreductase activity, incorporation or reduction of molecular oxygen
L-tryptophan decarboxylase activity
N-acetylgalactosamine 4-sulfate 6-O-sulfotransferase activity
palmitoyl-CoA 9-desaturase activity
glucosamine 6-phosphate N-acetyltransferase activity
lysozyme activity
peptide lactyltransferase (ATP-dependent) activity
pyridoxal 5'-phosphate synthase (glutamine hydrolysing) activity
transmembrane receptor protein kinase activity
acyl-CoA delta11-(Z)-desaturase activity
oxidoreductase activity, reduction of molecular oxygen to water
diamine N-acetyltransferase activity
3'-phosphoadenosine 5'-phosphosulfate binding
RNA-DNA hybrid ribonuclease activity
epidermal growth factor binding
D-mannose binding
DNA polymerase activity
polysaccharide binding
telomerase activity
stearoyl-CoA 9-desaturase activity
0
1
2
3
4
log(odds ratio of HTgs over non-HTgs)

c)


-log(p-value)
10
20
number of HTgs
5
10
15
protein secretion by the type III secretion system
bacterial-type flagellum assembly
carbon utilization
mitochondrial RNA catabolic process
viral release via disruption of host outer membrane
purine-containing compound catabolic process
1-deoxy-D-xylulose 5-phosphate biosynthetic process
negative regulation of cytolysis
glyceraldehyde-3-phosphate metabolic process
rhamnose biosynthetic process
reductive pentose-phosphate cycle
regulation of leukocyte mediated immunity
regulation of interleukin-1 production
establishment of competence for transformation
symbiont-mediated perturbation of host programmed cell death
tryptophan biosynthetic process
glycine biosynthetic process
alanine metabolic process
glycyl-tRNA aminoacylation
argininosuccinate metabolic process
0
1
2
3
4
log(odds ratio of HTgs over non-HTgs)

d)

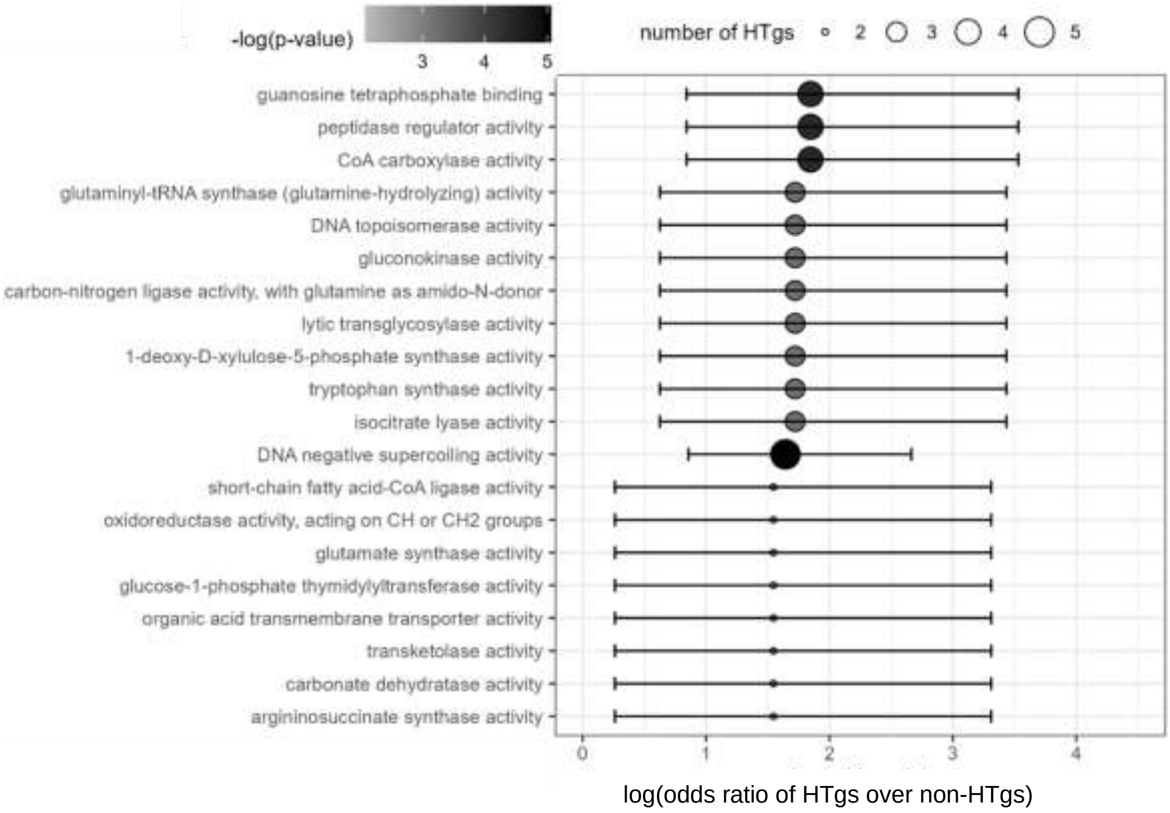

-log(p-value)
3
4
5
number of HTgs
2
3
4
5
guanosine tetraphosphate binding
peptidase regulator activity
CoA carboxylase activity
glutaminyl-tRNA synthase (glutamine-hydrolyzing) activity
DNA topoisomerase activity
gluconokinase activity
carbon-nitrogen ligase activity, with glutamine as amido-N-donor
lytic transglycosylase activity
1-deoxy-D-xylulose-5-phosphate synthase activity
tryptophan synthase activity
isocitrate lyase activity
DNA negative supercoiling activity
short-chain fatty acid-CoA ligase activity
oxidoreductase activity, acting on CH or CH2 groups
glutamate synthase activity
glucose-1-phosphate thymidylyltransferase activity
organic acid transmembrane transporter activity
transketolase activity
carbonate dehydratase activity
argininosuccinate synthase activity
0
1
2
3
4
log(odds ratio of HTgs over non-HTgs)

Figure 11. Enrichment analysis results of GO terms most significantly associated with HGT. Whiskers depict 95% confidence intervals for the odds ratio (OR). Odds ratios are logarithmically scaled. The size of the circle is proportional to the number of HTgs annotated with the corresponding GO term. The shade of grey represents a negative base-10 logarithmic value of the p-value of the Fisher's exact test for each GO term. **a, b)** GO terms of molecular function and biological process in *S. domuncula*, respectively; **c, d)** GO terms of molecular function and biological process in *E. subterraneus*, respectively.

Within GO terms of biological processes, the most pronounced GO term in *E. subterraneus* HTgs was the bacterial-type flagellum assembly. Annotated genes linked to bacterial-type flagellum assembly lie on 9 scaffolds with the median length of 75,351 bp, they contain up to 19 introns with the median of 3 introns per gene. Furthermore, these genes were not expressed in the analyzed RNAseq libraries. Other notable examples were: regulation of interleukin-1 production, establishment of competence for transformation, symbiont-mediated perturbation of host programmed cell death, tryptophan biosynthetic process, and protein secretion by the type II secretion system. Except genes annotated with the regulation of interleukin-1 production GO term, all genes related to mentioned GO terms were not expressed in the analyzed libraries. Moreover, most of these genes did not contain intronic regions and are relatively short. Two genes linked with regulation of interleukin-1 production had a higher level of relative expression – 42 and 21 FPKM in day 1 of primorph growth, and 27 and 16 in day 10 FPKM, respectively.

