# Identification and characterization of transposable elements in sponges (Porifera)

**Kuzman, Maja**

University of Zagreb

Faculty of Science

Maja Kuzman

# IDENTIFICATION AND CHARACTERIZATION OF TRANSPOSABLE ELEMENTS IN SPONGES (PORIFERA)

DOCTORAL DISSERTATION

Supervisor:

Prof. dr. sc. Kristian Vlahoviček

Zagreb, 2020

Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Maja Kuzman

# IDENTIFIKACIJA I KARAKTERIZACIJA POKRETNIH GENETIČKIH ELEMENATA U SPUŽVAMA (PORIFERA)

DOKTORSKI RAD

Mentor:

Prof. dr. sc. Kristian Vlahoviček

Zagreb, 2020

Ovaj je doktorski rad izrađen u Grupi za bioinformatiku pod vodstvom prof. dr. sc. Kristiana Vlahovičeka, u sklopu Sveučilišnog poslijediplomskog doktorskog studija Biologije pri Biološkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu.

I thank my mentor for supporting me whenever needed and including me in most of his collaborations. I can not imagine learning and experiencing as much as I have without the opportunities presented to me during this PhD.

The bioinformatics group and all its members had an enormous impact on both my professional and personal life in the previous five years. Rosa, Filip, Dunja, Antonio, and all the past members thank you first of all for your friendship, but also for patience, discussions and help with the thesis and bioinformatics but also life related topics in general. I learned a lot. :) I thank my students Paula, Kristian, Moreno, Eva, Zoe and Anamaria for their enthusiasm, motivation and hard work. It was a real pleasure working with all of you!

I thank my collaborators, especially Marina, Bojana, Guillaume, Petr, Paz and Rosa for providing me the opportunity to work with them. I learned a lot during those collaborations and they definitely enriched this PhD experience.

I am deeply grateful and very lucky to meet (and befriend) such amazing people during summer schools and workshops I attended. NGSchool is awesome. German, Kasia, Gienio, Andrey (and again Anamaria :) ) – thank you for your friendship, random thoughts, fruitful discussions and a lot of help with the thesis! Guys from IGNITE, thank you for including me to great workshops and free time activities!

I thank my „real-life“ friends for their support. Thank you Dora and Franke for useful comments on the thesis :) I especially thank my life long friends Iva and Korša for sharing the passion for biology and sharing life advice.

I thank my family; mom, dad and sister for an amazing life. You taught me by your own example that the most valuable traits a person should have are hard work, enthusiasm and good attitude and everything else will come along.

Finally, I thank Boris for making me try harder to become a better person every day. I can't imagine my life without you.

University of Zagreb
Faculty of Science
Department of Biology

Doctoral thesis

# IDENTIFICATION AND CHARACTERIZATION OF TRANSPOSABLE ELEMENTS IN SPONGES (PORIFERA)

MAJA KUZMAN

Transposable elements constitute a large portion of most eukaryotic genomes and play an important role in transcriptional regulation, shaping gene regulatory networks and genome evolution. Due to a lack of quality genome assemblies of non-model organisms, not much is known about the diversity, distribution and the role of transposable elements across Metazoa. In this thesis I present high quality genome assemblies for the sponge species *Eunapius subterraneus* and *Suberites domuncula*. I use computational genomics techniques to identify, characterize and compare transposable elements in the publicly available sponge genomes and in-house assembled genomes by analysing the abundance and conservation of the transposable elements in different sponge species. I assess the impact of transposable elements on the evolution of the host genomes by analysing their contribution to genome organization and correlation with gene expression. Finally, I present a catalog of the homologs of piRNA pathway in sponges and analyse their transcriptional activity.

University of Zagreb, Faculty of Science, Department of Biology
(112 pages, 43 figures, 20 tables, 187 references, original in English)

Keywords: Transposons, genome assembly, porifera, computational genomics
Supervisor: Kristian Vlahoviček, PhD, professor
Reviewers: Helena Ćetković, PhD, Scientific advisor
Petr Svoboda, PhD, professor
Rosa Karlić, PhD, assistant professor

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Doktorska disertacija

# IDENTIFIKACIJA I KARAKTERIZACIJA POKRETNIH GENETIČKIH ELEMENATA U SPUŽVAMA (PORIFERA)

MAJA KUZMAN

Prijenosni genetički elementi (transpozoni) sačinjavaju velik dio većine eukariotskih genoma i igraju važnu ulogu u regulaciji transkripcije, oblikovanju regulatornih mreža gena i evoluciji genoma. Zbog nedostatka kvalitetno sklopljenih genoma ne-modelnih organizama, ne zna se mnogo o raznolikosti, rasprostranjenosti i ulozi transpozona u skupini Metazoa. U ovom radu predstavljam visokokvalitetne sklopove genoma za vrste spužva *Eunapius subterraneus* i *Suberites domuncula*. Korištenjem tehnika računalne genomike identificiram, karakteriziram i uspoređujem pokretne elemente u javno dostupnim genomima spužvi i interno skupljenim genomima analizom brojnosti i očuvanja transpozona u različitim vrstama spužvi. Procjenjujem utjecaj transpozona na evoluciju genoma domaćina analizom njihovog doprinosa u organizaciji genoma i povezanosti s ekspresijom gena. Na kraju predstavljam katalog homologa piRNA puta u spužvama i analiziram njihovu transkripcijsku aktivnost.

Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Biološki odsjek
(112 stranica, 43 slike, 20 tablica, 187 literaturnih navoda, jezik izvornika: engleski)

Keywords: transpozoni, sklapanje genoma, spužve, računalna genomika
Mentor: dr.sc. Kristian Vlahoviček, redoviti profesor
Ocjenitelji: dr.sc. Helena Ćetković, znanstvena savjetnica, IRB
Petr Svoboda, PhD, redoviti profesor, IMG
dr.sc. Rosa Karlić, docentica, PMF

# Prošireni sažetak

## I. Uvod

Pokretni genetički elementi (transpozoni) i sekvence izvedene iz njih čine velik dio mnogih eukariotskih genoma, a otkriveni su i u različitim prokariotskim vrstama. Generalno se mogu podijeliti u 2 glavne skupine. Elementi skupine I koriste RNA kao posrednik u svom mehanizmu kopiranja i lijepljenja, dok se elementi skupine II sastoje od DNA transpozona koji se mobiliziraju na način rezanja i lijepljenja. Obje skupine sadrže elemente koji se mogu klasificirati kao autonomni, ako kodiraju proteine potrebne za vlastitu (retro)transpoziciju ili ne-autonomni, ako za svoj prijenos koriste proteine koje proizvode autonomni transpozoni.

Pokretni genetički elementi utječu na genome svojih domaćina na različite načine zbog svoje sposobnosti kretanja i umetanja u nova mjesta u genomu. Ponekad njihov utjecaj nije odmah jasan – ovo se dogodi kada su umetnuti u nefunkcionalnu DNA. U drugim prilikama njihov je utjecaj daleko od suptilnog. Često su odgovorni za velike genomske ekspanzije i povećanje genomske raznolikosti. U ekstremnim primjerima transpozoni čine 85% genoma. Iako veliki dio njih akumulira mutacije i degradira se nakon umetanja što ih čini nesposobnim za transpoziciju, u većini genoma neke od transpozonskih obitelji još su uvijek aktivne. Postoje mnogi primjeri molekularnog pripitomljavanja, u kojima LTR retrotranspozoni sudjeluju u širokom rasponu staničnih regulatornih procesa, pa čak i evoluciji novih gena. Transpozoni također mogu stvarati alternativne transkripte, a njihovi regulatorni elementi mogu sudjelovati u cis-regulaciji gena domaćina.

Tri su glavne strategije u identifikaciji i anotaciji transpozona: anotacija na temelju postojećih baza podataka, *de novo* anotacija na genomu i *de novo* anotacija na neprocesiranim podacima. Trenutno je zlatni standard anotacija bazirana na dostupnim bazama podataka pomoću alata RepeatMasker, gdje se sekvence uspoređuju s poznatim ponavljajućim sekvencama ili motivima. Ovaj pristup neće uspješno identificirati nove, nepoznate transpozone, te se njegovo korištenje preporučuje samo za organizme za koje je dostupna sveobuhvatna baza repetitivnih elemenata. Njihova uspješna identifikacija uvelike ovisi o dostupnosti i kvaliteti složenih genoma i dostupnosti konsenzusnih sekvenci transpozona srodnih vrsta.

Spužve (Porifera) su koljeno koje se najranije odvojilo od ostatka pripadnika skupine Metazoa. U adultnom obliku su višestanične, heterotrofne i sesilne životinje. Glavna

karakteristika čitavog koljena je postojanje brojnih pora, po čemu je skupina i dobila ime (lat. *porifera* = poronoše). Žive isključivo u vodi i oslanjaju se na održavanje konstantnog protoka vode kako bi iz nje filtrirale mikroskopske čestice hrane i izmjenjivale plinove. Sadrže specijalizirane stanice koje nisu organizirane u tkiva i organe, i nespecijalizirane stanice koje se mogu transformirati u različite tipove stanica. Iako im nedostaju mnoge složene morfološke osobine pronađene kod pripadnika skupine Bilateria, njihovi transkriptomi otkrivaju veliku genetsku složenost.

Do danas su poznati samo rijetki opisi transpozona u spužvama, a budući da većina dostupnih studija opisuje nalaze specifičnih elemenata, sveobuhvatna studija svih transpozona i njihov utjecaj na evoluciju genoma spužvi još uvijek nedostaje. Potencijalni razlog nedostatka ovakve studije su nedostupnost kvalitetno složenih genoma spužvi. Iako koljeno sadrži preko 9000 različitih vrsta, do danas je složeno i javno dostupno tek pet genoma. Najbolje složen genom spužve je genom vrste *Amphimedon queenslandica,* a osim njega složeni su i genomi vrsta *Sycon ciliatum, Tethya wilhelma* i *Oscarella pearsei.* Nedavno je objavljen genom vrste *Ephydatia muelleri* složen do razine kromosoma.

Razlozi za nedostupnost kvalitetnih genoma razumljivi su ako razmotrimo proces sekvenciranja i slaganja genoma. U svojim počecima, genomi su bili sekvencirani metodama takozvane prve generacije sekvenciranja. Karakteristika očitanih sljedova dobivenih ovom tehnologijom je visoka točnost (oko 99%) i duljina (do 900 parova baza po očitanom slijedu), no iznimno niska ukupna duljina svih očitanih sljedova. Druga generacija sekvenciranja uvelike je unaprijedila ukupnu duljinu svih očitanih sljedova koja se može postići u jednom eksperimentu. Za razliku od prve generacije gdje je prosječan genom u jednom eksperimentu moguće pročitati samo djelomično, korištenjem metoda druge generacije sekvenciranja prosječan genom može odjednom biti pročitan više stotina puta opetovano. Ovako ogromno poboljšanje nažalost dolazi sa svojom cijenom - duljina pojedinačnog očitanih sljedova dobivenog u eksperimentu druge generacije sekvenciranja drastično je manja i iznosi tek oko stotinu baza.

Metode treće generacije sekvenciranja rješavaju oba problema prethodnih generacija. Duljina očitanih sljedova je ponovno velika, i to puno bolja nego prije - prosječan slijed je sada dugačak nekoliko tisuća baza, te je ukupna duljina usporediva s ukupnom duljinom dobivenom u eksperimentima druge generacije sekvenciranja. Glavni izazov koji ova generacija sekvenciranja donosi je vrlo visok postotak pogreške u očitanim sljedovima. Dok je pogreška očitanih sljedova dobivenih metodama prve dvije generacije sekvenciranja manja od 1 %, u

prosječnom slijedu dobivenom trećom generacijom sekvenciranja nije čudno vidjeti pogreške u oko 20 % baza.

Proces slaganja genoma dugačak je i netrivijalan, a uključuje pronalazak preklapanja između očitanih sljedova dobivenih prilikom sekvenciranja i temeljem njih sklapanja dugih neprekinutih nizova koji odgovaraju dijelovima početnog genoma. Razvijeni su razni algoritmi za sklapanje očitanih sljedova ovisno o tehnologiji sekvenciranja kojom su dobiveni. Dok korištenje kratkih očitanih sljedova proizvedenih eksperimentima druge generacije sekvenciranja rezultira neprekinutim nizovima nukleotida s visokom točnošću, njihova je duljina često tek neznatna u usporedbi s duljinom genoma, te je tako složen genom fragmentiran. Korištenje očitanih sljedova dobivenih eksperimentima treće generacije sekvenciranja za rezultat ima puno manji broj puno duljih neprekinutih sljedova no u tom slučaju je njihova točnost manja u usporedbi s prethodnima. Iz ovih razloga su najbolji rezultati slaganja genoma često dobiveni kombinacijom ovih metoda.

Upotrebom metoda druge i treće generacije sekvenciranja složila sam genome visoke kvalitete za vrste spužvi *E. subterraneus* i *S. domuncula*. Zbog problematičnog slaganja repetitivnih regija genoma, prvo sam složila dugačke očitane sljedove dobivene metodama sekvenciranja treće generacije te sam naknadno ispravila pogreške u dobivenim neprekinutim sljedovima korištenjem kratkih očitanih sljedova koje sam prethodno složila u iznimno točne ali kraće neprekinute sljedove. Definirala sam potencijalno nepouzdane regije u složenim genomima i uzela ih u obzir u daljnjim analizama.

U složenim genomima i svim javno dostupnim genomima spužvi identificirala sam transpozone korištenjem metoda koje se oslanjaju na postojeće konsenzusne sekvence transpozona, i posebno korištenjem metoda koje otkrivaju transpozone *de novo*. Isključila sam iz daljnje obrade transpozone koji se nalaze u potencijalno nepouzdanim regijama genoma. Ostale transpozone sam anotirala usporedbom s postojećim konsenzusima transpozona. Nakon identifikacije i karakterizacije, usporedila sam brojnost i raspodjelu transpozona među genomima spužvi. Procijenila sam utjecaj na evoluciju genoma spužvi analizom njihovog doprinosa u organizaciji genoma, analizom korelacije s ekspresijom gena i analizom homologa piRNA puta koji služi u obrani protiv transpozona. U radu sam koristila isključivo tehnike računalne genomike.

# II. Metode

## Izvori podataka

Biološki uzorci spužve *E. subterraneus* prikupljeni su u špilji Tounjčica u blizini Ogulina. Uzorci vrste *S. domuncula* prikupljeni su u Rovinju. DNA (i RNA za vrstu *E. subterraneus*) izolirane su iz kulture primorfa. DNA je sekvencirana metodama druge (Illumina HiSeqX) i treće (Oxford Nanopore Technologies) generacije sekvenciranja. Genomi ostalih spužvi preuzeti su iz javno dostupnih publikacija koje opisuju genome, a genom spužve *A. queenslandica* preuzet je iz baze Ensembl (verzija 47). Ostali korišteni podaci preuzeti su s javno dostupne baze SRA, a eksperimenti su dostupni pod pristupnim brojevima SRP055403 i ERP001229.

## Slaganje genoma spužvi *Eunapius subterraneus* i *Suberites domuncula*

Pročistila sam očitane sljedove dobivene metodama druge i treće generacije kako bih uklonila adaptere i dijelove niske kvalitete. Tako pročišćene očitane sljedove dobivene trećom generacijom sekvenciranja složila sam računalnim programom Flye koji je pogodan za slaganje nepouzdanih dugačkih sljedova, a temelji se na algoritmu koji slaže sljedove pomoću grafa ponavljanja. Ovim slaganjem dobiveni su dugi neprekinuti sljedovi čija je pouzdanost niska zbog prisutnosti pogrešaka.