GO terms of molecular function in *E. subterraneus* were all linked to a relatively small number of HTgs (less than five), which were mostly not expressed in the analyzed libraires. One gene, linked to carbonate dehydratase activity, had a relative expression level of 32 and 31 FPKM in in day 1 and day 10 of primorph growth, respectively.

In *S. domuncula*, the analyzed GO terms of biological processes were linked to a small number of genes (less than six). One half of these genes do not contain introns. None of the gene had the expression level higher than 5 RPKM.

Within the GO terms of molecular function in *S. domuncula,* the GO term for DNA polymerase activity dominated with 109 linked genes located on 81 scaffolds. However, none of these genes had the expression level higher than 5 RPKM. Other abundant GO terms included telomerase activity, RNA-DNA hybrid ribonuclease activity and 3'-phosphoadenosine 5'-phosphosulfate binding. They were all attributed to genes which are spread on several scaffolds, with expression levels not higher than 5 RPKM.

4.7. Detecting homologues of horizontally transferred genes

After performing BLAT similarity search against the available sponge genomes, I filtered the resulting alignments for E-value, calculated the length percentage of HTgs aligning with a genomic region for each of the sponge and presented the results as a heatmap (**Figure 12.**). *S. domuncula* HTgs aligns best with *Tethya californiana*. Other sponge species showed an observable amount of clustered HTgs. *Oopsacas minuta* showed the fewest alignment with *S. domuncula*. On the other hand, *E. subterraneus* HTgs aligned best with *Ephydatia muelleri*. Other sponge species showed an observable amount of clustered HTgs. *Oopsacas minuta* exhibited the fewest alignment with *E. subterraneus*.

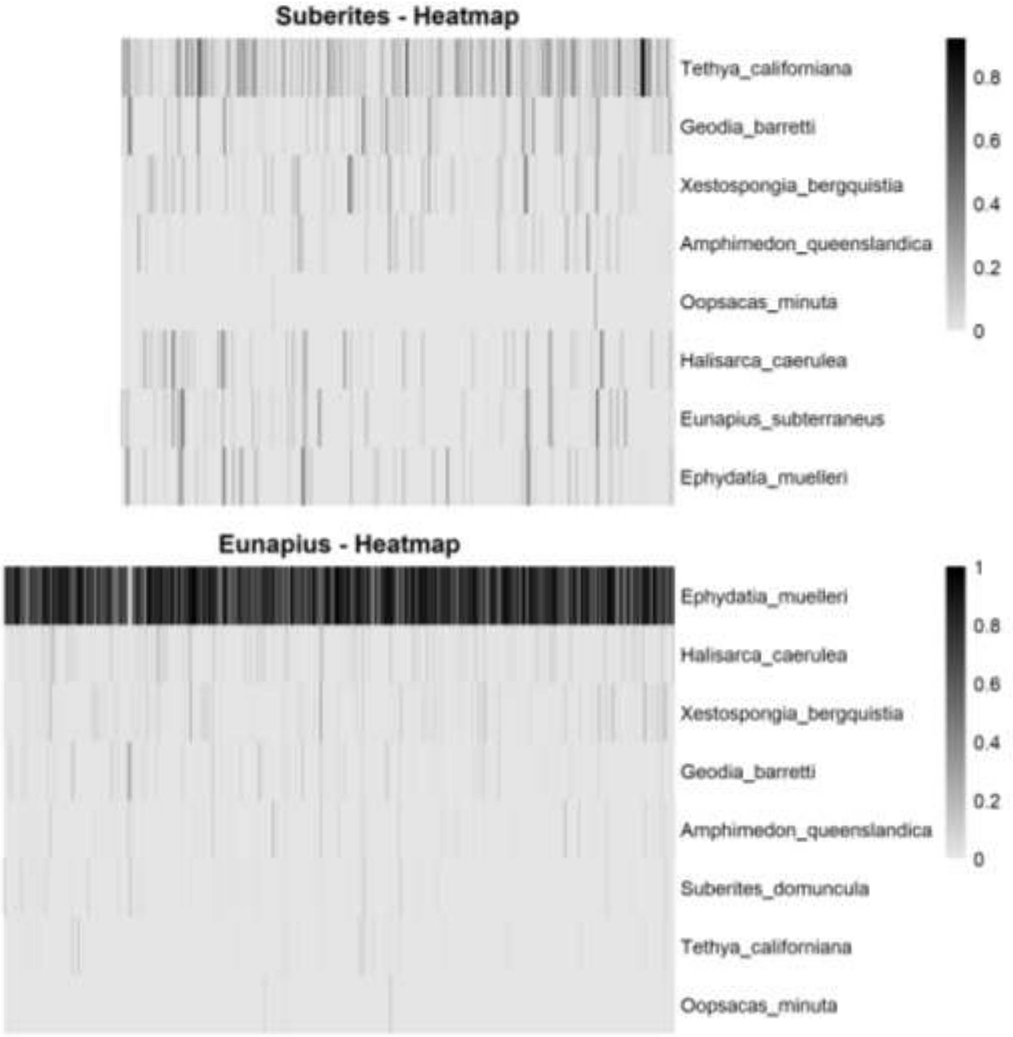


Figure 12. Detecting homologues of HTgs in available sponge genomes. Heatmaps represent the length percentage of HTgs aligning with sponge genomes for *S. domuncula* (up) and *E. subterraneus* (down). The length percentage of cumulative alignment is depicted with shades of grey.

A **c**ladogram derived from the NCBI Taxonomy Browser is shown in **Figure 13.** The species *E. subterraneus* is most related to *Ephydatia muelleri*, while *S. domuncula* is most related to *Tethya californiana*. *Oopsacas minuta* is the most distant sponge species to both *S. domuncula* and *E. subterraneus*. This information was used to identify most conserved HTgs shown on the **Figure 14.**

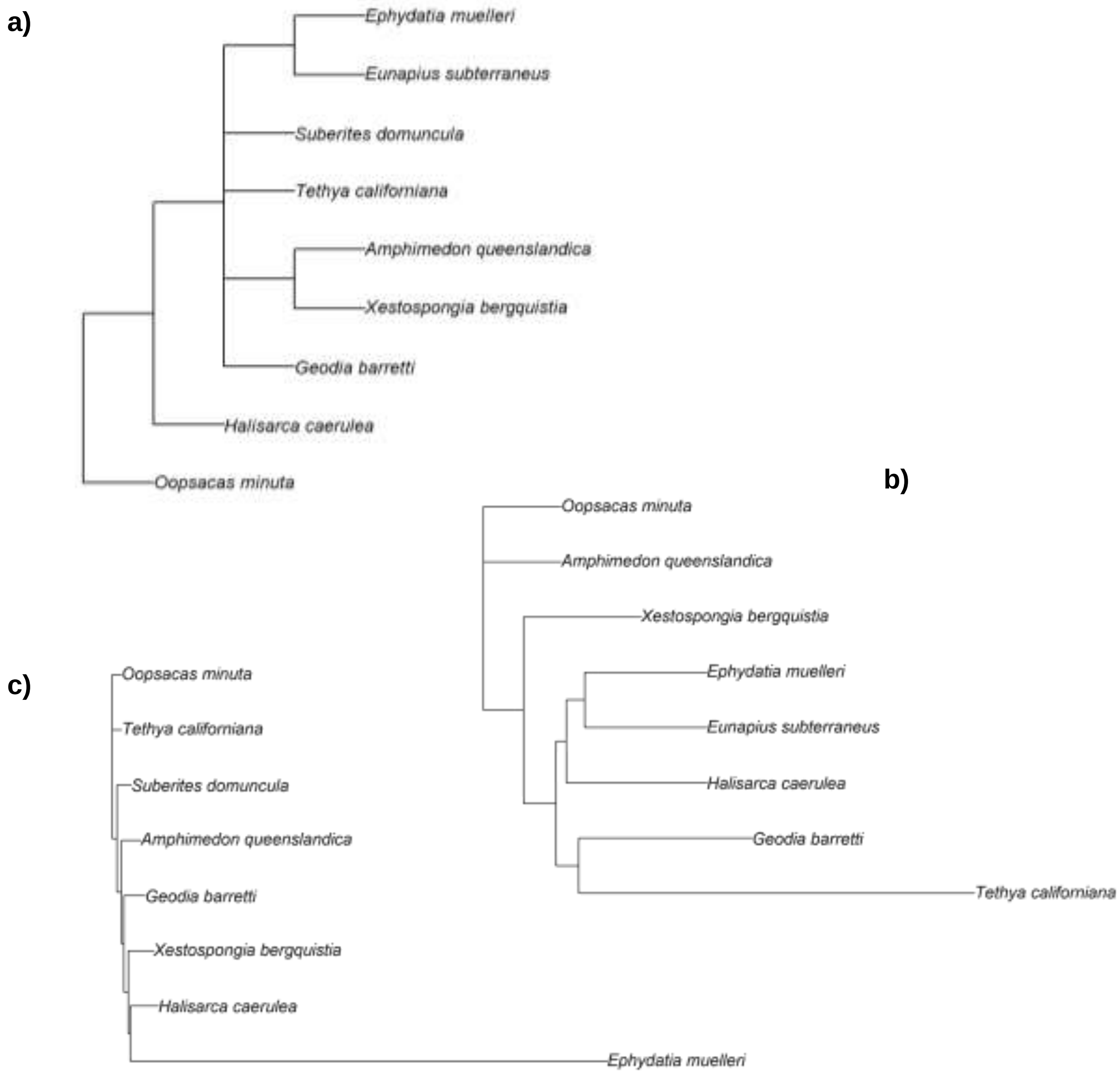


Figure 13. Sponge species relation. a) Cladogram of sponge species derived from NCBI Taxonomy Browser, **b)** dendrogram of sponge species based on homologues of HTgs found in *S. domuncula*, **c)** dendrogram of sponge species based on homologues of HTgs found in *E. subterraneus*.

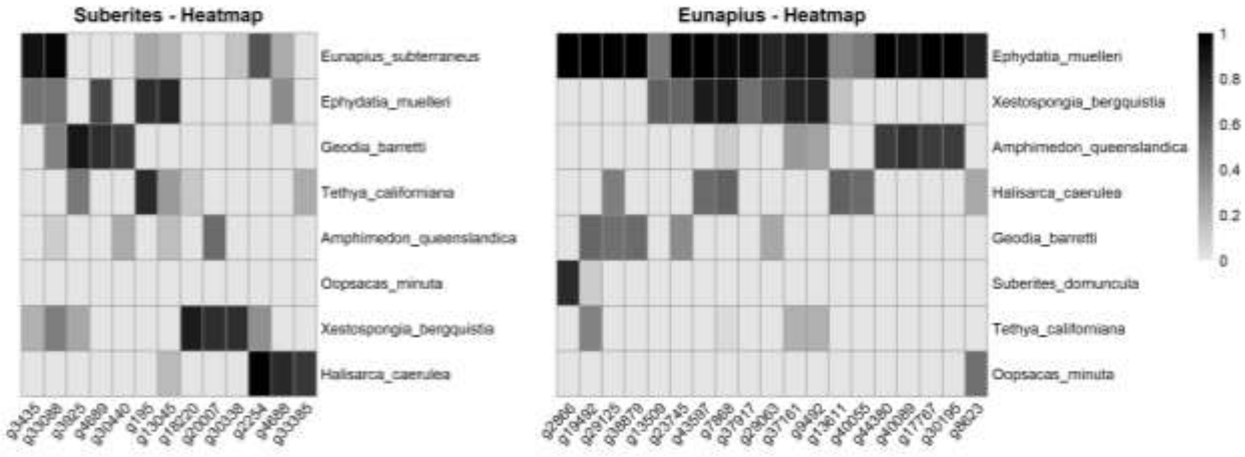


Figure 14. Detecting most conserved homologues of HTgs. The plot shows HTgs that aligned with more than one sponge genome, with at least one alignment covering more than 50% of HTgs, excluding the most related sponge: *Tethia californiana* for *S. domuncula* and *Ephydatia muelleri* for *E. subterraneus*. Heatmap representing the length percentage of HTgs that align with available sponge genomes for *S. domuncula* (left) and *E. subterraneus* (right). The percentage of cumulative alignment length is depicted with shades of grey.

Conserved HTgs in *S. domuncula* showed a scarce association with GO terms. The exception to this was the “involvement in DNA metabolism” GO term. In *E. subterraneus*, most conserved HTgs were associated with several enriched HTG GO terms, such as DNA metabolism and metabolic cycles like urea cycle and tricarboxylic acid cycle. However, none of the conserved HTgs were associated with bacterial-type flagellum assembly or interleukin-1 production.

5. DISCUSSION

Horizontal gene transfer (HGT) has been recorded in prokaryotes since the second half of the 20th century. Its role in acquiring variability, necessary for adaptation to new conditions, was repeatedly demonstrated. Initially, research focused on the spread of antibiotic resistance through horizontal transfer (**Akiba et al. 1960**), but with the recognition of the widespread nature of this phenomenon, interest expanded to studies in evolution and ecology (**Isbell et al. 2013**). Horizontal gene transfer in eukaryotes has long been neglected due to its rare occurrence and the lack of immediately apparent phenotypes. The distinct gene architecture, compartmentalization of the eukaryotic cell, multicellularity, and separation of germline cells from somatic cells reduce the efficiency of HGT from prokaryotes to eukaryotes. Sponges are a good model for studying HGT for several reasons. First, the low interdependence and specialization of sponge tissues increase interaction with endosymbionts: prokaryotes are brought into close contact with the progenitor cells of sponge gametes, increasing the likelihood of HGT in the germline and the retention of acquired genes across successive generations. Furthermore, sponges are of significant ecological importance as filter feeders, constantly interacting with microbial communities.

The limited exploration of HGT in eukaryotes is not solely due to its low occurrence rate in nature but also due to the increased focus on a small number of model organisms. This represents a significant challenge because HGT research relies on the availability of annotated genome sequences. Currently, the NCBI database lists 117 publicly available sponge genomes of varying quality and completeness, but only 8 of these are annotated reference genomes. In my research, I used the assembled genomes of *Suberites domuncula* and *Eunapius subterraneus*. At the beginning of the analysis, I processed these genomes by assessing their quality and removing contaminants. Compared to the annotated reference sponge genomes available in the NCBI database, the genomes initially showed moderate contiguity, rendering the analysis more difficult.