Kako bi ispravila pogreške u dobivenim neprekinutim sljedovima koristila sam iznimno točne, ali kratke očitane sljedove dobivene metodama sekvenciranja druge generacije. Sljedove sam složila računalnim programom SPAdes u de Bruijnov graf. Čvorovi tog grafa predstavljaju kratke genomske sljedove duljine k (tzv. k-meri), a dva čvora su međusobno povezana bridom ako im je zajednički slijed duljine k-1. Put koji prolazi kroz više bridova u takvom grafu daje relativno dug neprekinuti slijed koji predstavlja dio genoma. Ovako dobiveni neprekinuti slijed je visoke točnosti i ne sadrži pogreške, no u pravilu su ovakvi putevi kroz de Bruijnov graf puno kraći nego neprekinuti sljedovi dobiveni slaganjem očitanih sljedova treće generacije sekvenciranja. Zbog toga sam iskoristila vrlo točne puteve u de Bruijnovom grafu kako bih ispravila pogreške u dugačkim neprekinutim sljedovima.

Ovako složenim genomima procijenila sam kvalitetu slaganja i odredila dijelove niske kvalitete usporedbom s početnim očitanim sljedovima. Usporedila sam opće karakteristike i dovršenost složenih genoma s ostalim javno dostupnim genomima spužvi korištenjem računalnih programa QUAST 5.0.2 i BUSCO 4.0.5. QUAST služi za određivanje općih značajki genoma - njegove duljine, fragmentiranosti i prosječne duljine neprekinutih sljedova. BUSCO se koristi za procjenu dovršenosti genoma na temelju broja pronađenih očuvanih ortologa gena koje u genomu bilo kojeg pripadnika skupine Metazoa očekujemo pronaći u jednoj kopiji.

## Identifikacija i karakterizacija transpozona u genomima spužvi

Identificirala sam ponavljajuće elemente u svim genomima spužvi korištenjem računalnog alata RepeatMasker koji traži sličnosti genomskih sekvenci s poznatim ponavljajućim elementima koji su pohranjeni u bazi podataka RepBase. Zbog činjenice da ne postoji baza ponavljajućih elemenata specifičnih za spužve, mnogi su ponavljajući elementi ovim pristupom ostali neotkriveni. Iz tog sam razloga koristila *de novo* pristup kako bih definirala konsenzusne sekvence ponavljajućih elemenata u svakom od genoma spužvi zasebno, i iskoristila dobivene knjižnice kao predložak prema kojem je RepeatMasker odredio potencijalne ponavljajuće elemente.

Potencijalne ponavljajuće elemente koje sam odredila na ovaj način sam pročistila s obzirom na kvalitetu složenosti genoma u regiji u kojoj su nađeni. Pročišćene ponavljajuće elemente sam anotirala usporedbom s poznatim konsenzusima transpozona dostupnim u bazama Dfam i RepBase24.11 korištenjem računalnog alata RepeatClassifier. Ponavljajuće elemente sam grupirala u skupine: jednostavni ponavljajući sljedovi, regije niske kompleksnosti, DNA transpozoni, transpozoni koji se repliciraju mehanizmom kotrljajućeg kruga (eng. *Rolling circle, RC*), retrotranspozoni koji sadrže duga terminalna ponavljanja (eng. *Long terminal repeats, LTR*) i retrotranspozoni koji su dugi i raspršeni po genomu (eng. *Long interspersed nuclear element, LINE*). Ponavljajući sljedovi koji nisu grupirani niti u jednu od definiranih skupina anotirani su kao zasebna skupina nepoznatih elemenata. Koristeći računalni alat LTR_retriever, skupinu LTR retrotranspozona sam dodatno podijelila na podskupine s obzirom na očuvanost integriranog elementa u cjelovite elemente koji sadrže internu sekvencu, usamljene cjelovite LTR elemente bez interne sekvence i ostale, neodređene elemente.

## Procjena doprinosa transpozona evoluciji genoma spužvi

Kako bih procijenila doprinos transpozona evoluciji genoma spužvi, prvo sam anotirala genome spužvi s obzirom na kodirajuće sekvence. Iz publikacija koje su opisivale genome sam preuzela anotaciju genskih regija, s označenim intronima i egzonima. Za spužve *S. ciliatum* i *O. pearsei,* kao i za spužve koje sam složila sama genski modeli nisu bili dostupni te su napravljeni pomoću računalnog alata BRAKER2. Usporedila sam brojnost i količinu gena, te zasebno introna, egzona i intergenskih regija u genomima svih spužvi.

Procijenila sam doprinos transpozona evoluciji genoma spužvi analizom doprinosa organizaciji genoma u smislu preklapanja s genima, (i zasebno sa egzonima i intronima) i intergenskim regijama. Odredila sam očuvanost svake skupine transpozona analizom stope mutacija elemenata određenih u genomu u odnosu na konsenzusni element, i usporedila je

između različitih skupina i vrsta. Zasebno sam analizirala stope mutacija LTR elemenata s obzirom na očuvanost elementa u genomu (cjelovit/ zaseban ili neodređen).

Kako bih odredila postoji li povezanost između mjesta integracije određenog tipa transpozona i ekspresije gena, analizirala sam ekspresiju gena u vrsti *E. subterraneus* prema grupama gena podijeljenim s obzirom na skupinu transpozona koji je u njih integriran, i prema regiji u koju je element integriran (egzon ili intron). Dodatno sam za LTR elemente usporedila razinu ekspresije gena s obzirom na mjesto integracije elementa (radi li se o prvom / zadnjem ili unutarnjem egzonu). Zasebno sam odredila postoji li razlika u razini ekspresije gena koji su kodirani transpozonima u odnosu na gene koji nisu kodirani transpozonima. Sve spomenute analize napravila sam i vizualizirala samostalno u računalnom programskom okruženju R, verziji 4.1.

Kako bih odredila jesu li u spužvama aktivni geni uključeni u borbu protiv transpozona, odredila sam homologe piRNA puta u spužvama. Provjerila sam njihovu ekspresiju u spužvi *E. subterraneus*, i tokom deset različitih faza embrionalnog razvoja i životnog ciklusa spužve *A. queenslandica*.

# III. Rezultati

## Genomi vrsta *E. subterraneus* i *S. domuncula* i usporedbe s ostalim genomima spužvi

Složila sam genome spužvi *E. subterraneus* i *S. domuncula* i uklonila iz njih neprekinute sljedove koji su prepoznati kao bakterijski. Složene neprekinute sljedove sam dodatno obradila pomoću visoko točnih sljedova kako bih umanjila broj pogrešaka. Završna verzija genoma *E. subterraneus* sadrži 3339 neprekinutih sljedova ukupne duljine 185.5 megabaza (Mb). Složeni genom vrste *S. domuncula* sadrži 833 neprekinuta slijeda čija je ukupna duljina 101.3 Mb.

Razina dovršenosti genoma spužve *S. domuncula* superiorna je u usporedbi s ostalim dostupnim genomima spužvi. Njegova procijenjena dovršenost temeljem analize postojanja očuvanih jedinstvenih ortologa za skupinu eukariotskih organizama je 95.3%, te 91.3% za skupinu jedinstvenih ortologa koje očekujemo naći u skupini Metazoa. Procijenjena dovršenost genoma *E. subterraneus* također je visoka - on zauzima treće mjesto od analiziranih sedam genoma. Od očekivanih jedinstvenih ortologa u skupini eukariotskih organizama, kod njega je nađeno 88.3%, a od ortologa koje očekujemo naći kao jedinstvene u skupini Metazoa nađeno je 81.7%.

Broj gena u analiziranim genomima spužvi kreće se od 18906 u genomu *O. pearsei*, do 47022 u genomu *E. subterraneus*. Geni su u prosjeku relativno kratki, s prosječnom duljinom 3277 baza i medijanom duljine 1566 baza. Od ostalih genoma posebno odskače *S. ciliatum* čiji medijan duljina gena iznosi čak 3123 baze. Ova analiza nije potpuna zbog nejednake dovršenosti analiziranih genoma - postoji korelacija između prosječne duljine gena i fragmentiranosti genoma, te je moguće da se boljim slaganjem genoma zaključak o prosječnim duljinama gena promijeni. Ovo se da naslutiti i pregledom raspodjela duljina gena u genomu *S. domuncula* koji je najbolje složen s obzirom na genski sastav, te pokazuje medijan duljine gena mjerljiv sa puno većim genomom *E. muelleri* (2673 u odnosu na 2234) iako su oba genoma složena sa usporedivo niskom fragmentiranosti.

Pregledom duljina introna, egzona i intergenskih regija zaključila sam da je ukupan postotak introna i intergenskih regija u genomu pozitivno koreliran s veličinom genoma, dok udio egzona u genomu ostaje relativno stabilan s obzirom na veličinu genoma. Genom vrste *E. muelleri* pokazuje najveći postotak intergenskih regija u genomu što je moguća posljedica korištenja različitih tehnologija sekvenciranja koje su omogućile iznimno dobru razinu složenosti tih regija, inače teških za slaganje. S obzirom na to, moguće je da će u budućnosti

bolja složenost intergenskih regija i ostalih spužvi dovesti do dodatnog povećanja u postotku intergenskih regija. Spužva *S. ciliatum* pokazuje zanimljivo velik postotak introna u genomu, i omjer broja nukleotida koji grade egzone / introne 0.26, dok je taj omjer u ostalim spužvama veći od 0.7. *S. ciliatum* je jedina spužva koja pokazuje razliku u raspodjeli duljina prvog introna u odnosu na ostale introne, što je u višim organizmima obično povezano s postojanjem regulatornih regija u prvom intronu.

## Identifikacija i karakterizacija transpozona u genomima spužvi

Odredila sam ponavljajuće elemente u genomima spužvi na dva načina. Korištenje konsenzusnih sljedova poznatih ponavljajućih elemenata kao predloška rezultiralo je u identifikaciji u prosjeku 6.28% baza kao ponavljajućih. Nakon što sam za svaku vrstu odredila konsenzuse *de novo* i koristila dobivene knjižnice u identifikaciji ponavljajućih elemenata, postotak baza identificiranih kao ponavljajuće porastao je na 36.3% u prosjeku. Ponavljajuće elemente sam pročistila prema kvaliteti složenosti genoma u regijama gdje su elementi nađeni. Pročišćene elemente sam anotirala i grupirala u skupine. Elementi koji ne pripadaju transpozonima posebno su grupirani u skupine jednostavnih ponavljanja i regija niske kompleksnosti. Elementi koji pripadaju transpozonima su grupirani u skupine DNA transpozona, RC, LINE i LTR. Ponavljajući elementi koji nisu grupirani u niti jednu skupinu objedinjeni su u jednoj skupini i anotirani kao nepoznati.

Elementi čija je skupina poznata raspodijeljeni su različito u genomima različitih spužvi. Većina ponavljajućih elemenata (čija klasa nije neodređena), čak 59.3%, vrste *O. pearsei* spada u jednostavna ponavljanja. Vrsta *S. ciliatum* pokazuje sličan trend: 42.7% elemenata poznate klase spada u jednostavna ponavljanja, a 38.8% spada u tip LINE transpozona. Transpozoni čine otprilike tri četvrtine poznatih ponavljajućih elemenata u ostalim genomima, a u vrsti *E. muelleri* čine čak 86.5%. Zanimljivo je da u vrsti *A. queenslandica* RC elementi čine 25% svih poznatih ponavljajućih elemenata, dok u vrsti *S. ciliatum* nije nađen niti jedan element ove skupine, a u vrstama *E. muelleri* i *E. subterraneus* su nađeni u iznimno niskom postotku (<1%). Generalno je broj baza koji su anotirani kao ponavljajući elementi koreliran s veličinom genoma, no ova je korelacija statistički značajna samo za skupinu LINE elemenata, čiji je ukupan postotak relativno nizak.

## Procjena doprinosa transpozona evoluciji genoma spužvi

Kako bih procijenila utjecaj transpozona na evoluciju genoma spužvi, prvo sam analizirala njihov doprinos organizaciji genoma u svim spužvama. Transpozoni su sačinjavali od 3.6% svih egzona u vrsti *O. pearsei,* 9.5% u *S. ciliatum* do čak 46% svih egzona u vrsti *E. muelleri,* dok su u ostalim spužvama skupine Demospongia sačinjavaju 15-29% baza eksona. Sličan trend vidljiv je i u intronima. Najmanji postotak introna od svih spužvi bio je građen od transpozona u vrsti *O. pearsei*, dok je najveći postotak baza introna imao preklapanje s ponavljajućim elementima u vrsti *E. muelleri* (45.2%). Ponavljajući elementi grade različit udio intergenskih regija, od 26.1% u vrsti *O. pearsei*, do čak 65.3% u *E. muelleri*.

Usporedbom s očekivanim brojem baza u svakoj od regija pokazala sam da postoji jasan trend deplecije transpozona u kodirajućim regijama u genomima svih analiziranih spužvi, kao i konzistentno veći udio transpozona u intergenskim regijama od očekivanog s pretpostavkom slučajnog rasporeda transpozona u genomu. S druge strane, u većini spužvi postoji deplecija transpozona u intronima, dok ih je u intronima *S. ciliatum* više od očekivanog. Bilo bi zanimljivo u budućnosti analizirati je li takvo obogaćenje transpozonima povezano s neobično dugačkim intronima u *S. ciliatum* u odnosu na ostale spužve. Također je zanimljivo da većina transpozona pokazuje depleciju u kodirajućim regijama a obogaćenje u intergenskim regijama, dok je za LTR transpozone u vrstama *S. ciliatum* i *S. domuncula* ovaj trend obrnut što daje naslutiti da postoji skupina gena kod kojih je došlo do iznimno uspješne ko-opcije te su preuzeli dio kodirajuće sekvence od LTR elemenata, ili se pak integracija dogodila nedavno.

Analizirala sam potencijal aktivnosti transpozona određivanjem stopa mutacija u odnosu na konsenzusnu sekvencu. Najmanju stopu mutacija pokazali su LTR elementi u *S. domuncula, E. subterraneus,* i *S. ciliatum* te RC elementi u *E. muelleri* i LINE elementi u *O. pearsei*. Činjenica da su najmanje mutirani znači da su se ovi elementi u prosjeku najmanje promijenili nakon što se prva integracija dogodila u pojedini genom, iz čega možemo zaključiti da u ovim skupinama treba tražiti potencijalno aktivne elemente. Generalno iako su nađene statistički značajne razlike u stopi mutacija za mnoge skupine transpozona ovisno o regiji u genomu u koju su integrirani, statistička značajnost većinom proizlazi iz velikog broja opservacija, a ne velike razlike. Osobno smatram zanimljivom razliku koja je pronađena u svim spužvama za LTR elemente a posebno je očita u vrsti *S. ciliatum*, za koje je stopa mutacija niža u egzonima nego u intergenskim regijama i intronima. Naime, u ostalim skupinama elemenata je ovaj trend obrnut, elementi koji su ugrađeni u egzone pokazuju veću stopu mutacija nego

elementi istih skupina ugrađeni u introne ili intergenske regije. Ovaj trend u LTR elementima posebno je očit za “solo” LTR elemente.

Generalno, ekspresija gena koji su kodirani transpozonima je niža u svim spužvama nego ekspresija gena koji nisu kodirani transpozonima. Također, geni u kojima je transpozon integriran preferencijalno u introne imaju u prosjeku višu ekspresiju nego geni u kojima je integracija transpozona preferencijalno u egzonima. Iako generalno LTR elementi slijede ovo pravilo, geni u kojima LTR elementi pune duljine sudjeluju kao 3’egzon imaju u prosjeku višu ekspresiju nego geni u kojima je LTR element pune duljine ugrađen u intron.