In the analysis of HGT from prokaryotes to sponges, I placed a significant emphasis on minimizing false positive HTgs, which was reflected in frequent filtering and the application of stringent thresholds. I began by filtering the initial genomes for contamination, employing relatively strict MEGAN parameters. For protein-based detection, I applied an AI threshold and an additional alignment length threshold of 150 amino acids. For nucleotide-based detection, in addition to standard E-value filtering, I removed regions that overlapped with any eukaryotic regions. Given that genome fragmentation was a major issue, particularly in *E. subterraneus*, I filtered horizontally transferred genes (HTgs) based on their genomic positions after identifying the union of protein-based and nucleotide-based HTgs. Scaffolds that were too short or those in which HTgs comprised the majority (≥ 50%) of the gene content were excluded from further analysis, as they could represent contamination missed by MEGAN. While no further filtering was performed, the interpretation of results considered several factors, including the number of introns, gene length, intron and exon lengths, relative expression, and gene ontology associated with individual genes.

While multiple filtering steps increased the likelihood of missing some HTgs, I considered this necessary given the abundance of sponge endosymbionts that could have contaminated the sequencing sample and the moderate quality of the assembled genomes. The initial filtering of the assembled genomes for viral and prokaryotic contamination removed 3% and 6% of the *S. domuncula* and *E. subterraneus* genome, respectively. The contamination filtering was based on the alignment of scaffold regions with the non-redundant protein database. Due to the prokaryotic origin of putative HTgs, these methods may remove a significant number of bona-fide HTgs and classify their scaffolds as contamination. I attempted to control for this by adjusting the default MEGAN parameters. This included setting the scaffold coverage threshold to 75%, lowering the probability of false-positive contamination filtering. Accordingly, the differences in length distributions between contaminated and uncontaminated scaffolds suggest accurate contamination removal.

The primary method for detecting HGT involves comparing deviations in the phylogenetic distribution of genome segments and identifying sequences that deviate from expected patterns. Confirming a suspected HGT event requires identifying a topological incongruence between a well-supported gene tree and the established species tree (**Lacroix & Citovsky 2016**). Although it is unlikely for most organisms within a group to lose a gene, deviations from phylogenetic distribution are not definitive evidence of HGT, and the phylogenetic approach must be supported by other methods. The phylogeny of sponges remains a subject of ongoing debate, particularly regarding the branching patterns of their major groups and their relationships to other non-bilaterian animals (**Wörheide et al. 2012**). Therefore, this study did not rely on phylogenetics but instead employed methods based on sequence similarity and alignments. While these methods do not provide define evidence for HGT, I believe that combining the protein-based and nucleotide-based approaches along with applying stringent filtering thresholds successfully removed most false positives HTgs, which was my primary focus. As such, this research represents a survey of potential HTgs in sponge genomes, rather than definite evidence for HGT events.

Detection of HGT using the protein and nucleotide approaches complements each other. Due to the degeneracy of the nucleotide code, a greater accumulation of mutations in a gene is required for detection at the amino acid sequence level. Thus, nucleotide-based methods are generally limited to detecting recent HGT events, while protein-based methods can detect HGT that occurred earlier in evolutionary history. The higher number of HTgs detected via the protein-based pipeline suggests earlier HGT events. In nucleotide sequences, it is likely that enough mutations have accumulated to render alignment with related prokaryotic genomes undetectable under the applied filtering conditions – a significant number of ancient HTgs could have adapted sufficiently to the sponge genomes so the detection methods based on sequence similarity could no longer differentiate them from non-HTgs.

A comparison of the studied sponges reveals that a smaller number of HTgs was detected in *E. subterraneus*. It would be overly simplistic to interpret this as evidence of a lower rate of HGT in this sponge due to its habitat in freshwater caves, where life is sparse. Without metagenomic studies of the cave systems in the Croatian karst, such a hypothesis cannot be substantiated. In fact, some caves have been shown to exhibit exceptionally high biodiversity (**Moutaouakil et al. 2024**). A significant obstacle to drawing conclusions is the fragmentation of the genome, which makes it difficult to accurately assess the extent of HGT in *E. subterraneus*.

The analysis of exon characteristics of detected HTgs revealed that they generally contain fewer exons, which aligns with expectations, particularly in the context of HGT from prokaryotes to eukaryotes. Eukaryotes are characterized by a large number of introns of varying length. In contrast, the prokaryotic gene expression system mostly lacks introns, as RNA molecules are not processed after transcription, and transcription and translation occur simultaneously. Following HGT from prokaryotes to eukaryotes, one way for the gene architecture to adapt to the host system is through the insertion of introns. The distributions of per-gene number of exons in analyzed HTgs and non-HTgs differ significantly, with HTgs having a slightly smaller per-gene number of exons. Unfortunately, due to the limited research on sponges, there is little data on the rate of genomic changes in sponges, particularly regarding intron insertion rates in genes acquired via HGT. The observed difference in the range of exon numbers between HTgs and non-HTgs in each sponge may suggest similar timelines of HGT events. Since both sponges belong to the class *Demospongiae*, it is plausible that these events occurred in a common ancestor. However, this hypothesis is contradicted by the analysis of potential homologues, which identified few shared homologues between the two studied sponges. Notably, genes with five or less exons were significantly more common among HTgs compared to non-HTgs in both species. This could support the hypothesis of foreign origin for these genes, as it suggests an initial lack of introns that are gradually introduced over evolutionary time. The low expression levels of such genes might seem to contradict their evolutionary integration, as it implies that these genes have no functional role in the sponge genome due to a lack of transcription. However, it is possible that these genes were simply not expressed in the analyzed sponge samples or that the expression could not be detected due to the isolation of solely polyadenylated RNAs.

I found significant differences in gene, exon and intron lengths among HTgs and non-HTgs. The significantly shorter introns in *E. subterraneus* and significantly longer exons in both sponges are not surprising. The shortness of introns in *E. subterraneus*, analogous to the smaller number of introns, suggests an initial lack of introns, which are gradually incorporated over evolutionary time. Long exons could be influenced by the small number of inserted introns in the ancestral prokaryotic gene, resulting in longer exonic parts. Longer introns are important in eukaryotes because, in addition to transcription, introns play a role in the modular evolution of the gene product by keeping exons apart and lowering their recombination linkage. This has been shown in protein-coding genes, where the presence of introns promotes recombination between exons, termed exon shuffling, leading to the creation of new functional products while preserving domain functionality (**Patthy 2021**). This hinders recombination within exons, which can disrupt the function of the domain they encode and consequently affect the gene product. All in all, the detected HTgs contained fewer and shorter introns, likely resulting in longer exons, which implies adaptation to eukaryotic expression system is a long-lasting in process.

Interestingly, in *S. domuncula* introns were significantly longer. Together with codon usage, which showed non-significant difference with codon usage between HTgs and non-HTgs, this would implicate higher rate of adaptation to sponge expression system. Low expression data did not agree with this although questionable quality of Nanopore reads could be the reason why certain low expression genes went undetected.

Lastly, HTgs in *S. domuncula* are significantly longer due to both significantly longer introns and exons. However, why HTgs are longer than non-HTgs in *E. subterraneus* is not so clear. Shorter introns and longer exons would imply that the length of HTgs is simply the result of opposing trends which have, in total, led to significantly longer HTgs. Greater gene length is costly to maintain, so based on the current data, no clear explanation can be inferred for why the greater length of HTgs would be evolutionarily advantageous.

The GC content results support the foreign origin of HTgs. The comparison of GC content between HTgs and non-HTgs shows a significant distinction, which was expected given the distribution pattern of GC content and the large number of genes used in the statistical tests. However, GC content does not necessarily confirm that HTgs are of prokaryotic origin, as prokaryotes exhibit a very broad range of GC content, with distributions ranging from 20% to 70% – this wide range of GC content has also been detected in sponge microbial communities (**Storey et al. 2020**). Nevertheless, the GC content distribution of HTgs in both studied sponge species displays several different peaks, supporting the hypothesis of a heterogeneous origin for HTgs, potentially due to the diverse microbial communities that sponge species harbor. The heterogeneity of GC content distribution is less pronounced in *E. subterraneus*, which might be caused by genome fragmentation or a significant amount of false-negative HTgs.

The analysis of codon usage revealed that, overall, codon usage discriminates HTgs from non-HTgs in *E. subterraneus*, while no such differences were found in *S. domuncula*. Trends were observed when analyzing individual codons, especially in the species *E. subterraneus*, where codons more homogeneous with regard to the GC or AT bases were used. Although the trend favors AT base pairs, it is interesting to note that all codons containing the CG dinucleotide show a higher usage in HTgs in *E. subterraneus*. This might interplay with CpG methylation system in sponges, but the importance of CpG methylation within CDS is still under inspection in mammals (**Creasey & Tauber 2024**), with scarce information on CpG methylation in sponges. By now it has been shown that *A. queenslandica* has a hypermethylated genome when compared to vertebrates (**de Mendoza et al. 2019**). Furthermore, in *Ephydatia muelleri*, close relative of *E. subterraneus*, the presence and functional importance of a methyl-cytosine binding domain (MBD) containing protein has been proved (**Cramer et al. 2017**). This would explain the general tendency of non-HTgs towards AT bases and once again classify HTgs in *S. domuncula* as more adapted to sponge translational system than HTgs in *E. subterraneus.* The difference in codon usage observed and prevalent CG dinucleotide in *E. subterraneus* would suggest recent HGT events or a specific regulatory function.

While non-HTgs exhibited significantly higher expression in both species, a substantial number of HTgs was expressed in both *E. subterraneus* libraries. This number was significantly lower for the RNA library isolated from *S. domuncula*. Considering that the utilized RNA libraries were not specifically created for the purpose of this thesis, as well as the relatively low quality of *S. domuncula* ONT reads, the expression analysis should be treated with caution. On the one hand, it is possible that a significant number of unexpressed HTgs are generally not transcribed. On the other hand, unexpressed HTgs may code for regulatory RNAs lacking polyadenylated 3' ends which were not sequenced in these libraries. Finally, it is possible that HTgs represent protein-coding genes that, due to insufficient adaptation to the sponge's expression system, exhibit extremely low and undetectable expression.

The functional analysis and homologue analysis of HTgs support the functional importance and evolutionary conservation of HTgs in sponges. The functional analysis was based on strong matches with a database of proteins with known functions and could suggest the association of a gene with a specific function. However, these associations should be confirmed by experimental data. It should not be ignored that HTgs most significantly associated with several GO terms did not exhibit any expression, especially in *S. domuncula*. In *E. subterraneus*, a slightly higher number of these genes exhibited significant expression in both analyzed RNA libraries. These genes were associated with functions related to the GO term "DNA biosynthetic process" such as DNA polymerase regulation, endonuclease activity, and telomerase activity. Additionally, functions related to cycles like the urea cycle and tricarboxylic acid cycle were also prominent, suggesting possible metabolic roles.

The results of the homologues analysis suggest a notable degree of conservation of HTgs in analyzed sponge species. Moreover, based on the amount of conserved HTgs, it was possible to cluster analyzed sponges and compare the obtained clusters with the taxonomic relationships. This enables comparisons between general sponge relatedness and species clustering resulting from the presence of HTgs.

The analysis of conservation of HTgs in *S. domuncula* revealed clustering that closely resembles the taxonomic tree of related sponges. Moreover, the heatmaps showed that not a lot of genes are preserved in many sponge species but in fact the conservation is mostly pairwise among the analyzed species. These results suggest that the majority of detected HGT events in *S. domuncula* took place in the common ancestor of the analyzed species. This scenario suggests that the pool of acquired genes later evolved in their corresponding host, producing patterns which could successfully reconstruct the taxonomic relations among the analyzed sponge species.

In contrary, the taxonomic relationship reconstruction based on HTgs was not possible in *E. subterraneus.* Relations among *E. subterraneus* and sponges other than *Ephydatia muelleri* were non-reconstructable based on the detection of HTgs found in *E. subterraneus.* Only a small fraction of HTgs could have been common to all sponges, but it cannot be assumed to which extant. Interestingly, the vast majority of HTgs from *E. subterraneus* were present in *E. muelleri*. The high degree of similarity and close relationship between these two sponges, which has already been shown in similar research (**Bodulić 2020**), was most probably the cause of this finding.

GO terms associated with functions of most conserved HTgs in *S. domuncula* were "DNA biosynthetic process" and "binding". HTgs associated with GO terms that could represent potential contamination, such as "bacterial-like flagellum assembly," were not conserved in other sponge species and did not exhibit any expression in the analyzed RNA libraries. However, these genes spanned multiple scaffolds, while containing several relatively long introns. As such, the origin and potential roles of these genes remains unclear.

The GO enrichment analysis of most conserved HTgs in both sponges revealed an increased presence of HTgs associated with the "DNA biosynthetic process" function. Genes associated with this function are widespread and essential for living organisms, making them present in every sponge endosymbiont which could be a potential donor of HTgs. In a similar study of HGT in the *Amphimedon queenslandica* genome, HTgs involved in secondary metabolite metabolism and genes related to stress conditions were detected. The HTgs found in my research did not exhibit a significant enrichment of the stated GO terms. This could be a consequence of different HGT patterns observed in these sponge species. On the other hand, this could result from differing genome assembly quality, as well as a more stringent filtering strategy implemented in my analysis. This could also be a consequence of relatively weak contamination filtering performed in the stated study of HGT in *Amphimedon queenslandica*.

6. CONCLUSION

Horizontal gene transfer represents the process of sudden acquisition of new traits that can bridge evolutionarily distant groups. While HGT has been extensively studied in prokaryotes, the number of studies on HGT in eukaryotes is substantially lower. So far, only one systematic study on the detection and analysis of HGT in sponges has been published. For this reason, I focused on the detection and analysis of HGT in two sponge species, *Suberites domuncula*, a cosmopolitan marine sponge, and *Eunapius subterraneus*, a Croatian endemic of karst groundwater. I identified a large number of horizontally transferred genes (HTgs) by combining different approaches. Through a series of filtering steps, I reduced the impact of contamination and increased the relevance of the obtained results.

The architecture of the detected HTgs was analyzed and found to differ significantly from other genes in terms of gene, exon and intron length. Significantly lower per-gene exon number, higher exon and lower intron lengths of HTgs could be explained by the prokaryotic origin of these genes, which exhibited a varying incorporation into the organization of sponge genomes. This is further pronounced by the multimodal GC content distribution and low expression of HTgs. The clustering of genes by codon usage found exclusively in *E. subterraneus* suggests notable degrees of adaptation to the sponge translational system. The relevance of the detected HTgs was confirmed by a combination of similarity analysis of the detected HTgs in related sponges, as well as functional and expression analyses. Enrichment analysis of HTgs revealed a significant presence of regulating proteins involved in DNA metabolism. It has not been clarified whether this regulation occurs at the protein level or through gene silencing, but their conservation in related sponges suggests their functional importance.