Na kraju, s obzirom da u mnogim organizmima postoji piRNA put koji služi za obranu protiv transpozona, istražila sam kako taj put izgleda u spužvama. Od 14 gena koji sudjeluju u piRNA putu u čovjeku, osam homologa je nađeno u gotovo svim spužvama, a ostali su nađeni kao homolozi u barem jednoj. Svi pronađeni homolozi u vrsti *E. subterraneus* bili su među 40% najviše eksprimiranih gena, a homolog gena PIWIL1 bio je u gornjih 10% po razini ekspresije. Ovo otkriće je zanimljivo jer je taj gen inače eksprimiran u zametnim stanicama. S obzirom na ovo, provjerila sam kakva je ekspresija homologa piRNA gena u različitim životnim fazama spužve *A. queenslandica* i pokazala da su homolozi gena PIWIL1 i DDX4 iznimno aktivni u većini razvojnih faza, pa i u odraslom obliku ove spužve. Ove činjenice zajedno daju naslutiti da su transpozoni u spužvama aktivni i u ostalim fazama razvoja a ne samo nekim, što će biti vrlo zanimljivo istražiti u budućnosti.

# 1 Introduction

The phylum Porifera consists of over 9200 species and there are currently only five publicly available sponge genomes (Riesgo *et al.*, 2014); (Kenny *et al.*, 2020); (Francis *et al.*, 2017); (Fortunato *et al.*, 2014); (Nichols *et al.*, 2012), and no research exists comparing the transposable elements from this diverse phylum. Only sparse descriptions of transposable elements in sponges exist currently and there are two main reasons for this. The first and most important one is the lack of high quality sponge genomes. Second reason is that our knowledge about the types of transposable elements in this diverse group is limited. As a consequence, methods for identification of transposable elements that rely on the currently known transposable elements will underestimate their diversity and abundance, while *de novo* methods might fail to detect novel transposable elements (TEs) due to poor assemblies of the available genomes.

## 1.1 Objectives

In the thesis, I will use techniques of computational genomics to assemble the genomes of two sponge species and explore transposable elements in all available sponge genomes. First, I will produce quality drafts of *de novo* genome assemblies for two sponge species, *Eunapius subterraneus* and *Suberites domuncula* whose genomes were not published previously. Since the repetitive regions of the genome are especially difficult to assemble, I will use both erroneous long reads and accurate short reads to assemble the genomes. I will identify and annotate potential transposons in produced assemblies and all publicly available sponge genome assemblies using both *de novo* and repository based methods. The transposon abundances will be compared between sponge genomes. I will characterise their impacts on genome evolution of sponges by assessing the contribution to genome organization, analysing the correlation with gene expression and analysing the presence and expression of the homologs of the piRNA pathway.

## 1.2 Genome sequencing

*As a bioinformatician, I have had the opportunity to discover amazing things about how the cells function. While most knowledge took years to accumulate and to grasp, there were a few eye opening facts that came almost in a flash. One of them was the astonishing revelation that the majority of available genomes are not nearly as finished or reliable as I naively expected them to be. In the following section I will in short describe the reasons for this.*

The human genome project was one of the greatest feats of exploration in history, with the goal to sequence and understand the genome of humans. It started in 1990 and took 13 years and 2.7 billion dollars to get to the result - a draft assembly (*The Cost of Sequencing a Human Genome*, no date; Lander *et al.*, 2001). However, it has spurred an extraordinary progress in genome sequencing technologies. In 2003, at the time when the first draft of the human genome was published, there were only few species with available genomes: 38 bacteria, one fungus (*Saccharomyces cerevisiae*), two invertebrates (*Caenorhabditis elegans* and *Drosophila melanogaster*) and one plant (*Arabidopsis thaliana*), all with relatively small and simple genomes (Lander, 2011). Today there are 18992 bacterial and only 459 Eukaryotic complete and published genomes, while 127829 bacterial and 4440 Eukaryotic genomes are in "permanent draft" status according to GOLD (Mukherjee *et al.*, 2019).

*Table 1. Characteristics of different sequencing generations.*

| | Generation of sequencing | | |
|---|---|---|---|
| Hallmark | First | Second | Third |
| Fragment lengths | - | <50kb | No limit |
| Sequence bias | GC and AT underrepresented | | None, more erroneous homopolymers |
| Error rate | <0.3% | <0.1% | ~15% |
| Error profile | mismatch | mismatch | insertions (5%), deletions (8%), mismatches (5%) |
| Data per run (Gb) | <0.002 | <1800 | 1 |
| Reads per run | 96 | Hundreds of millions | Thousands |
| Read length | <900 | <300 | 30kb |

The reason for such sparse availability and low completeness of genomes is simple - currently available technologies are not able to determine the sequence of the entire chromosomes at once. Instead, the DNA has to first be fragmented into a large number of fragments with appropriate size for sequencing. The basic steps of all DNA sequencing

experiments are the same: after DNA is extracted and purified from the cells of interest, it is sheared into smaller fragments and prepared for sequencing (the step is referred to as the template preparation or library preparation), after which the sequence is obtained and analysed (Metzker, 2010). Currently three generations of sequencing technologies exist which differ by characteristic approaches they take to solve each step. I summarize the hallmarks of each generation in Table 1. Values represent typical values for each generation of sequencing, but they vary depending on the sequencing platform used and should be taken as an approximation. For more detailed information see references (Jain *et al.*, 2015; Reuter, Spacek and Snyder, 2015; Kchouk, Gibrat and Elloumi, 2017; Amarasinghe *et al.*, 2020).

In the first and second generation of sequencing, after random shearing the DNA fragments are clonally amplified to enhance the signal needed to determine the sequence. Alternatively, in the third generation of sequencing the fragments themselves are used directly as templates with no prior amplification. In the case where amplification is needed, the fragments are selected to match a predefined size. In second generation sequencing experiments, sequences of both ends of the fragment can be determined, which is called "paired-end" sequencing. In case of fragments longer than several kilobases, the fragment has to be circularized and cut into a smaller fragment first, after which the ends of this smaller fragment are sequenced. Reads obtained this way are paired, but their relative orientation is opposite than in the original fragment. To denote this, such paired sequencing is called "mate-pair" sequencing. Both pair-end and mate-pair sequencing are useful because they result in pairs of sequences, with information about the approximate distance between them. The amplification of fragments was first done in bacterial clones, but was updated to cell-free systems where it is carried out by a polymerase (emulsion PCR (Dressman *et al.*, 2003), solid state amplification (Fedurco *et al.*, 2006) and other). The goal of amplification is to produce a population of identical templates (clones, thus this step is called clonal amplification), all of which will be sequenced. In first generation sequencing, only one template is clonally amplified and sequenced in a single reaction, and in parallel up to 96 templates could be sequenced in a single sequencing run. Reads (fragments with sequence determined) obtained by first generation sequencing have the length of up to 900 nucleotides. Both the first and the second generation of sequencing require amplification of the original DNA fragments which regularly introduces biases in the resulting reads. Both AT- and CG-rich regions of the genome are often underrepresented because they are more difficult to amplify than the rest of the genome under the same amplification conditions (Su *et al.*, 1996; Dohm *et al.*, 2008; Metzker, 2010).

Another consequence of amplification affecting second generation sequencing experiments is that the reliability of the sequence drops with increasing length of the read. In short, for each clonally expanded population of fragments, the order of nucleotides is determined by addition of fluorescently labelled nucleotides. The nucleotide complementary to the first nucleotide in the template binds to all templates of the same population, while the other nucleotides are washed away. At this stage, all the fragments from the population have the same nucleotide bound to them, and emit the same fluorescent color. They are imaged and the emitted signal is averaged to discern the identity of the base in the first position. This cycle is repeated several dozens of times, and in each new cycle, the following base is determined by averaging the fluorescent signal from the population of templates. The incorporation of bases is not perfect, e.g. some of the templates will not have a nucleotide incorporated during one cycle, but will instead incorporate this nucleotide in the next cycle, which is known as dephasing. Since the bases are determined by averaging the signal for the entire population of templates, if an error occurs in some of them, it will be passed on to next cycles, and added to previous errors, which will deteriorate the signal quality for later bases in the sequenced fragment. On the other hand, due to the fact that the signal for each base read in second generation sequencing is averaged over the entire population of reads originating from the same template, the overall estimated error rate for this type of sequencing is under 1% (Reuter, Spacek and Snyder, 2015), and the errors are mostly single base mismatches. Moreover, each base is assigned a quality score based on the purity of the measured signal, which is used later in read processing to remove low quality parts from reads. However, the drop in quality imposes restrictions on the read length, so the main characteristic of this sequencing generation is a large number of relatively short reads, with a major decrease in price per megabase compared to first generation. Second generation sequencing technologies enabled parallel sequencing of hundreds of millions (up to 5 billion) different fragments in a single experiment (Kchouk, Gibrat and Elloumi, 2017), but the size of each read has decreased to hundreds of nucleotides (typically 75 - 300). First generation sequencing is less affected by the problem of drop in quality with increasing read length, since the sequences are separated based on their lengths during the electrophoresis which precedes basecalling. In this case the lengths of the reads are capped at around 900 nucleotides because the electrophoresis does not efficiently separate large fragments that differ in lengths by a single nucleotide. The main problem with the first generation sequencing is the low output. A single experiment typically generates 0.0021 Gb of data per run, while in the second generation a single run can produce up to 1500 Gb of data (Kchouk, Gibrat and Elloumi, 2017). While short read

sequencing is cost-effective and accurate, the small size of the produced reads complicates the task of reconstructing the original molecules.

Third generation sequencing produces reads with lengths several orders of magnitude higher compared to previous generations. Third generation sequencing technologies use the DNA fragments directly as templates and by doing so, completely avoid PCR biases introduced during whole genome amplification. Two technologies dominate the field of long read sequencing: PacBio (Pacific Biosciences) single molecule real time sequencing (SMRT), and Oxford Nanopore Technologies (ONT, also referred to as "nanopore sequencing"). SMRT uses immobilized polymerase (Eid *et al.*, 2009) and detects fluorescence upon incorporation of differently labelled nucleotides to the circularized template. The fluorescence is recorded as a movie, and the template is read several times, after which a consensus is made. The lengths of the reads are limited by longevity of the polymerase, with current average around 30 kbp (Hebert *et al.*, 2018; Amarasinghe *et al.*, 2020). Nanopore technology uses biological nanopores which are only wide enough for single stranded DNA molecules to pass through. They "read" the sequence of DNA by measuring the ionic current fluctuations which occur because different nucleotides confer different resistances to the flow of ions between two sides of the nanopore (Jain *et al.*, 2016; Rang, Kloosterman and de Ridder, 2018). The signal is not straightforward to decipher, so different machine learning approaches are used to infer the base at some position from the signal. Base-calling of nanopore sequences is an area of active research and there are 23 different tools for this purpose alone (Amarasinghe *et al.*, 2020). This process has a high influence on the error rates. The manufacturers report raw nanopore reads error rates to be under 5%, while independent evaluation of the human genome-derived raw nanopore data shows 9-18% error rate (Jain *et al.*, 2018; Bowden *et al.*, 2019). The base-calling is highly dependent on the datasets used to train the machine learning algorithms, and since the currently available basecallers are trained on a mixture of human, yeast and bacterial data, the effective accuracy of nanopore reads originating from other species will most likely be lower (Amarasinghe *et al.*, 2020). Although the errors of nanopore reads are most frequently indels (around 12% of bases), substitutions are also non-neglectable (5% of bases) (Jain *et al.*, 2015). They are not uniformly distributed across the genome and occur more frequently in homopolymers.

In summary, first generation sequencing produces very accurate reads of length up to 900 bases, but the output is very low. Due to the need of whole genome amplification which is routinely done by PCR, first and second generation sequencing produce reads that do not represent the genome well - regions with very high AT or GC content are underrepresented. Second generation sequencing enabled massively parallel sequencing, resulting in hundreds of

millions of very accurate reads which cover the entire genome dozens of times. However, the short length of the reads complicates the downstream analysis. Third generation sequencing avoids the PCR bias and produces reads which have lengths few orders of magnitude greater than previous generations, but at a cost of an order of magnitude higher error rate.

## 1.3 De novo genome assembly

*As the reader might anticipate by now, the process of genome assembly is definitely not trivial. It often involves millions of 200 bases long and several thousands of kilobases-long 85% accurate reads, and the expected final result is around 20 continuous sequences (chromosomes) of a few hundred mega bases in length. A very common explanation of the genome assembly process is the analogy with a 200 million pieces puzzle. I agree with this, and would like to add that some of the pieces in the box are missing, a lot of them are extremely similar to each other, some are unfortunately not entirely correct, and there is a non-negligible percent of pieces belonging to another puzzle mixed in your box. And only if you are lucky have you seen a similar puzzle partially solved before.*

Genome assembly is the process of composing small fragments of DNA sequence into a representation of the original chromosomes from which the fragments derive. There are different algorithms used for short and long read assembly, but with a common goal - to find most probable overlaps between reads and use this information to make a longer sequence. Longest possible continuous sequences of identified nucleotides are called contigs. If the sequencing was performed in paired-end/mate-pair mode and there is information about distances between pairs of reads, this information can be used to order the contigs into larger sequences, where the exact sequence between contigs is not known, but the distance between them is determined by distance between pairs in sequencing which map to ends of contigs. Such sequence constructs are called scaffolds (see Figure 1). As an example, the human genome was first assembled and published as a high quality draft assembly, in 2001, covering ~90% of the euchromatic genome and interrupted by around 250,000 gaps (Lander *et al.*, 2001). Currently assembled version of the human genome (Genome Reference Consortium Human Build 38 patch release 13, GRCh38.13) contains 3,099,706,404 bases assembled into 998 contigs, ordered into 472 scaffolds. Gaps still include mostly repetitive regions that could not be reliably cloned or assembled, heterochromatic sequences, including the large centromeres and the short arms of acrocentric chromosomes (Lander, 2011). In August of 2019, a first complete telomere-

to-telomere assembly of human chromosome X was published assembled *de novo* from data sets which covered the chromosome over 150 times (Miga *et al.*, 2019).

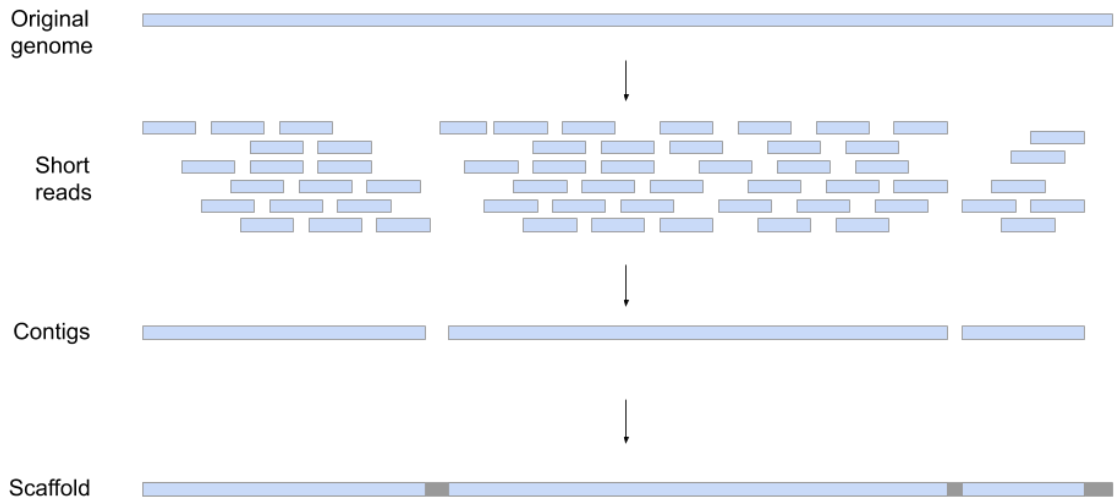



*Figure 1. Schema of the general steps in genome assembly. The original genome is fragmented and sequenced as short reads. In the assembly process the overlapping reads are assembled into contigs. If there is an estimate of the distance between the contigs, they are connected with gaps into scaffolds.*

### 1.3.1 De novo genome assembly algorithms

Today there are several different approaches in genome assembly algorithms; de Bruijn graph (DBG) and string graph approaches are often used in second generation sequence assembly, while the third generation seemingly benefits from overlap-layout-consensus and repeat graphs approaches. In the remainder of this chapter I will briefly explain general ideas behind them and outline some of the problems associated with them.