Finally, pattern of HTgs conserved in *S. domuncula* successfully reconstructed the taxonomic relation of analyzed sponges. This implies the presence of most of these genes in the common ancestor of analyzed sponges, while this was not the case with *E. subterraneus*. This finding aligns with several features that suggest the adaptation of *S. domuncula* HTgs to sponge translational system, such as intron length and codon usage.

7. LITERATURE

Acuña, R., Padilla, B. E., Flórez-Ramos, C. P., Rubio, J. D., Herrera, J. C., Benavides, P., ... & Rose, J. K. (2012). Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proceedings of the National Academy of Sciences*, *109*(11), 4197-4202. doi: 10.1073/pnas.1121190109

Akiba, T., Koyama, K., Ishiki, Y., Kimura, S., & Fukushima, T. (1960). On the mechanism of the development of multiple drug-resistant clones of Shigella. *Japanese journal of microbiology*, *4*(2), 219-227. doi: 10.1111/j.1348-0421.1960.tb00170.x

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410. doi: 10.1016/S0022-2836(05)80360-2

Anindya, R. (2022). Cytoplasmic DNA in cancer cells: several pathways that potentially limit DNase2 and TREX1 activities. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, *1869*(8), 119278. doi: 10.1016/j.bbamcr.2022.119278

Archibald, J. M. (2015). Endosymbiosis and eukaryotic cell evolution. *Current Biology*, *25*(19), R911-R921. doi: 10.1016/j.cub.2015.07.055

Arias, C. F., Acosta, F. J., Bertocchini, F., Herrero, M. A., & Fernández-Arias, C. (2022). The coordination of anti-phage immunity mechanisms in bacterial cells. *Nature communications*, *13*(1), 7412. doi: 10.1038/s41467-022-35203-7

Arnqvist, G., & Rowe, L. (2005). *Sexual conflict* (Vol. 27). Princeton university press.

Avery, O. T., MacLeod, C. M., & McCarty, M. (1944). Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *Die Entdeckung der Doppelhelix*, *97*. doi: 10.1084/jem.79.2.137

Bağcı, C., Patz, S., & Huson, D. H. (2021). DIAMOND+ MEGAN: fast and easy taxonomic and functional analysis of short and long microbiome sequences. *Current protocols*, *1*(3), e59. doi: 10.1002/cpz1.59

Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Hocking, T., Schwendinger, B. (2024). data.table: Extension of `data.frame`. R package version 1.16.2. https://CRAN.R-project.org/package=data.table

Bedek, J., Bilandžija, H. i Jalžić, B. (2008). Ogulinska špiljska spužvica Eunapius subterraneus Sket et Velikonja, 1984, rasprostranjenost i ekologija vrste i staništa. *Modruški zbornik, 2* (2), 103-130. Preuzeto s https://hrcak.srce.hr/79074

Bell, J. J. (2008). The functional roles of marine sponges. *Estuarine, coastal and shelf science*, *79*(3), 341-353. doi: 10.1016/j.ecss.2008.05.002

Bergquist, P. R. (2001). Porifera (Sponges): Recent Knowledge and New Perspectives. *Encyclopedia of Life Sciences. John Wiley & Sons, New York.* doi: 10.1002/9780470015902.a0001582.pub2

Bergstrom, C. T., Lipsitch, M., & Levin, B. R. (2000). Natural selection, infectious transfer and the existence conditions for bacterial plasmids. *Genetics*, *155*(4), 1505-1519. doi: 10.1093/genetics/155.4.1505

Bhagwat, M., Young, L., & Robison, R. R. (2012). Using BLAT to find sequence similarity in closely related genomes. *Current protocols in bioinformatics*, *37*(1), 10-8. doi: 10.1002/0471250953.bi1008s37

Bilandžija, H., Bedek, J., Jalžić, B. i Gottstein, S. (2007). The morphological variability, distribution patterns and endangerment in the Ogulin cave sponge Eunapius subterraneus Sket & Velikonja, 1984 (Demospongiae, Spongillidae). *Natura Croatica, 16* (1), 1-17. Preuzeto s https://hrcak.srce.hr/13515

Bodulić, K. (2020). Računalna analiza dugih nekodirajućih RNA ogulinske špiljske spužvice (Eunapius subterraneus) (Diplomski rad). Zagreb: University of Zagreb, Faculty of Science. Retrieved from https://urn.nsk.hr/urn:nbn:hr:217:310016

Brigulla, M., & Wackernagel, W. (2010). Molecular aspects of gene transfer and foreign DNA acquisition in prokaryotes with regard to safety issues. *Applied microbiology and biotechnology*, *86*, 1027-1041. doi: 10.1007/s00253-010-2489-3

Brunet, T., & King, N. (2017). The origin of animal multicellularity and cell differentiation. *Developmental cell*, *43*(2), 124-140. doi: 10.1016/j.devcel.2017.09.016

Buchfink, B., Reuter, K., & Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature methods*, *18*(4), 366-368. doi: 10.1038/s41592-021-01101-x

Bushnell, B. (2014) BBMap: A Fast, Accurate, Splice-Aware Aligner. United States. (https://www.osti.gov/servlets/purl/1241166).

Chamberlain, S. & Szocs, E. (2013). taxize - taxonomic search and retrieval in R. F1000Research, 2:191. URL: https://f1000research.com/articles/2-191/v2

Chamberlain, S., Szoecs, E., Foster, Z., Arendsee, Z., Boettiger, C., Ram, K. … & Grenié, M. (2020) taxize: Taxonomic information from around the web. R package version 0.9.98. https://github.com/ropensci/taxize

Conaco, C., Tsoulfas, P., Sakarya, O., Dolan, A., Werren, J., & Kosik, K. S. (2016). Detection of prokaryotic genes in the Amphimedon queenslandica genome. *PLoS One*, *11*(3), e0151092. doi: 10.1371/journal.pone.0151092

Cramer, J. M., Pohlmann, D., Gomez, F., Mark, L., Kornegay, B., Hall, C., ... & Williams Jr, D. C. (2017). Methylation specific targeting of a chromatin remodeling complex from sponges to humans. *Scientific Reports*, *7*(1), 40674. doi: 10.1038/srep40674

Creasey, L. D., & Tauber, E. (2024). Interconnected Codons: Unravelling the Epigenetic Significance of Flanking Sequences in CpG Dyads. *Journal of Molecular Evolution*, *92*(3), 207-216. doi: 10.1007/s00239-024-10172-1

de Mendoza, A., Hatleberg, W. L., Pang, K., Leininger, S., Bogdanovic, O., Pflueger, J., ... & Lister, R. (2019). Convergent evolution of a vertebrate-like methylome in a marine sponge. *Nature ecology & evolution*, *3*(10), 1464-1473. doi: 10.1038/s41559-019-0983-2

Dessimoz, C., Margadant, D., & Gonnet, G. H. (2008). DLIGHT–lateral gene transfer detection using pairwise evolutionary distances in a statistical framework. In *Annual International Conference on Research in Computational Molecular Biology* (pp. 315-330). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-540-78839-3_27

Dyer, S. C., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Barrera-Enriquez, V. P., ... & Yates, A. D. (2025). Ensembl 2025. *Nucleic Acids Research*, *53*(D1), D948-D957. doi: 10.1093/nar/gkae1071

Eichinger, L., Pachebat, J. A., Glöckner, G., Rajandream, M. A., Sucgang, R., Berriman, M., ... & Kuspa, A. (2005). The genome of the social amoeba Dictyostelium discoideum. *Nature*, *435*(7038), 43-57. doi: 10.1038/nature03481

Elek, A., Kuzman, M., Vlahovicek, K. (2024). coRdon: Codon Usage Analysis and Prediction of Gene Expressivity. doi: 10.18129/B9.bioc.coRdon

Eme, L., Gentekaki, E., Curtis, B., Archibald, J. M., & Roger, A. J. (2017). Lateral gene transfer in the adaptation of the anaerobic parasite Blastocystis to the gut. *Current Biology*, *27*(6), 807-820. doi: 10.17632/pktp3hggf7.1.

Engelstädter, J. (2017). Asexual but not clonal: evolutionary processes in automictic populations. *Genetics*, *206*(2), 993-1009. doi: 10.1534/genetics.116.196873

Ereskovsky, A., Melnikov, N. P., & Lavrov, A. (2024). Archaeocytes in sponges: simple cells of complicated fate. *Biological Reviews*. doi: 10.1111/brv.13162

Erwin, D. H. (2015). Early metazoan life: divergence, environment and ecology. Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1684), 20150036. doi: 10.1098/rstb.2015.0036

Evidently AI Team (2024) Classification metrics guide: How to use classification threshold to balance precision and recall. *Evidently AI*. https://www.evidentlyai.com/classification-metrics/classification-threshold (Last updated: October 1, 2024)

Fukasawa, Y., Ermini, L., Wang, H., Carty, K., & Cheung, M.-S. (2020). LongQC: A Quality Control Tool for Third Generation Sequencing Long Read Data. G3: Genes, Genomes, Genetics, 10(4), 1193–1196. https://doi.org/10.1534/g3.119.400985

Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., & Stanke, M. (2024). BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Research*. doi: 10.1101/gr.278090.123

Gao, Z., Waggoner, D., Stephens, M., Ober, C., & Przeworski, M. (2015). An estimate of the average number of recessive lethal mutations carried by humans. *Genetics*, *199*(4), 1243-1254. doi:

Goldfarb, T., Kodali, V. K., Pujar, S., Brover, V., Robbertse, B., Farrell, C. M., ... & Murphy, T. D. (2025). NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Research*, *53*(D1), D243-D257. doi: 10.1093/nar/gkae1038

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., ... & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research*, *36*(10), 3420-3435. doi: 10.1093/nar/gkn176

Griffith, F. (1928) The significance of pneumococcal types. *Journal of Hygiene* 27, 113–159 doi: 10.1017/s0022172400031879

Guindon, S., & Perriere, G. (2001). Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. *Molecular biology and evolution*, *18*(9), 1838-1840. doi: 10.1093/oxfordjournals.molbev.a003972

Harcet, M., Bilandžija, H., Bruvo-Mađarić, B., & Ćetković, H. (2010). Taxonomic position of Eunapius subterraneus (Porifera, Spongillidae) inferred from molecular data–A revised classification needed?. *Molecular Phylogenetics and Evolution*, *54*(3), 1021-1027. doi: 10.1016/j.ympev.2009.12.019

Hotopp, J. C. D., Clark, M. E., Oliveira, D. C., Foster, J. M., Fischer, P., Torres, M. C. M., ... & Werren, J. H. (2007). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science*, *317*(5845), 1753-1756. doi: 10.1126/science.1142490

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome research*, *17*(3), 377-386. doi: 10.1101/gr.5969107

Isbell, F., Reich, P. B., Tilman, D., Hobbie, S. E., Polasky, S., & Binder, S. (2013). Nutrient enrichment, biodiversity loss, and consequent declines in ecosystem productivity. *Proceedings of the National Academy of Sciences*, *110*(29), 11911-11916. doi: 10.1073/pnas.1310880110

Johnston, C., Martin, B., Fichant, G., Polard, P., & Claverys, J. P. (2014). Bacterial transformation: distribution, shared mechanisms and divergent control. *Nature Reviews Microbiology*, *12*(3), 181-196. doi: 10.1038/nrmicro3199

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome research*, *12*(4), 656-664. doi: 10.1101/gr.229202

Klasson, L., Kambris, Z., Cook, P. E., Walker, T., & Sinkins, S. P. (2009). Horizontal gene transfer between Wolbachia and the mosquito Aedes aegypti. *BMC genomics*, *10*, 1-9. doi: 10.1186/1471-2164-10-33

Kolde, R. (2019). pheatmap: Pretty Heatmaps. R package version 1.0.12, https://CRAN.R-project.org/package=pheatmap

Kong, L. Z., Kim, S. M., Wang, C., Lee, S. Y., Oh, S. C., Lee, S., ... & Kim, T. D. (2023). Understanding nucleic acid sensing and its therapeutic applications. *Experimental & Molecular Medicine*, *55*(11), 2320-2331. doi: 10.1038/s12276-023-01118-6

Kowalczykowski, S. C., Dixon, D. A., Eggleston, A. K., Lauder, S. D., & Rehrauer, W. M. (1994). Biochemistry of homologous recombination in Escherichia coli. *Microbiological reviews*, *58*(3), 401-465. doi: 10.1128/mr.58.3.401-465.1994

Kumari, N., Kaur, E., Raghavan, S. C., & Sengupta, S. (2024). Regulation of pathway choice in DNA repair after Double-strand Breaks. *Current Opinion in Pharmacology*, 102496. doi: 10.1016/j.coph.2024.102496

Kuznetsov, D., Tegenfeldt, F., Manni, M., Seppey, M., Berkeley, M., Kriventseva, E. V., & Zdobnov, E. M. (2023). OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research*, *51*(D1), D445-D451. doi: 10.1093/nar/gkac998

Lacroix, B., Tzfira, T., Vainstein, A., & Citovsky, V. (2006). A case of promiscuity: Agrobacterium's endless hunt for new partners. *TRENDS in Genetics*, *22*(1), 29-37. doi: 10.1016/j.tig.2005.10.004

Lacroix, B., & Citovsky, V. (2016). Transfer of DNA from bacteria to eukaryotes. *MBio*, *7*(4), 10-1128. doi: 10.1128/mbio.00863-16

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., et al. (2013) Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol 9(8): e1003118. doi: 10.1371/journal.pcbi.1003118

Lederberg, J., & Tatum, E. L. (1946). Gene recombination in Escherichia coli. Nature, 158(4016). doi: 10.1038/158558a0

Lee, S., Lewis, D. E., & Adhya, S. (2018). The developmental switch in bacteriophage λ: a critical role of the Cro protein. *Journal of molecular biology*, *430*(1), 58-68.doi: 10.1016/j.jmb.2017.11.005

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100. doi: 10.1093/bioinformatics/bty191

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923-930. doi: 10.1093/bioinformatics/btt656

Matoničkin, I., Habdija, I & Primc-Habdija, B. (1998). Beskralješnjaci: biologija nižih avertebrata. *Školska knjiga*, Zagreb, 190-205.