The overlap-layout-consensus solves the assembly problem by constructing a graph where the reads are represented by nodes in the graph. In the overlap phase, pairwise alignments for nodes are calculated and each two nodes are connected if they overlap with predetermined length and identity. Edges hold the information on the number of shared nucleotides between reads. In the layout phase, the graph is simplified - redundant edges are removed based on transitivity and all paths that visit each node only once are found. Finally, the most likely paths through the graphs are found in the consensus step. This approach requires a computationally intensive step of overlapping reads both in the first and last step, so it becomes inefficient when using a large number of reads. For those reasons, DBG (Pevzner, Tang and Waterman, 2001;

Butler *et al.*, 2008; Chaisson and Pevzner, 2008; Zerbino and Birney, 2008; Bankevich *et al.*, 2012) or string graph (Myers, 2005; Simpson and Durbin, 2010) - a faster variant of the OLC algorithm, are mostly used in second generation sequence assembly where the number of reads is counted in hundreds of millions.

In the de Bruijn graph approach, every read is divided into successive overlapping sub-reads of length k, called k-mers. All k-mers become nodes in a graph, in such a way that each k-mer is represented only once. Nodes are connected in the graph whenever they are successive k-mers in some read. This leads to a graph where nodes are connected in a way that the last k-1 bases in one node match the first k-1 bases of another node. The assembly problem is solved by solving a problem of finding the path through the graph, in such a way that each connecting edge is traversed only once. Although some of the information is lost while constructing a graph because the reads of size n are divided into k-mers of sizes usually around n/2, high coverage of sequencing guarantees that most k-mers are represented in multiple reads, which makes the assembly possible. An important consideration is the choice of the k-mer size (Chen *et al.*, 2017). A small k-mer will decrease the number of edges in a graph, but will produce many vertices which complicate the path reconstruction and lead to more fragmented assembly. If the k-mer choice is too small, it might introduce chimeras in the final contigs. A larger k is also preferred with repetitive regions in mind, where a k-mer longer than a repeat will alleviate the problem of a complicated graph (Chikhi and Medvedev, 2014). However, larger k-mer requires higher overlap between reads so the number of vertices will drop and we might lose some connections we would have with lower k-mer value.

### 1.3.2 Problems accompanying a de novo assembly of short reads

The first problem that the de Bruijn graph approach encounters is handling errors. It relies on read correctness, and does not handle high error rates efficiently. Erroneous reads create low coverage k-mers which are added as additional nodes to the graph, and branches connecting to them might create ambiguous paths and result in a fragmented assembly. The problem becomes more pronounced taking into account the presence of polymorphisms. This is why even short reads with their <1% error rate are corrected prior to graph construction (Kelley, Schatz and Salzberg, 2010; Chikhi and Medvedev, 2014; Chen *et al.*, 2017). Due to non-uniform coverage of sequencing caused by PCR bias, some of the paths might be underrepresented and proclaimed errors. Two types of regions are most difficult to correct in second generation sequencing reads: regions with low coverage of second generation reads (both GC-rich and AT-rich regions), and regions in direct proximity of highly repetitive patterns

(Heydari *et al.*, 2017) . Another common issue is the presence of contamination (e.g. reads originating from different species) in the sequences. Here, the repeated regions might be similar in some of the species present in the sample which will produce more branches in the graph and will lead to fragmented assembly due to ambiguity. All of those potential problems can be alleviated by higher sequencing coverage.

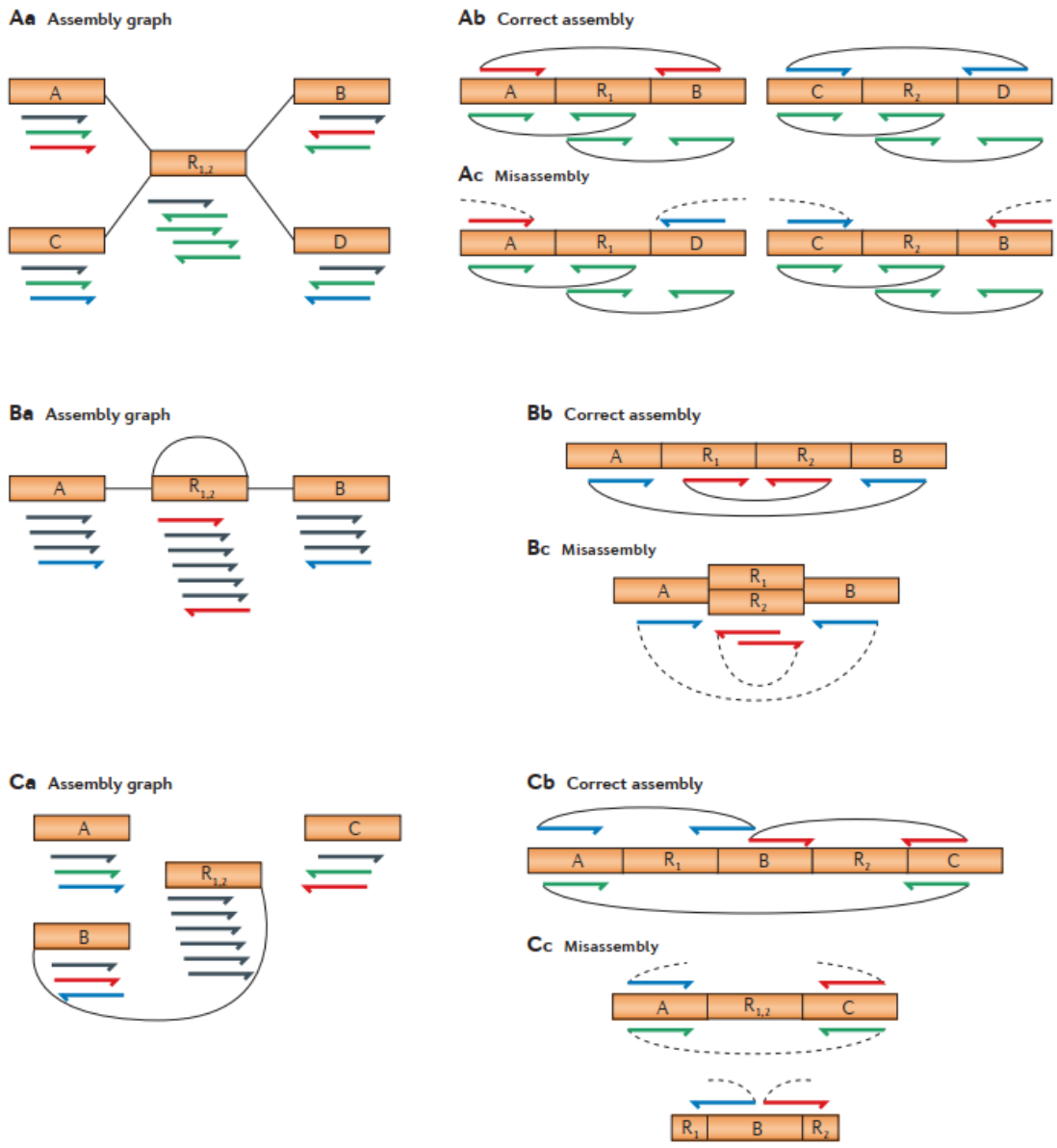



*Figure 2. Types of errors found in de bruijn assembly graphs derived from short reads. Adapted from (Treangen and Salzberg, 2011).a-assembly graph, b- correct assembly, c- misassembly, A-rearrangement error caused by repeat, B- A collapsed tandem repeat, C- collapsed interspersed repeat, R1,R2 - repeat*

The biggest technical challenge that accompanies a short read *de novo* assembly is the assembly of the repetitive regions of the genome, (Treangen and Salzberg, 2011). Whenever a repeat is longer than read length, it will not be assembled into a single contig. Instead, it will

create shorter contigs which will correspond to the sequence flanking the repeat and start of the repeat. Another contig will harbour the end of the repeat and sequence flanking it. Depending on the size of the repeat, there will be many short contigs representing the sequences found in different repeat copies throughout the genome. In worst cases, if the assembler we chose was less conservative, it will produce erroneously collapsed repeats and misassembled rearrangements (Phillippy, Schatz and Pop, 2008; Pop and Salzberg, 2008). These problems can be partially resolved by the use of paired-end and mate-pair distance information, but are ultimately solvable only by read lengths longer than lengths of the repeats, so third generation sequencing provides a great opportunity to finally solve the genomes.

### 1.3.3 Long read assembly

Third generation sequencing reads have a much higher error rate, and they are several orders of magnitude longer than the second generation reads, so the assembly approaches needed to be modified accordingly. Since the number of long reads per experiment is generally smaller than the number of short reads, overlap-layout-consensus methods are usually used in the assembly step, although modified de Bruijn graphs (ABruijn graphs and repeat graphs) are shown to work well in assembly of error-prone long reads (Lin *et al.*, no date; Kolmogorov *et al.*, 2019). In k-mer based approaches, errors are avoided in assembly by avoiding overlaps which rely on rare k-mers, although this might result in failure to detect some correct overlaps. More commonly, assembly approaches rely on overlapping and making a consensus of long reads.

The process of long read assembly given high quality data is very fast and cheap. Recently, an assembler was published which assembled 11 human genomes *de novo* with more than 99.9% identity in under 6 hours per genome on a single processing node (Shafin *et al.*, 2020). The “trick” was in the data - they achieved 63x nanopore coverage for each genome of which 6.5× coverage was in reads >100 k. Such a good data set is often not available, and the resulting accuracy depends on an unreliable process of finding overlaps between erroneous reads.

Because of the high error rate, nanopore reads are usually corrected either before further analysis, during the assembly process, or the final assembly is corrected (polished). The correction is either done by using only long reads (non-hybrid methods), or by using accurate short reads (hybrid methods). In case when the errors are corrected by non-hybrid methods, corrected reads will still have errors which exhibit a bias for indels in homopolymeric regions due to the fact that the errors are not uniformly distributed (Wenger *et al.*, 2019). Another

approach is to use highly accurate short reads to correct nanopore reads. Hybrid error correction methods either map the short reads directly to long reads and correct them, or first assemble them or build a de Bruijn graph and use the graph or assembled contigs to correct the long reads. Assembly based methods usually outperform alignment based methods because there are less ambiguities in mapping longer reads then with using shorter ones. For comprehensive evaluation of error correction methods for long reads, see (Zhang, Jain and Aluru, no date; Fu, Wang and Au, 2019; Lima *et al.*, 2019).

In practice, hybrid error correction works better with assembly than non-hybrid methods alone (Mahmoud *et al.*, 2019), but the regions of the genome which are highly similar to each other (e.g. transposons, pseudogenes, non-coding regions) might not be corrected because of their low mappability (Kolmogorov *et al.*, 2019). This will happen since it is very difficult to discern true few-percent true differences in sequence from nanopore errors which average around 15%. Also, regions which are underrepresented in the original short reads set might be corrected less efficiently (Mahmoud *et al.*, 2019).

After the assembly, it is common to "polish" the genome in order to improve the base accuracy of scaffolds, or in other words to correct any errors which were not successfully corrected either before or during the assembly process. Since the polishing step is basically error correction, approaches used in error correction are used now as well. Some tools recommend iterative polishing with the rationale that more reads will be uniquely mappable on more accurate assembly. However, too many iterations of polishing sometimes reduce the quality of the assembly (Miller *et al.*, 2018).

In summary, different sequencing generations provide different advantages - short reads ensure base level accuracy, while long reads improve the scaffolding into chromosomes. The problem remains in assembly of repetitive regions, which when assembled with short reads, can be collapsed, cause misassembles and often are the major culprit for fragmented assembly. While second generation only assembly is mostly fragmented due to simple repeats, assemblies based only on long reads can generally produce highly complete genomes. However, this depends on the coverage of long reads and genome complexity, since third-generation based assembly gaps originate primarily from LTR elements and satellites (Peona *et al.*, no date). Using a combination of both approaches will often improve the assembly but the repetitive regions will again be corrected less efficiently due to their low complexity and problematic mappability of short reads to them. As a conclusion, there is no silver bullet for a *de novo* genome assembly and it will most likely require a lot of data, time, patience and careful manual examination.

### 1.3.4 Assembly quality assessment

*I once read that the genome assembly process is like a teenager: It has bold claims on what it is capable of doing, but makes embarrassing mistakes while actually doing it. There are a plethora of available tools for de novo genome assembly and most of them will produce output when given some input. Deciding on which one is the right for you will require careful examination of the produced assembly in the light of your expectations and needs for downstream analysis.*

There are many assemblers available and due to the different algorithms used, they will produce different assemblies starting from the same data. If the referent genome is available, the produced assemblies can be compared with it and we can calculate how close they are to the theoretical limit on completeness and correctness given a data set (Mikheenko *et al.*, 2018). However, the existence of a good referent genome is a rare luxury, and other metrics are devised to summarize the quality of the assembly. Total length of the assembly is the sum of all lengths of scaffolds produced in the assembly. (When referring to assembled genomes, I will use the terms scaffolds for all assembled sequences, both true scaffolds and contigs, regardless of the existence of gaps in them.) Since our goal is to reconstruct the original genome, the total length should be as close as possible to the length of the genome, and the number of scaffolds ideally corresponds to the original haploid number of DNA molecules per cell in the sample. In reality, the number of scaffolds is orders of magnitude higher than expected and varies greatly between assemblies, which is often referred to as contiguity. Instead of comparing the total number of scaffolds, a more stable measure, N50 is commonly used. It is defined as the length of the scaffold which will in summation with all scaffolds longer than itself surpass the half of the total assembly length. It is similar to median scaffold size, but gives more weight to longer scaffolds. More generally, Nx is defined as the scaffold length which in summation with all longer scaffolds surpasses x% of the total assembly length, while Lx is defined as the smallest number of scaffolds needed to achieve Nx. Related measures NGx and LGx are analogously defined with respect to the total genome length. Unfortunately, the degree of contiguity varies among different genomes and assemblies, and is not well correlated with genome correctness (Salzberg *et al.*, 2012). Other measures solve this problem by measuring the annotation completeness based on the number of genes that can be predicted in an assembly (Brůna, Lomsadze and Borodovsky, no date; Hoff *et al.*, 2019). Since the number of genes is variable among different species, it is common to estimate completeness based on a smaller set

of genes we expect to find (e.g. universal single copy orthologs) (Parra, Bradnam and Korf, 2007; Simão *et al.*, 2015; Seppey, Manni and Zdobnov, 2019). Different tools exist that map the short reads back to the assembly to score the whole assembly (Clark *et al.*, 2013; Rahman and Pachter, 2013; Yang *et al.*, 2019), while some score each position of the assembled sequence and automatically detect misassembly (Hunt *et al.*, 2013; Muggli *et al.*, 2015; Wu *et al.*, 2017). If other data is available, (e.g. transcripts) it is informative to map them to the assembled genome as well. Since repetitive regions are usually more problematic to assemble than gene rich regions, there are metrics that evaluate the assembly quality by assessing the quality of assembly of a group of interspersed repeats or piRNA clusters, and by analysing the abundance, SNPs and internal deletions of transposable elements (Ou, Chen and Jiang, 2018; Wierzbicki *et al.*, 2020). As a conclusion, there is no straightforward way to determine the quality of an assembly and different metrics should be used keeping in mind the goal of the research.

## 1.4 Transposable elements

Transposable elements (TEs) are parts of DNA which can change position within a genome. Transposable elements and sequences derived from them constitute a large portion of most eukaryotic genomes (Kazazian, 2004) and have also been detected in various prokaryotes. They are broadly divided into 2 main classes (Finnegan, 1989). Class I elements use a copy and paste mechanism of mobilization by an RNA intermediate, while Class II consists of DNA transposons which mobilize in a cut and paste fashion. Both classes contain elements that can be classified as autonomous or non-autonomous, either encoding proteins necessary for their (retro)transpositions, or utilizing proteins produced by elements of autonomous class. With the increase in the number of species with assembled genomes grew the diversity of discovered transposable elements. It has become difficult to place newly discovered transposable elements into the existing 2 class system. Class I is further divided into four subclasses: long terminal repeat flanked elements (LTRs), non-LTR elements (which contain autonomous long interspersed elements (LINEs) and non-autonomous short interspersed elements (SINEs)), direct inverted repeats (DIRS), and Penelope-like elements (PLEs). Class II elements are grouped into three subclasses: terminal inverted repeat bearing elements (TIRs), Helitrons, and Mavericks (Polintons) (Wicker *et al.*, 2007). Each subclass is further divided into subgroups (superfamilies) of monophyletic origin (e.g. Ty3/gypsy, Ty1/copia, Tc1/mariner and other). Finally, elements are grouped into subfamilies and families which all represent the remnants of

a common single ancestral transposable element (Britten and Kohne, 1968), and can be represented by the same consensus sequence.

Transposable elements impact the genomes of their hosts in various ways due to their ability to move through the genome and insert into new positions. They are often responsible for major genomic expansions and increase in genomic diversity (Morgulis *et al.*, 2006; Piskurek and Jackson, 2012). In the most extreme examples transposable elements constitute 85% of the genome (Schnable *et al.*, 2009). Although large fraction of them accumulate mutations and truncation events after insertion, rendering them incompetent for transposition, in most genomes some families are still active and can generate new insertions. Even in the inactive form, they are a source of regulatory elements for the host genome through which they drive genome evolution. The potential impact of transposable elements to the host genome is largely determined by the insertion location in the genome. In eukaryotes, transposable elements, much like retroviruses, often exhibit a preference for sites of integration and can target active genes (Sultana *et al.*, 2017; Lucic *et al.*, 2019) which can result in host adaptation or pathogenicity. In humans, genes derived from retroelements are involved in many important biological processes, including pluripotency (Lu *et al.*, 2014), placenta formation, X chromosome inactivation (Lyon, 2006), the immune system (Young, 2016) and cancer (Wang *et al.*, 2007; Young, 2016; Rodriguez-Martin *et al.*, 2020). Transposable elements can also create alternative transcripts, while their regulatory elements can be co-opted for cis-regulation in host regulatory pathways (Chuong, Elde and Feschotte, 2017). I will briefly outline the characteristics of each class which make them so widespread and successful.

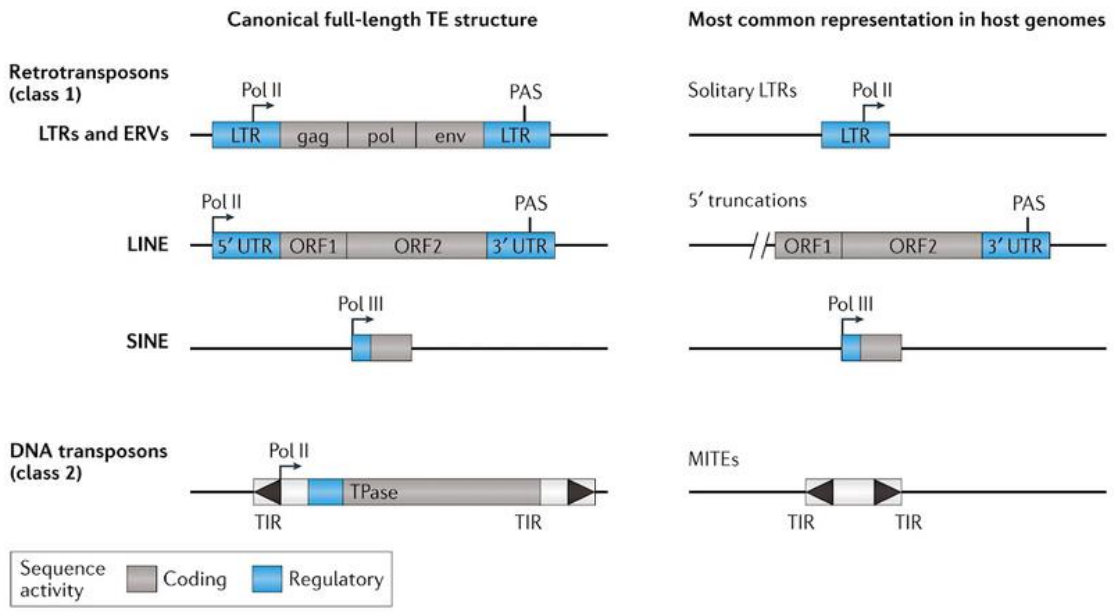



*Figure 3. Schematic of major transposable element (TE) classes and their typical genetic organization. Adapted from (Chuong, Elde and Feschotte, 2017)*

## 1.4.1 Class I elements

### 1.4.1.1 Non-LTR retrotransposons

Class I non-LTR retrotransposons are still active in many mammals today. It is estimated that there are 80-100 active L1 elements in the human genome (Brouha *et al.*, 2003) and that a new L1 insertion occurs in the germ cell of 1 in 50 individuals. (Ostertag *et al.*, 2002). They shape the mammalian genome in many ways, reviewed in (Kazazian, 2004). They cause mutations by insertions and structural rearrangements by homologous recombination. However, they are constructive in a number of ways. They can repair double stranded breaks by acting as a bandage to the broken DNA and inserting in the break. They can move 3' or 5' proximal sequences to a new genomic location. These 3'/5' transductions are possible because of a weak RNA cleavage signal at the 3'end of an L1 element, and the existence of a promoter upstream of a full length 5' end of an L1 element. L1 reverse transcriptase occasionally switches a template from L1 RNA to other small nuclear RNA. This results in production of new chimeric retrogenes. L1 elements are also present as exons in nearly 200 protein coding sequences (Li *et al.*, 2001), and alter gene expression by providing a promoter which results in an alternative transcription start site (Nigumann *et al.*, 2002). Although they preferentially act in cis (Wei *et al.*, 2001), occasionally they are able to trans mobilize nonautonomous SINEs. Alu elements

are 300z bp long SINEs that do not encode any proteins and yet have expanded to 1.1 million copies, comprising 11% of the human genome (Ostertag and Kazazian, 2001). Alu elements are also associated with upregulated gene expression, and have the largest overall regulatory impact on the human genome (Zeng *et al.*, 2018).

#### 1.4.1.2 LTR retrotransposons

Class I LTR retrotransposons, termed endogenous retroviruses (ERVs) are integral part of most eukaryotic genomes. They comprise about 8% of the human and 10% of the mouse genome, but are more frequently exapted compared to other transposable elements (Chuong *et al.*, 2013; Xie, Donohue and Birchler, 2013; Kannan *et al.*, 2015). In their full form, they typically consist of two identical 5' and 3' LTRs, surrounding ORFs that encode proteins essential for their replication, including gag and pol and an ORF encoding an envelope protein (env) that is usually degraded (see Figure 3). The key to why they are particularly suitable for co-option can be explained by their structure. Their LTRs harbour a range of regulatory regions, including different combinations of transcription factor binding sites, PolII promoter elements (TATA box), transcription start site and may also contain a splice donor (Thompson, Macfarlan and Lorincz, 2016). Although in many cases soon after the integration into the host genome the internal region of the element is deleted by recombination of LTRs, those "solo" LTRs still carry preserved regulatory elements, unlike LINEs which are mostly truncated. In the human genome, we find 577 000 of such "solo" elements (Friedli and Trono, 2015), and it is estimated that they have contributed to around 20% of functional binding sites for many transcription factors in human and mouse, including OCT4, SOX2 and NANOG (Sundaram *et al.*, 2014). Some classes of LTRs hold a preserved splice donor site which renders them suitable for exaptation as alternative promoters (Peaston *et al.*, 2004; Lamprecht *et al.*, 2010). Whole genome studies have found them to be overlapping with up to 80% of human lncRNAs, mostly serving as exons, but occasionally also co-opted as promoters (Kannan *et al.*, 2015). Solo LTRs in intergenic regions are potential sources of novel lncRNA and protein coding genes (Friedli and Trono, 2015; Franke *et al.*, 2017). They have also been co-opted as enhancers, which is not surprising in light of the potential to bind trans-acting factors due to the presence of multiple transcription factor binding sites.

### 1.4.2 Class II elements

Class II of transposable elements are present across the eukaryotic tree of life (Feschotte and Pritham, 2007). This class consists of DNA transposons which use the transposase enzyme to move through the genomes in a cut and paste fashion. DNA transposons are flanked by terminal inverted repeats (TIRs, see Figure 3). Transposase recognises TIRs and excises the transposon which is then inserted into the new location. While inserting into the new location, few nucleotides within the integration site are duplicated (target site duplications, TSD) and are characteristic for each DNA transposon. Class II elements can be divided into 2 subgroups based on their TSD, TIRs and sequence: Tc1/*mariner*, PIF/*Harbinger*, hAT, Mutator, Merlin, Transib, P, *piggyBac* and CACTA belong to subclass I, while *Helitron* and *Mavericks* replicate in a different fashion and are classified as subclass II. For a detailed review of class II elements, see (Feschotte and Pritham, 2007; Muñoz-López and García-Pérez, 2010).

Tc1/mariner superfamily is the most widely distributed family of transposable elements, present from rotifers, through insects to mammals (Robertson, 1993; Plasterk, Izsvák and Ivics, 1999; Arkhipova and Meselson, 2005), however, only a few of them are known to be active. The transposases of this superfamily do not show a lot of sequence conservation on nucleotide levels, but all have two characteristic domains: amino-terminal helix-turn-helix motive and carboxy-terminal catalytic motif (D-[92aa]-D-[31-39aa]-D/D-[92aa]-D-[31-39aa]-E). They do not require host factors to transpose, so it is not surprising that there are many proposed cases of horizontal transfer among different species within the same and even different phyla (Lampe *et al.*, 2003; Casse *et al.*, 2006; Laha *et al.*, 2007). Proposed potential vectors include parasites (Houck *et al.*, 1991) and viruses (Houck *et al.*, 1991; Hartl, Lohe and Lozovskaya, 1997). PiggyBac transposons are also present in various taxa (Sarkar *et al.*, 2003), although probably inactive. They contain a single 1.8kb ORF encoding a transposase, and 13bp long TIRs. hAT (*hobo*/*Ac*/*Tam3*) superfamily are found in eukaryotes, and encode a transposase flanked by 5-7bp TIRs. The transposase contains DDE motif and a DNA binding domain.

Transposases from DNA transposons have been used widely to manipulate an organism's genome. Compared to viral delivery systems, they are inexpensive, non-immunogenic and easy to purify (Ivics *et al.*, 2009), and some can be excised without altering the original genome (Yusa *et al.*, 2009). They efficiently transpose transgenes up to 10 kb, offer a variety of integration site preferences and are in general a promising tool for a variety of genomic studies (Muñoz-López and García-Pérez, 2010).

### 1.4.3 Balance between expression and repression of transposable elements

The rate at which TEs transpose is an important driver of genome evolution. Transposition of TEs is an imprecise process which results in large scale deletions, duplications and inversions, so they are often associated with large chromosome rearrangements. Moreover, structural variation is a direct consequence of recombination events that occur between highly homologous dispersed genomic sequences which remain even after the TE loses the capacity to mobilize (Bourque *et al.*, 2018). From an evolutionary perspective through the viewpoint of a transposable element, an ideal scenario is to be expressed in the germline, and not the somatic cells (Haig, 2016). Overexpression of transposable elements in the somatic cells might lead to an overall fitness disadvantage for the host, which would in turn reduce the probability that the transposable element will be fixed in the population. On the other hand, transposable elements which are expressed in the germline would be propagated to the next generation, and the deleterious ones will have been selected against (Calvi and Gelbart, 1994; Kano *et al.*, 2009).

Hosts have developed multiple mechanisms to restrict transposable elements expression (Goodier, 2016; Liu *et al.*, 2018), which include small RNA, chromatin and DNA modification pathways (Molaro and Malik, 2016). There is also evidence that transposons themselves are down-regulated by their own transposase - since two functional molecules of transposase are necessary to perform transposition, inactive transposase originating from mutated site acts as negative inhibitor (Lohe and Hartl, 1996; Lohe, De Aguiar and Hartl, 1997; Claeys Bouuaert *et al.*, 2013).

Small RNA pathways are an important defence against viruses and transposable elements. They include several processes that utilize short RNAs to target and manipulate complementary nucleic acids. There are three distinct small RNA pathways which act against viruses and transposable elements (Obbard et al., 2009): viRNA pathway which as a defence against virus sequences, miRNA pathway acts in posttranscriptional control of gene expression (Bushati and Cohen, 2007), and the piRNA pathway. The piRNA pathway operates in germ cells and targets transposable elements (Hartig, Tomari and Forstemann, 2007; Watanabe *et al.*, 2015; Tóth *et al.*, 2016; Meseure and Alsibai, 2020).

### 1.4.4 Identification and annotation of transposable elements

There are three main strategies in identification and annotation of transposable elements: repository based annotation, *de novo* annotation from the assembled genome, and *de novo* annotation from the raw reads (Goerner-Potvin and Bourque, 2018). Most widely used strategy is the repository based annotation using RepeatMasker, where sequences are queried against a database of consensuses of known transposable element sequences or motifs. This approach requires the availability of a comprehensive repository of transposable elements and relies on its completeness. The number of species for which there exists a high quality database of consensus sequences is limited (Lerat, Rizzon and Biémont, 2003; Hubley *et al.*, 2016), as it requires years of manual curation. RepeatMasker by default uses the Repbase Update repository. It contains more than 38,000 consensus repeat sequences, 90% of which are originating from 134 species (Bao, Kojima and Kohany, 2015). The remaining 10% encompass around 700 additional species. This is by no means a complete set of TEs and the use of RepeatMasker often fails in identification of novel, previously unknown transposable elements. In particular, when the first complete published sponge genome (that of *Amphimedon queenslandica*) is masked with RepeatMasker, only 3.51% of the genome is masked (*Amphimedon queenslandica Annotation Report*, no date), while 23.66% of it is predicted to be repetitive using a *de novo* approach (Morgulis *et al.*, 2006). This is the reason why it is recommended to use RepeatMasker only for organisms for which a comprehensive repeat library is available.

*De novo* approaches offer the solution to this problem. Tools belonging to this category can be used on genomes for which there is no repeat library, and they use consensus seeds, pairwise similarities or oligonucleotide counts to detect potential transposable elements (Bao, 2002; Price, Jones and Pevzner, 2005). Although they can be used to annotate additional new repeats not detected by RepeatMasker, some of the results might be false positives. For example, when applied to the human genome, nearly two thirds of the genome is annotated as repetitive (Koning *et al.*, 2011). The last group of tools uses low-coverage sequencing data and detects overrepresented sequences directly on raw reads. They rely on the high abundance of transposable elements over other sequences, so their abundance can be detected even in reads of low coverages (Novák, Neumann and Macas, 2010; Goubert *et al.*, 2015; Chu, Nielsen and Wu, 2016). Although such tools represent a great opportunity to detect transposable elements in species which lack a high quality assembly, they are less sensitive than aforementioned

approaches and might miss elements of low abundances (Goerner-Potvin and Bourque, 2018). In summary, computational pipelines which combine multiple approaches seem to give optimal results (Ou *et al.*, 2019; Flynn *et al.*, 2020).