Mitrikeski, P. T. (2013). Yeast competence for exogenous DNA uptake: towards understanding its genetic component. *Antonie Van Leeuwenhoek*, *103*(6), 1181-1207. doi: 10.1007/s10482-013-9905-5

Moutaouakil, S., Souza-Silva, M., Oliveira, L. F., Ghamizi, M., & Ferreira, R. L. (2024). A cave with remarkably high subterranean diversity in Africa and its significance for biodiversity conservation. *Subterranean Biology*, *50*, 1-28. doi: 10.3897/subtbiol.50.113919

Mukherjee, S., Seshadri, R., Varghese, N. J., Eloe-Fadrosh, E. A., Meier-Kolthoff, J. P., Göker, M., ... & Kyrpides, N. C. (2017). 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature biotechnology*, *35*(7), 676-683. doi: 10.1038/nbt.3886

Müller, W. E., Wiens, M., Batel, R., Steffen, R., Schröder, H. C., Borojevic, R., & Custodio, M. R. (1999). Establishment of a primary cell culture from a sponge: primmorphs from Suberites domuncula. *Marine Ecology Progress Series*, *178*, 205-219. doi: 10.3354/meps178205

O'Leary, N. A., Cox, E., Holmes, J. B., Anderson, W. R., Falk, R., Hem, V., ... & Schneider, V. A. (2024). Exploring and retrieving sequence and metadata for species across the tree of life with NCBI

Datasets. *Scientific data*, *11*(1), 732. doi: doi: 10.1038/s41597-024-03571-y. PMID: 38969627; PMCID: PMC11226681. Generated January 23, 2025

Oxford Nanopore Technologies. (2025). Dorado: Oxford Nanopore's Basecaller. GitHub. https://github.com/nanoporetech/dorado

Pagès, H., Aboyoun, P., Gentleman, R., DebRoy, S. (2024). Biostrings: Efficient manipulation of biological strings. R package version 2.74.1. doi: 10.18129/B9.bioc.Biostrings

Paradis, E., Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyzes in R. *Bioinformatics*, *35*, 526-528. doi:10.1093/bioinformatics/bty633

Patthy, L. (2021). Exon shuffling played a decisive role in the evolution of the genetic toolkit for the multicellular body plan of metazoa. *Genes*, *12*(3), 382. doi: 10.3390/genes12030382

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing,Vienna, Austria. https://www.R-project.org/

Ravenhall, M., Škunca, N., Lassalle, F., & Dessimoz, C. (2015). Inferring horizontal gene transfer. *PLoS computational biology*, *11*(5), e1004095. doi: 10.1371/journal.pcbi.1004095

Sayers, E. W., Beck, J., Bolton, E. E., Brister, J. R., Chan, J., Connor, R., ... & Pruitt, K. D. (2025). Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Research*, *53*(D1), D20-D29. doi: 0.1093/nar/gkae979

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch, I. Mizrachi. 2019. *GenBank. Nucleic Acids Res*, *48*, D84-D86. doi: 10.1093/nar/gky989

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., ... & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, *2020*, baaa062. doi: 10.1093/database/baaa062

Schönknecht, G., Chen, W. H., Ternes, C. M., Barbier, G. G., Shrestha, R. P., Stanke, M., ... & Weber, A. P. (2013). Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*, *339*(6124), 1207-1210. doi: 10.1126/science.1231707

Sorek, R., Zhu, Y., Creevey, C. J., Francino, M. P., Bork, P., & Rubin, E. M. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. *Science*, *318*(5855), 1449-1452. doi: 10.1126/science.1147112

Souiba, Z., Santín, A., Kader Yettefti, I., Abdessadek, M., & El Amraoui, B. (2023). Biochemical composition and antioxidant properties of the sponge Suberites domuncula from two regions of Moroccan Mediterranean waters. *International Aquatic Research*, *15*(4), 345-359. doi: 10.22034/IAR.2023.1994517.1495

Storey, M. A., Andreassend, S. K., Bracegirdle, J., Brown, A., Keyzers, R. A., Ackerley, D. F., ... & Owen, J. G. (2020). Metagenomic exploration of the marine sponge Mycale hentscheli uncovers multiple polyketide-producing bacterial symbionts. *MBio*, *11*(2), 10-1128. doi: 10.1128/mBio.02997-19

The UniProt Consortium (2025) UniProt: the Universal protein knowledgebase in 2025. *Nucleic Acids Research*, 2025, 53.D1: D609-D617. doi: 10.1093/nar/gkae1010

Upreti, C., Kumar, P., Durso, L. M., & Palmer, K. L. (2024). CRISPR-Cas inhibits plasmid transfer and immunizes bacteria against antibiotic resistance acquisition in manure. *Applied and Environmental Microbiology*, *90*(9), e00876-24. doi: 10.1128/aem.00876-24

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*.

Wickham, H., Vaughan, D., Girlich, M. (2024). tidyr: Tidy Messy Data. R package version 1.3.1, https://CRAN.R-project.org/package=tidyr

Wörheide, G., Dohrmann, M., Erpenbeck, D., Larroux, C., Maldonado, M., Voigt, O., ... & Lavrov, D. V. (2012). Deep phylogeny and evolution of sponges (phylum Porifera). *Advances in marine biology*, *61*, 1-78. doi: 10.1016/B978-0-12-387787-1.00007-6

Würtele, H., Little, K. C. E., & Chartrand, P. (2003). Illegitimate DNA integration in mammalian cells. *Gene therapy*, *10*(21), 1791-1799. doi: 10.1038/sj.gt.3302074

Yue, J., Sun, G., Hu, X., & Huang, J. (2013). The scale and evolutionary significance of horizontal gene transfer in the choanoflagellate Monosiga brevicollis. *BMC genomics*, *14*, 1-10. doi: 10.1186/1471-2164-14-729

Zhang, M., Zhang, T., Yu, M., Chen, Y. L., & Jin, M. (2022). The life cycle transitions of temperate phages: regulating factors and potential ecological implications. *Viruses*, *14*(9), 1904. doi: 10.3390/v14091904

Zhu, Q., Kosoy, M., & Dittmar, K. (2014). HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. *BMC genomics*, *15*, 1-18. doi: 10.1186/1471-2164-15-717

Zinder, N. D., & Lederberg, J. (1952). Genetic exchange in Salmonella. Journal of bacteriology, 64(5), 679-699. doi: 10.1128/jb.64.5.679-699.1952

CURRICULUM VITAE

Luka Buršić, born in 2000, attended Vodnjan Elementary School and the Science Gymnasium in Pula. At the same time, he attended the Ivan Matetić Ronjgov Music School in Pula and graduated in 2019 as a pianist. During his elementary and high school education, he participated in various national competitions in mathematics, biology, chemistry, the Croatian language, piano, and solfeggio.

In 2019, he enrolled in the undergraduate program in Molecular Biology at the Faculty of Science, University of Zagreb. In 2022, he began his graduate studies in Molecular Biology at the Faculty. During his studies, he worked as a physics demonstrator at the Faculty and later as a tour guide at the Nikola Tesla Technical Museum. He furthered his training through a phylogeny workshop in Mainz and an internship in the research group of Prof. Dr. Iva Tolić on mitotic spindle in tumor cells. In 2025, he is conducting his master's thesis in the research group of Prof. Dr. Kristian Vlahoviček.