## 1.5 Porifera

### 1.5.1 General characteristics of sponges

Sponges (Porifera) are the first lineage to have branched off from other Metazoa about 750 million years ago (Wörheide *et al.*, 2012; Telford, Moroz and Halanych, 2016; Feuda *et al.*, 2017). The most obvious characteristic of the animals within this phylum is the existence of numerous pores and channels throughout their bodies (*porifera* = pore-bearing). They are multicellular, heterotrophic and sessile animals in the adult form. They live exclusively in water and rely on the maintenance of the water flow to exchange gasses and to feed by filtering microscopic food particles from the water through the pores. They harbour specialized cells which are not organized into tissues or organs, and unspecialized cells which can transform into different cell types. Porifera are supported as a monophyletic group by morphological evidence that includes biphasic life cycle, filter-feeding sessile lifestyle, and the existence of pinacocytes, choanocytes and an aquiferous system (Ax, 2012; Hooper and van Soest, 2012). The monophyly of sponges as a phylum is also supported by molecular evidence (Wörheide *et al.*, 2012).

The phylum Porifera is divided into 4 classes, Demospongia, Calcarea, Hexactinellida, and Homoscleromorhpa. Demosponges are the most diverse and well studied group, comprising about 85% of all described species and are characterized by silica spicules, collagen-derived skeletal structures and the presence of spongin (Hooper and van Soest, 2012). Sponges from the class Calcarea are distinguished by extracellular calcite spicules while those from Hexactinellida produce siliceous skeletons and have syncytial tissue. Homoscleromorpha is a small class with less than 100 known species, which are all marineencrusting or lobate animals with a smooth surface, usually occurring at shallow depths. All sponges are marine except for the family Spongillidae, a group of Demospongiae which made its transition to freshwater around 250-300 million years ago (Schuster *et al.*, 2018).

Sponges reproduce sexually, releasing sperm cells into the water and fertilizing ova. They can reproduce asexually by regeneration from fragments if the fragments include right cell types. Some species can reproduce by budding. Many freshwater species and some marine species produce internal buds of unspecialized dormant cells called gemmules. They are often

produced in inhospitable environmental conditions and can form completel new organisms when the conditions improve. Although the cultivation of sponges is challenging in laboratory conditions, some species form multicellular aggregates after dissociation, called primmorphs (Custodio *et al.*, 1998; Le Pennec *et al.*, 2003). These primmorphs show organization in the structure of the aggregates and can be cultured for months.

There are 4 main phases of the sponge life cycle. It starts when the ova are fertilized by sperm cells and an embryo develops into a swimming larva. The swimming larva is called a pre-competent larvae until it gains competence for settlement. Post settlement, the larva develops into a juvenile and subsequently into an adult sponge with a fully functional aquiferous system. The embryonic development is also divided into distinct phases. The embryos in the early cleavage stages are found mostly on the edges of the brood chamber and are recognisable by milky white color. Through a series of asymmetric and asynchronous cell divisions, a solid blastula with cells of uniform sizes is formed. Different cell types in blastula organize into layers in a process that is considered to be gastrulation (Leys and Degnan, 2005). At the end of gastrulation embryos have a brown beige colour and show anterior-posterior asymmetry. Pigment cells initially distributed throughout the outer layer migrate to the posterior pole forming a spot, and later migrate outwards forming a narrow pigment ring (Adamska *et al.*, 2007).

Sponges lack many complex morphological traits found in bilateria, but their transcriptomes reveal a large genetic complexity (Harcet *et al.*, 2010; Riesgo *et al.*, 2014). Although the phylum Porifera is composed of over 9200 species (Van Soest *et al.*, 2012) , there are only 5 publicly available draft genomes. Three belong to the class Demospongiae: *Amphimedon queenslandica* (Riesgo *et al.*, 2014), *Ephydatia muelleri* (Kenny *et al.*, 2020) and *Tethya wilhelma* (Francis *et al.*, 2017), one to the class Calcarea; *Sycon ciliatum* (Fortunato *et al.*, 2014), and one to the class Homoscleromorpha - *Oscarella pearsei* (Nichols *et al.*, 2012). There are also two draft hologenomes available - that of *Stylissa carteri* and *Xestospongia testudinaria* (Ryu *et al*., 2016). The average genome size in the sponge phylum is 200Mb, but the genome size varies 17-fold, ranging from 40Mb in *Tethya actinia* to 600Mb in *Mycale laevis* (Jeffery, Jardine and Gregory, 2013).

### 1.5.2 Transposable elements in the phylum Porifera

Descriptions of transposable elements in sponges are sparse. Long terminal repeats-retrotransposon Baikalium-1 was discovered in species endemic to Lake Baikal (*Lubomirskia baicalensis*, *S. baicalensis* and *B. bacilifera*), as well as cosmopolitan species *Spongilla sp*. and

*Ephydatia sp.* collected in rivers with direct contact with Lake Baikal, but not elsewhere. Baikalium-1 is found in proximity to the silicatein-A1 gene, along with several other mobile genetic elements. The authors hypothesize that adaptation to freshwater habitat could be conferred by high transpositional activity of the ancestral silicatein gene (Wiens *et al.*, 2009). A different study identified LTR retrotransposon belonging to BEL/Pao subclass in the genome of *Amphimedon queenslandica*, in only 24 copies. Around 80 % of the potential LTR elements found in this study failed to be classified (de la Chaux and Wagner, 2011). Another study found that one third of the genes which are marked to be potentially horizontally transferred to the *A. queenslandica* genome were enriched in TE-derived sequences. Over a half of those sequences were unclassified, a quarter belonged to copia LTR retrotransposons, and the remainder was assigned as DNA transposon Helitrons (Higgie, no date; de la Chaux and Wagner, 2011). Lastly, there is evidence of miniature transposable elements (MITEs) in the *A. queenslandica* genome. The authors report 3800 and 1700 copies of Queen1 and Queen2 elements, respectively. They are mostly located in intergenic regions as well as introns, providing potential to become splice donors and acceptors (Erpenbeck *et al.*, 2011). Since most of the available studies describe findings of specific elements, a comprehensive study of all transposable elements and their impact on the evolution of the genome of sponges is still missing.

Currently there exists only one analysis exploring Piwi expression in sponges in the light of stem cells. Piwi homologs were identified in a freshwater sponge Ephydatia fluviatilis. They were expressed in cells with archeocyte and choanocyte cell morphological features. Archeocytes are pluripotent cells, and choanocytes are food-entrapping cells with the ability to transform into archeocytes under specific circumstances and to give rise to gametes (mostly sperm) indicating that they maintain pluripotent stem cell-like potential. (Funayama et al. 2010).

# 2 Materials and methods

## 2.1 Samples and sequencing

Biological samples for *Eunapius subterraneus* were collected from Tounjčica Cave near Ogulin, Croatia. Samples for *Suberites domuncula* were collected in Rovinj, Croatia. Samples were stored in ethanol. Part of the tissue samples was dissociated and cultured into aggregates of cells called primmorphs. DNA from primmorphs was isolated using Genomic Tip 100 (Qiagen) protocol and stored in DNase-free water at -20°C until sequencing by second generation sequencing technology. DNA used for nanopore sequencing was extracted from tissue samples and isolated using the QIAGEN Blood & Cell Culture DNA set.

Libraries from primmorph samples were prepared by random fragmentation followed by 5' and 3' adapter ligation using TrueSeq Nano DNA (350) kit. Adapter-ligated fragments were PCR amplified and gel purified. Each fragment was amplified into a clonal cluster by bridge amplification and was sequenced on HiSeqX sequencer in a paired-end fashion. Table 2 lists raw sequencing statistics.

*Table 2. Raw sequencing statistics for HiSeqX Illumina sequencing experiments. *Sequencing coverage was calculated with the assumed genome length of 200Mbp .*

| Sample ID | Sample | Total read bases (bp) | Total reads | Sequencing coverage | GC (%) | Q20 (%) |
|---|---|---|---|---|---|---|
| Esu001gPrim 57 | *Eunapius subterraneus* | 57,604,703,100 | 381,488,100 | 288 | 44.91 | 91.12 |
| Sdo001gPrim 65 | *Suberites domuncula* | 65,179,675,782 | 431,653,482 | 326 | 41.2 | 93.05 |

Libraries for nanopore sequencing were prepared from primmorphs and tissue with different sequencing kits, summarized in table 3. Basecalling was done using the Oxford Nanopore Technologies' basecalling algorithm with data processing toolkit Guppy.

*Table 3. Libraries from the nanopore sequencing experiments*

| Sample | Year | Number of runs | Sample origin | Library Kit |
|---|---|---|---|---|
| *Eunapius subterraneus* | 2015 | 1 | Tissue | SQK-MAP006 |
| *Eunapius subterraneus* | 2016 | 5 | Tissue | SQK-MAP006 |
| *Eunapius subterraneus* | 2017 | 17 | Primmorphs | rapid sequencing SQK-RAD004 |
| *Eunapius subterraneus* | 2018 | 28 | Primmorphs | SQK-LSK108 |
| *Eunapius subterraneus* | 2019 | 7 | Primmorphs | |
| *Suberites domuncula* | 2017 | 12 | Primmorphs | rapid sequencing SQK-RAD004 |
| *Suberites domuncula* | 2018 | 6 | Primmorphs | SQK-LSK108 |
| *Suberites domuncula* | 2019 | 1 | Primmorphs | |

RNA samples for *E. subterraneus* were isolated from primmorphs on day 1 and 10 of their growth. They were extracted with RNeasy Mini Kit (Qiagen).

## 2.2 Publicly available data

I used the publicly available sponge draft genomes, provided as supplements to the papers which described them. Genome for *Ephydatia muelleri* was downloaded from EphyBase (Kenny *et al.*, 2020), *Sycon ciliatum* from (Fortunato *et al.*, 2014), *Oscarella pearsei* from (Nichols *et al.*, 2012), *Tethya wilhelma* from (Francis *et al.*, 2017) and finally, *Amphimedon queenslandica (Riesgo et al., 2014)* was downloaded from Ensembl release 47.

The RNAseq data for the different stages of development of *A. queenslandica* was downloaded from SRA archive SRP055403.

I downloaded the assembly of the human genome from a nanopore-only data set (Canu 1.7 + WTDBG + Nanopolish) from https://genomeinformatics.github.io/na12878update/ . I downloaded the matching Illumina data set from the SRA project ERP001229.

## 2.3 Genome assembly

The schema for data preprocessing and assembly is shown on figure 4. In short, raw reads were preprocessed to improve the quality of the data sets. Nanopore reads were assembled separately into scaffolds. They were polished using a combination of high quality short Illumina reads and an assembly graph derived from high quality Illumina reads.

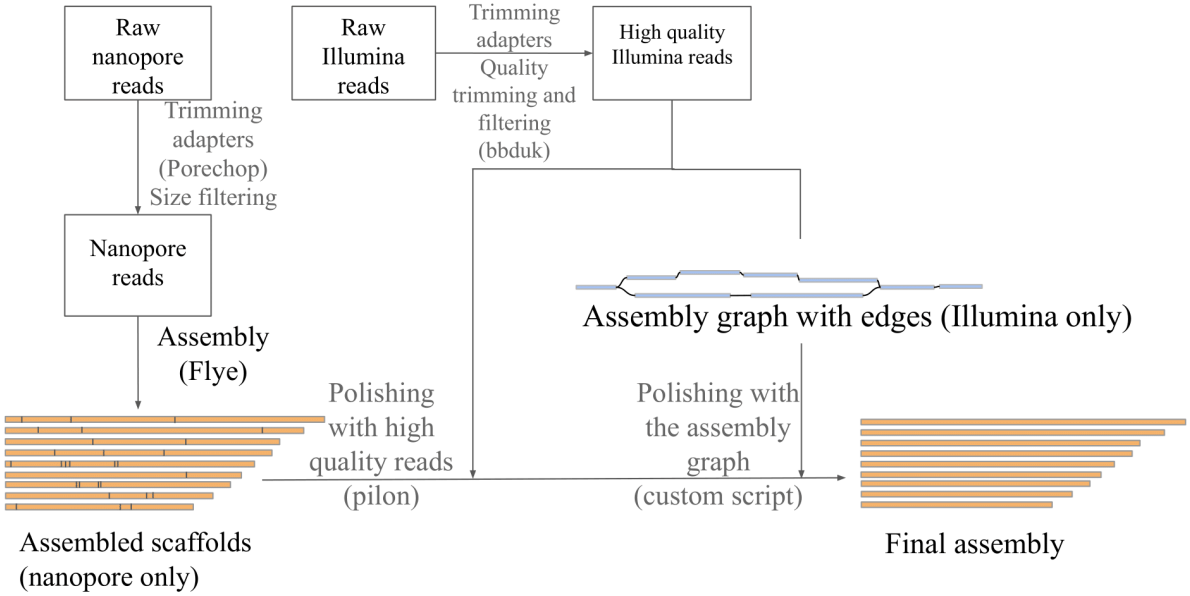



*Figure 4. Schematic diagram showing the process of data preprocessing and assembly.*

### 2.3.1 Preprocessing of the raw sequencing data

I preprocessed the raw Illumina reads as follows: firstly, I used bbduk, a tool included in (*BBMap*, no date) to remove adapters from sequences. Next, with the same tool I trimmed the bases with a quality score below 25 from both ends. Finally, I filtered out all the sequences which had an average quality of the remaining bases lower than 30 or were shorter than 50 nucleotides.

I used Porechop (rrwick, no date) to detect and remove adapters from the raw nanopore data called by Guppy basecaller, and split chimeric reads which have adapters in the middle of the read. I calculated the coverage using the approximate genome size of 200Mbp. For *S. domuncula* assembly I used the reads which were at least 5kbp long, while for *E. subterraneus* I kept all reads above 200bp, since more stringent filtering would greatly reduce the coverage.

### 2.3.2 Nanopore-only assembly

After the removal of adapters, quality trimming and filtering of the raw Illumina and nanopore reads, the genome was assembled in the following way. First, nanopore reads were assembled separately using Flye assembler (Kolmogorov *et al.*, 2019). Flye creates assembly graphs called repeat graphs, and solves them using information from reads. Relatively high coverage of the *S. domuncula* nanopore reads allowed for setting the minimum overlap parameter to 5000 bases. This setting controls the minimum overlap between reads and the increase of its value results in more contiguous assembly. The parameters used for this data set were set to expect the genome size of 200Mb, expect for minimum overlap between reads of 5kb and to use k-mer of size 17 to find overlaps between reads. Smaller k-mer value results in more sensitivity. Parameters used for the *E. subterraneus data set* were the same with the exception of minimum overlap between reads, which was set to 2000 because the coverage of the reads was a lot lower than in the *S. domuncula* dataset.

Additionally, I tested the influence of the pre-correction of nanopore reads with high quality Illumina sequences on the quality of the final assembly of *E. subterraneu*s. I used LoRDEC v0.9 (Salmela and Rivals, 2014) for correction of nanopore reads. LoRDEC corrects the nanopore reads by building a de Bruijn graph representing the short reads, and seeks a corrective sequence for each erroneous region in the long reads by traversing chosen paths in the graph. Paths in the graph are chosen based on k-mers in alignments of short reads which map to the assembly unambiguously. Not all k-mers are used - only those which have the abundance in the short reads between some predefined upper and lower limit. The lower limit is controlled by the parameter s, and the upper by parameter a. The correction is done iteratively in 3 rounds with high quality Illumina reads longer than 128 bases. In the first round a small k-mer is used which alleviates the problem of mapping the short reads to erroneous assembly (because a read mapping depends on the exact match in k-mers and lower k-mer will result in more matches). In the consecutive rounds, as the assembly is already corrected, it is common to use increased k-mer values and the value for low boundary of k-mer abundance. By increasing them, more reads are mapped to the genome unambiguously and will be used in the correction. The parameters I used were: -k 13 -s 5 -a 100000000 in the first round of correction, -k 21 -s 4 -a 100000000 in the second round and -k 31 -s 3 -a 100000000 in the final round. Next, I assembled the corrected nanopores in the same way as the non-corrected ones.

### 2.3.3 Assembly polishing

All assemblies were polished with both high quality Illumina reads (Walker *et al.*, 2014) and paths created from Illumina-only assembly graph. The polishing of the produced nanopore only assemblies was done in the following combinations:

1. Using only short reads in one iteration ("Reads only"),
2. Using only paths constructed from the assembly graph ("Paths only"),
3. Using short reads in the first round and the paths from the assembly graph constructed on the read-polished assembly in the second round ("Reads->Paths"),
4. Using the paths from the assembly graph in the first round and short reads in the second round ("Paths->Reads"),
5. Using short reads in two rounds consecutively ("Reads->Reads").

Polished assemblies were evaluated using the BUSCO 4.0.5 annotation completeness score estimated based on presence of universal single copy orthologs (Simão *et al.*, 2015). The BUSCO orthologs are defined as sets of genes from OrthoDB database of orthologs (www.orthodb.org) present in a single copy in more than 90% of the species. In this way a set of orthologs was defined for eukaryotes (eukaryota_odb10) and for metazoans (metazoa_odb10). I chose the best assembly for each species based on BUSCO score and used only this one in further analysis. The procedure is explained in more detail below.

#### 2.3.3.1 Polishing with high quality Illumina reads

I used pilon (Walker *et al.*, 2014) to polish the scaffolds with high quality Illumina reads alone. Pilon is a software that automatically improves draft assemblies by comparing the assembly with short reads which map to it. It identifies inconsistencies (including mismatches, indels and local misassemblies) between the genome and the evidence in the reads and attempts to correct them in the genome. Since pilon corrects the assembly based on the reads which map to it, it is critical to choose only the alignments of the reads which map unambiguously to the genome. To accomplish this, I mapped the reads and filtered them manually in the following way: reads were first mapped to the nanopore-only assembly using bwa-0.7.17 (Li and Durbin, 2009) with the default bwa mem parameters. The alignment was filtered with sambamba 0.6.1 (Tarasov *et al.*, 2015) to include only high quality alignments (all secondary and supplementary alignments, alignments with alignment quality 0, and all alignments which did not have a properly mapped paired read were removed). This was achieved by using the parameter: -F

“mapping_quality >= 0 and not (unmapped or secondary_alignment) and not ([XA] != null or [SA] != null) and proper_pair and not (mate_is_unmapped)”.

2.3.3.2 Polishing with paths from Illumina SPAdes derived assembly graph

The main problem with nanopore-only based assembly is that the error rate is relatively high compared to Illumina based assembly. On the other hand, the main problem of Illumina-only based assembly is that there are often too many possible paths in the assembly graph that connect the edges so there is no unambiguous path through them, which results in a fragmented assembly. Assembly graph is a representation of the genome assembly. Edges in the assembly graph represent contigs before repeat resolution. In other words, each edge represents a small part of the genome and exists in the assembly graph only once. However, the reverse is not true - an edge can appear in the genome multiple times, but in a different context. Thus, in an assembly graph identical repeats will be collapsed into a single edge but can be reconstituted if the sequence surrounding them is known. Illumina-only assembly commonly fails in reconstituting the genomic regions surrounding the repeats since the information it has is limited. The sequences are too short and do not hold enough information about the context of the repeats to unambiguously assemble those regions, and this is only partially compensated by the pair distance information. This is why the contigs and scaffolds produced are often short and the assembly fragmented.

My strategy was to use the nanopore-only assembled scaffolds as a guide which will enable us to find edges which are most similar to the assembly and search for paths between only those edges. This will reduce the number of edges in the graph and significantly reduce the number of possible paths through the graph. In turn, it will be possible to connect more edges unambiguously and this will result in a more contiguous assembly. Next, I will use those sequences (which I simply call paths) to correct the nanopore-only scaffolds and reduce the overall error rate for the assembly.

First, high quality Illumina reads were assembled using SPAdes genome assembler v3.14.0-dev (Bankevich *et al.*, 2012) into an assembly graph containing connected edges. As explained in the introduction, a choice of k-mer size used in the assembly will have an effect on the assembly graph. If a too short initial k-mer is used, it is possible that some false connections will be present in the graph. The final k-mer size used is the which ultimately determines how the graph will look like - a large k-mer requires higher overlap between reads so the number of vertices will drop and we might lose some connections we would have with

lower k-mer value. However, since k-mer longer than repeat will alleviate the problem of a complicated graph (Chikhi and Medvedev, 2014), I chose to increase the last k-mer size to 127. This means that two edges in a graph will be connected if they overlap by at least 127 nucleotides in reads which are 150 nucleotides long. Although this is a very stringent condition, the high coverage of reads allowed for it. The expected k-mer coverage can be computed as $C_k = C \cdot (R - K + 1)/R$, where C is the read coverage, R is the length of reads and K is the length of k-mer. Since the calculated read coverage for the genomes is around 150, the read length used is 150, and the last k-mer length used is 127, the expected coverage of k-mers on each edge is 26. SPAdes has an option to calculate the distribution of k-mer coverages over all the edges and use it to determine untrusted edges as those which have unusually low k-mer coverage on them. Thus, I used the SPAdes v3.14.0-dev (Bankevich *et al.*, 2012) with the parameters -k 33,55,77,99,111,127, and also used the --cov-cutoff auto option.

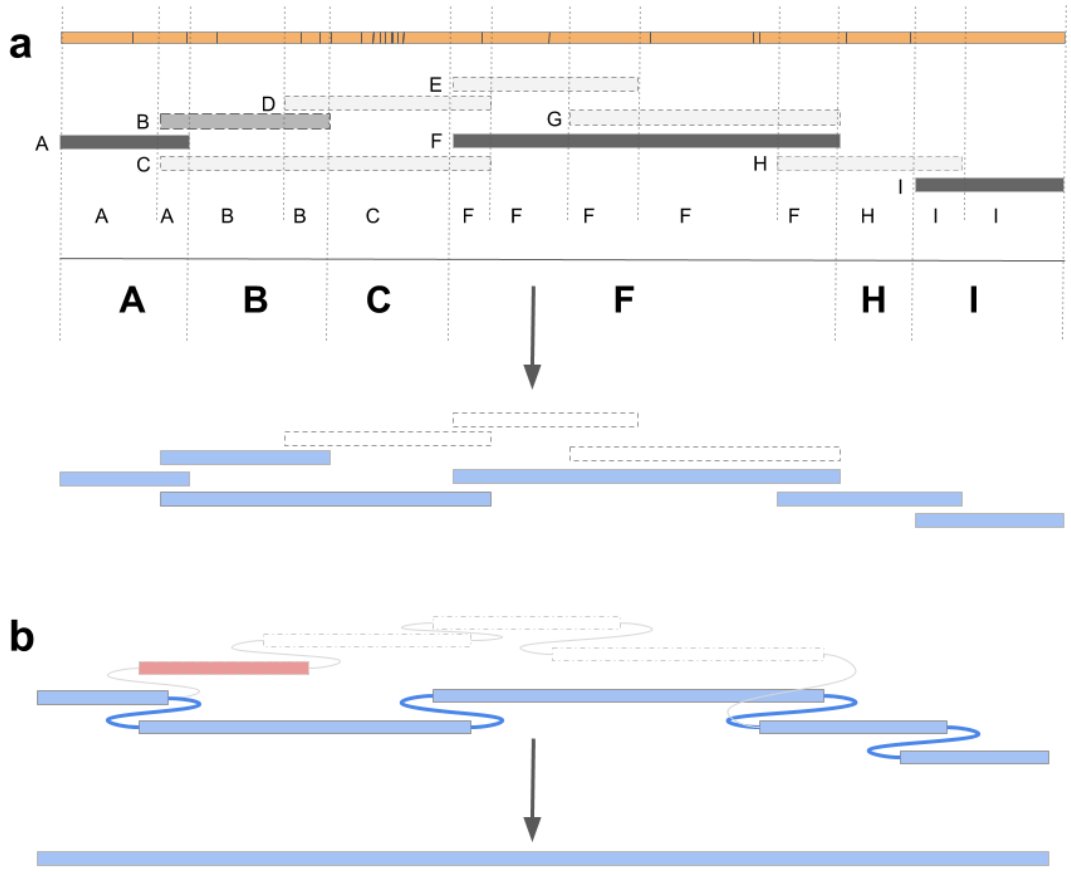



*Figure 5. Schema of the path correction algorithm. Orange rectangle represents nanopore derived scaffold. Blue rectangles represent edges in Illumina-derived assembly graph. a) Potential paths are constructed from the best mapped edges. Darker shade represents higher identity. b) Best mapping edges are queried against the assembly graph and unambiguous paths traversing best edges are returned as sequences.*

Next I needed to find edges in the produced SPAdes graph which are most similar to the nanopore-only assembly. To do this, for each region in each nanopore scaffold I found the sequence of edges in the Illumina graph which is most likely to represent this region of the nanopore scaffold. For example, in figure 5a, the nanopore scaffold is best represented by sequence of edges A:B:C:F:H:I. Since an assembly graph with those edges already exists, if there exists a path between then in the graph - they can be assembled into a longer sequence. Next, I analysed the connections between the edges which map to nanopore scaffold in the Illumina based SPAdes graph. If there exists an unambiguous path through the assembly graph which connects the adjacently mapped edges together, I used that sequence to refine the part of the nanopore assembly which was used as a scaffold (Figure 5b).

In more detail, to find candidate edges which map the best to nanopore scaffolds I first mapped the assembled (nanopore-based) scaffolds to the (Illumina-based) assembled edges with pblat (Wang and Kong, 2019). As the expected error rate of the nanopore-only assembly is relatively high, (higher than 10 percent given the low coverage of nanopores in the *E. subterraneus* data set) the process of finding edges in the Illumina based assembled graph which correspond to the nanopore scaffolds is not trivial. This is the reason why I relaxed the parameters of the blat search. The parameters I used were -maxIntron=100 -minIdentity=85 -extendThroughN, and they allowed the edges to be mapped to the assembled scaffolds even with up to 15% error rate. They also allowed mapping over unknown bases which might occur in the nanopore-only assembly, and mappingover indels which I expect to find in the assembled scaffolds. However, the problem is that by relaxing the parameters, I got a lot of edges mapped to the same position of the nanopore-only assembly. Thus, I filtered the results to retain only the alignments of edges which map the best to every position in the nanopore scaffold.

The filtering was done in the following way: I excluded alignments of all edges which are potentially assembled from contamination or sequencing errors and are unlikely to represent true sequences from the genome. Those edges can be discriminated from other edges by the coverage of k-mers which were used in the construction of the graph. All edges with the average k-mer coverage of 2 or less were removed from further analysis because this coverage cutoff was determined by SPAdes assembler to discriminate between trusted and non trusted edges.

Next, I filtered the potential set of best mapping edges by the alignment length. I removed all the edges which align with more than 1% offset in length to the reference (e.g. have more than 1% insertions or deletions). I also removed all alignments where an edge was mapped with less than 90% of its length, unless it was aligned to the edge of the scaffold - in this case it was retained.

The edges which were retained after the filtering steps were further processed as shown on the Figure 5a in order to find the best mapped edge for every position in the nanopore based scaffolds. I analysed all overlapping disjoins between all edges mapping to scaffold. In this context, a disjoin is part of the scaffold which has different alignments mapping to it when compared to preceding and following bases on the scaffold. In the figure, 9 edges map to a scaffold and define 13 disjoints. Only disjoins which are longer than some threshold (2 is default) are taken to consideration. For each disjoin I choose the edge which aligns to the scaffold with the highest percent identity to represent this disjoin.

Next, I merged the same consecutive edges into unique ones (A,A,B,B -> A,B). The sequence of edges defined this way represents potential edges which will be used to correct the scaffold. However, as shown in the example, this algorithm will not necessarily choose the best overall edges for every position in the scaffold since there might exist locally better aligned edges. This is why the sequence of edges is queried against a connected Illumina-based SPAdes derived graph. The search for unambiguous paths through the SPAdes graph was done by A. Prjibelski, a member of the SPAdes development team in CAB, SPbU. The result of this step is a set of paths through the graph which unambiguously connect the edges, respecting the order and distance between the edges. Thus, many false positive local best aligned edges are filtered out. Also, some edges are added if they are a part of an unambiguous path that is returned. This partially alleviates the mappability problem that is common in repetitive regions, and enables the recovery of edges which were unjustly filtered out in the previous steps. After the construction of the paths, I corrected the nanopore assembly by mapping the paths back to the assembly, choosing the correct alignments, and replacing the nanopore-based assembled sequence with the sequence from the paths that aligns to the assembly.

I mapped the paths to the assembly using the standard parameters for mapping nanopore reads, and not allowing for secondary alignments. This was done with minimap2 with the parameters: map-ont -c -p0.9 --secondary=no. Since I know which scaffold produced what path in the graph, I know the exact position in the assembly where I would expect the path to map back to. This is why after the mapping, all alignments which map to some position other than expected are filtered out before the correction. All the code for the determination of potential edges for the paths and the manual correction of the assembly with the paths can be found here: https://github.com/MaKuzman/SpongesTransposons/blob/master/.

I tested the accuracy of the produced paths on a small part of the human genome. To test the accuracy of the produced paths, I used the publicly available nanopore-only assembly, assembled with canu to produce paths. I selected 10 scaffolds ranging in ranks from 300th -

309, when ordered by length. Those scaffolds were arbitrarily chosen based on their size to cover around 1% of the genome. I mapped them to the human genome hg38 assembly with minimap2 (Li, 2018) using the standard parameters for mapping different assemblies one to another ( -x asm20 --secondary=no). I extracted the regions of the human genome where the nanopore selected scaffolds map using a custom R script. I processed the Illumina reads as described for the sponge data sets. I mapped them to the part of the human genome with minimap2 with standard parameters for mapping short reads and not allowing for any reads which are mapped somewhere else better (-x sr -c --secondary=no -Y) . I extracted only the reads that map to the part of the human genome with a custom bash script. I assembled those reads with SPAdes and created the paths from them as explained above. I mapped the paths to the original human genome with the standard parameters for mapping nanopore reads to the assembly (-x map-ont -c -p0.9 --secondary=no) and used a custom R script to calculate the accuracy for each path aligned to the genome based on its length. The nanopore-only canu assembly part was polished with paths as explained above. Quality of the nanopore only and the polished assembly was determined by Quast LG (Mikheenko *et al.*, 2018) using the human genome part as the reference, with the parameter --large for genome larger than 300Mb. The results of the accuracy calculations and quast results are added to the appendix of the thesis as Figure 31 and Table 14.

### 2.3.4 Assessment of quality

To filter out bacterial scaffolds I did the following: I used QUAST 5.0.2. to calculate and plot the GC content distribution in all available genomes in order to detect possible bacterial contamination on the genome level. I used MEGAN-LR (Huson *et al.*, 2018) to detect bacterial scaffolds in the assemblies of *E. subterraneus* and *S. domuncula*. MEGAN uses a modification of the lowest common ancestor algorithm for binning scaffolds onto the nodes of a given taxonomy based on alignments of known proteins to scaffolds. I downloaded the RefSeq database containing non-redundant protein sequences, updated 15.08.2019. from the NCBI. I used diamond 0.9.22 (Buchfink, Xie and Huson, 2015) program for protein alignment to align proteins from the RefSeq database to the assembled scaffolds. I increased the sensitivity of the search by modifying the parameters of diamond to blastx --sensitive --evalue 0.05 -F 15. I used MEGAN (Huson *et al.*, 2018) to assign taxonomy for the scaffolds and filter all the bacterial scaffolds from the assemblies of *E. subterraneus* and *S. domuncula*. Paula Štancl, a master student from our group, identified the bacterial scaffolds in other publicly available genomes. I removed all the bacterial scaffolds from the genomes.

I defined low quality regions for the assemblies of *E. subterraneus* and *S. domuncula* by mapping raw nanopore and high quality Illumina reads back to the genome. I defined a region to be low quality for nanopore data set if there were no nanopore reads aligned with more than 80% identity and 95% length of the read to this region, after removing first and last 5 nucleotides from each alignment. Similarly, I defined a genome region to be low quality for Illumina data set if it was not covered by any short read with at least 95% identity and 98% length of the read, after removing first and last 5 nucleotides from each alignment.

Nanopore reads were mapped with minimap2 (Li, 2018), using the standard parameters for mapping nanopore reads (-ax map-ont -Y -p0.95). I used sambamba 0.6.1 (Tarasov *et al.*, 2015) to filter the alignments: -F "not (unmapped or secondary_alignment) and not (chimeric or failed_quality_control)". I used paftools to convert sam to paf format. I used the minimap2 predefined default parameters for short reads for the mapping of Illumina reads. Sambamba was further used to filter the alignments with the parameter -F "not (unmapped or secondary_alignment) and not (mate_is_unmapped or failed_quality_control) and proper_pair". Again, I used paftools to convert sam to paf format. WIth a custom R script, I defined blacklisted regions of the genome, separately for nanopore and illumina alignments in the following way: The removal of the first and last nucleotides enables detection of chimeric regions in the genome, where reads align but are clipped in the same genomic position. I used low quality regions defined this way to filter out potential low quality transposons.

I compared the general characteristics of the produced assemblies of *E. subterraneus* and *S. domuncula* with the other publicly available sponge genomes, after the removal of bacterial scaffolds using QUAST 5.0.2 (Mikheenko *et al.*, 2018). Annotation completeness was assessed using BUSCO 4.0.5 (Seppey, Manni and Zdobnov, 2019).

### 2.3.5 *De novo* transcriptome assembly and annotation

I trimmed low quality ends from RNAseq data and filtered out all sequences which were shorter than 20 nucleotides after trimming or had an average quality lower than 20. I did this with bbduk using the parameters qtrim=rl trimq=3 minlength=20 minavgquality=20. I assembled the transcriptome *de novo* using rnaSPAdes (Bushmanova *et al.*, 2019) with default parameters. I annotated the transcripts using TransDecoder-v5.5.0 (see supplementary for (Haas *et al.*, 2013)) in the following way: First, I predicted the proteins from transcripts using the tool TransDecoder.LongOrfs. Next, I conducted a blastp search with extracted predicted proteins using diamond against a non redundant protein RefSeq database (version 15.08.2019.) with the parameters --max-target-seqs 1 --evalue 1e-5. I used the tool TransDecoder.Predict to predict

the likely coding regions from the transcripts with the parameter --retain_blastp_hits. The final coding regions predicted this way include both those regions that have sequence characteristics consistent with coding regions in addition to those that have demonstrated blast homology to known proteins.

## 2.4 Identification of potential transposable elements

I identified the potential transposable elements in the all the genomes using different approaches:

- Identification using RepeatMasker (*RepeatMasker Home Page*, no date), alone using the available RepBase24.11 repeat library
- Generating repeat library *de novo* from the assembled genomes with RepeatModeler2 (Flynn *et al.*, 2020) and combining it with RepeatMasker
- Additionally, for *S. domuncula* and *E. subterraneus* I also identified potential transposable elements from a subset of high quality short read data directly, using oligonucleotide (k-mer) content.
- Finally, for *E. subterraneus* I also identified the transcribed transposable elements from the de novo assembled annotated transcriptome.

### 2.4.1 Identification by comparison with available database

First, I used RepeatMasker (*RepeatMasker Home Page*, no date), version open-4.0.9 to search for repetitive regions in the genome. This version by default uses Dfam 3.0. database (Hubley *et al.*, 2016) to find repetitive sequences in the assembled genomes. Dfam is a database of families of repetitive DNA elements, in which each family is represented by a multiple sequence alignment and a profile hidden Markov model (HMM). Since Dfam has only a limited number of consensus sequences, I used a database which has a much larger number of consensus transposon sequences - RepBase. RepeatMasker searches for similarities of the sequences in the genome with the families of repetitive elements and reports their locations. Since I am working with eukaryotic genomes I disabled the search for bacterial insertion sequences using the parameter -no_is. I also provided information about the species using the parameter -species porifera.

### 2.4.2 Identification by comparison with de novo produced repeat libraries

Since I expect many transposable elements to be underrepresented in the available RepBase library because there is no curated publicly available library of repetitive elements in sponges, I used RepeatModeler2 (Flynn *et al.*, 2020) tool to detect potential transposable elements *de novo* from the assembled genomes. RepeatModeler2 is a computational pipeline that uses algorithms for repeat detection, RepeatScout (Price, Jones and Pevzner, 2005) and RECON (Bao, 2002), followed by consensus building and classification steps. RepeatScout identifies high frequency k-mers (repeat seeds), aligns the seeds and extends the alignment around them. This approach enables a quick detection of the youngest elements which did not diverge in sequences a lot. RECON performs sensitive but computationally intensive exhaustive intergenome alignments which enables the detection of older families. RepeatModeler2 uses both, in an iterative approach. It first subsamples a small portion of the genome (40Mb), and identifies the youngest elements. After identification, it masks the entire genome with the sequences it found, samples additional genome sequences of increasing total sizes (ranging from 3Mb to 3^5 Mb) and repeats the identification/masking process over 5 rounds. In this way, it produces a library of consensus repeat sequences de novo from the assembled genome.

The identification of LTR elements can be further improved by taking into account the structure of the LTR elements. This is why RepeatModeler2 also takes advantage of the LTRharvest (Ellinghaus, Kurtz and Willhoeft, 2008) and LTR_retriever (Ou and Jiang, 2018) tools which use information about LTR structure to create an improved library of LTR elements. Those structural information include length of LTRs, distance between them, similarity between the LTR pairs, existence of target site duplications and the existence of motif (5′-TG…CA-3′) within the LTRs to identify full length LTRs. In addition to the identification of the LTR elements which harbour the canonical TG/CA motif, LTR_retriever also identifies the non-canonical LTR-RTs (non-TGCA) with high sensitivity (91%), specificity (97%), accuracy (96%), and precision (90%), tested in rice (Ou and Jiang, 2018). It removes false positives by excluding tandem repeats, low complexity regions, sequences without identical target site duplications and sequences which harbour non-LTR transposases and protein coding sequences. It further excludes LTR elements which have other LTR elements nested inside their internal region (nested insertions). Finally, identified LTR elements are separated into LTR regions and internal regions for clustering by BLASTclust. Sequences are clustered together by the "80-90-100" rule; all sequences longer than 100 nucleotides are taken into consideration and clustered if they overlap by at least 80% sequence identity at the DNA level covering at

least 90% of the longest sequence. This step serves to remove redundancy and constructs the final LTR library. This library is then merged with the one created with RepeatModeler2.

In this way, RepeatModeler2 and LTR_retrivever are both used to construct a non-redundant repeat library in each of the sponge genomes. Those repeat libraries are then used as input to RepeatMasker (*RepeatMasker Home Page*, no date) to identify repetitive regions in the assembled genomes.

Since repetitive regions are difficult to assemble, there might exist some transposable elements which will not be detected in the assembled genomes. I used REPdenovo (Chu, Nielsen and Wu, 2016) to find potential transposable sequences directly from high quality short read data for *E. subterraneus* and *S. domuncula*. REPdenovo takes advantage of the fact that repetitive sequences appear many times in the reads and thus, only needs low coverage of reads - around 10x. To achieve this, I subsetted the high quality short read data set with reformat.sh (BBtools) using the parameter samplerate=0.1. REPdenovo finds highly occuring k-mers and extracts and assembles reads with those k-mers. I used blat with the parameter --minIdentity=80 to compare the consensuses built this way with consensuses built from the assembled genome. If a consensus overlapped with more than 80% of length and 80% identity with other consensuses, I counted it as a match. The results are added to the appendix.

## 2.5 Characterization of the transposable elements

### 2.5.1 Annotation of the repeat consensus library

Finally, repeatClassifier from the RepeatModeler2 package is used to annotate the consensuses based on similarity to known repeat proteins (Dfam3.0) and known repeat sequences (RepBase24.1). I used the annotated repeat libraries produced this way for each of the species separately as an input to RepeatMasker to detect the repeats in the genomes.

I used the regions defined as low quality to filter out the transposable elements which might be poorly assembled in the genomes of *E. subterraneus* and *S. domuncula*. Elements were filtered out if they overlapped in any number of bases with any of the low quality regions in the genome.

### 2.5.2 Identification of full length, solitary and pseudo-elements

Different classes of transposable elements can be found in the genomes in various forms, as explained in the introduction, for details see figure 3 and the reference (Chuong, Elde and Feschotte, 2017). I identified solo and intact LTR elements and full length and pseudoelements in the LINE repeat class. To identify solo and full length intact LTR elements, I used the library constructed with LTR_retriever in step 2, along with RepeatMasker (*RepeatMasker Home Page*, no date) to identify all hits in the genomes. LTR_retriever is then used to identify full length (intact) LTR elements and solitary elements ("solo" LTRs). In short, the LTR_retriever filters whole-genome RepeatMasker annotations based on the following structural features: Only annotation hits with alignment score higher than 300 and alignment length more than 100bp are retained. A hit is proclaimed to be "solo" LTR if it shows similarity to LTR element and no such LTR copy is present in the adjacent four annotations, or the difference of divergence between target LTR and other LTRs from the same family larger than 4%, and no internal regions located within 300 bp distance flanking the target. Complete LTR elements were identified if they satisfy the condition: structures of LTR-INT-LTR or LTR-INT-INT-LTR, the distance between annotation entries should be less than 300 bp, and the difference of divergence between LTRs less than 4% (Ou and Jiang, 2018).

## 2.6 Assessing the contribution to genome evolution

To analyse the contribution of transposable elements to the evolution of the genomes, I first annotated the genomes by defining the genes. Next, I divided the genomes into regions containing introns and exons. All regions which were not assigned to any gene were included in the intergenic region. I analysed the contribution of transposable elements to the organization of the genome by defining the overlaps between transposable elements and each region.

### 2.6.1 Annotating the genomes

I obtained the published annotations of the genomes when they were available. Gene models were downloaded for the genome of *Ephydatia muelleri* from the *E. muelleri* genomic resource accompanying the paper describing the genome (Kenny *et al.*, 2020). Gene models for *Amphimedon queenslandica* were downloaded from Ensembl (Howe *et al.*, 2020), for the assembly Aqu1. Gene models for *Tethya wilhelma* were downloaded from the supplementary material of the paper (Francis *et al.*, 2017).

Annotations for the genomes of *Sycon ciliatum* and *Oscarella pearsei* were not publicly available so the gene models for those sponge species along with the *S. domuncula* and *E. subterraneus* which I assembled, were made "in house". Genomes were annotated using the whole genome automated annotation pipeline BRAKER2 (Hoff *et al.*, 2019). BRAKER is a computational pipeline which uses GeneMark-ET (Lomsadze, Burns and Borodovsky, 2014) to find potential gene candidates *ab initio* and models the initial parameters for the model that predicts if a sequence is a gene or not. If RNA-seq reads are available, it uses the information from the splicing of the reads to improve the initial parameters based on the information about the position of introns and donor and acceptor sites. The gene-finding and parameter optimization then iteratively happen for several rounds until the parameters for the final model are optimized. BRAKER pipeline next uses the genes predicted by GeneMark-ET as an input of genes for the training of the AUGUSTUS gene prediction tool (Keller *et al.*, 2011). AUGUSTUS is among the most accurate gene prediction tools which integrates extrinsic evidence already in the gene prediction step (unlike GeneMark which uses it for parameter optimization only). For the genome of *E. subterraneus* RNAseq reads were available, and they were used as hints. Also, the transcriptome was assembled de novo and a protein coding set of sequences was extracted from it (explained previously in the methods), and this was also used to improve the gene prediction. *Sycon ciliatum* gene models were made with RNA hints, produced from publicly available data (SRA accession number PRJEB7138), and *O. pearsei* gene models were made with RNA and protein hints (PRJNA230477) by Juan Antonio Ruiz Santiesteban, a colleague from our group.

### 2.6.2 Defining the intron, exon and intergenic regions

I divided the genomes into introns, exons and intergenic regions using a custom R script in the following way; first, I collapsed all the gene models to remove alternative transcripts and define a set of all exons for each gene. Next, the regions between the exons were defined as introns for each gene. I assigned numbers to exons and introns based on their order in the gene in the coding orientation: first exon for the gene encoded on the "+" strand was the exon which appeared first when ordered by genomic location, and conversely, the first exon for the gene encoded on the "-" strand was the exon which appeared last from the set of exons in this gene when exons were ordered by genomic location. Analogously, the order was assigned to the rest of the exons and introns. All exons together for each species are considered as a single region marked as "exons". All introns together were considered "introns", and the regions of the genome not defined as genes were considered as an "intergenic" region.

### 2.6.3 Assessing the contribution to genome organisation

I analysed the impact of the transposable elements on the evolution of the sponge genomes by assessing the contribution of transposable elements to the genome organization. After I divided the genomes into exons, introns and intergenic regions, I calculated the abundance of transposable elements in each region, by summing up the length of overlapping fragments between TEs and the regions. The search of overlaps was done by Paula Štancl, a master student in our group. I calculated percent of total feature contributed by transposable elements by calculating the percentage of total number of bases in introns, exons and intergenic regions which overlap any transposable element.

I analysed abundances of transposable elements in introns, exons and intergenic regions. Since the lengths of the regions are not equal within a genome or among different genomes, I do not expect a uniform distribution of transposable elements neither within a single genome, nor between different genomes. For this reason, I assumed that the abundance of transposable elements in any region (intron/exon/intergenic) depends only on the total size of the region in the genome and the total length of transposable elements in this group. Thus, I calculated the expected abundance of transposable elements belonging to group X in the region Y by dividing the total length of all elements which belong to this TE group proportionally to the sizes of the regions in the genome. I calculated the enrichment of observed value over expected value as percent of the difference between the observed abundance and expected abundance compared to the expected abundance value. All the analysis was done using custom R scripts R scripts using R 4.1.0 (https://www.R-project.org/).

I assessed the conservation of transposable elements by determining the sequence divergences among elements of the same group of transposable elements compared to the consensus element. Only elements which were identified as at least 90% of the consensus length were taken into consideration. The sequence divergences were calculated by the RepeatMasker (*Website,* no date) in the process of identification of repetitive elements. I visualized the sequence divergences using custom R scripts using R 4.1.0 (https://www.R-project.org/ ) and the ggplot2 (Wickham, 2011) package. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge (where IQR is the interquartile range, or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge.

Next, I separately analysed the conservation of the LTR elements. The group of LTR elements was subdivided into “intact”, “solo” and “unassigned” subgroups (see 3.5.2 for

details). I analysed the contribution of the defined subgroups to genome organization, and their sequence divergences across different genomes in the same way as explained previously. All whiskers on the boxplots extend from the interquartile range to +- 1.5*interquartile range.

### 2.6.4 Expression analysis

I analysed the gene expression in the sponge *E. subterraneus* in the first and 10th day of primmorph formation. First, the raw RNA seq data was processed to remove the low quality ends from the reads. The bases were trimmed from the ends of the reads if their quality was lower than 3 in phred score. If the trimmed read was shorter than 20 nucleotides, it was discarded from the set. The trimming and filtering was performed by bbduk, a part of BBtools (http://sourceforge.net/projects/bbmap/ ).

The trimmed RNAseq reads were mapped to the genome of *Eunapius subterraneus* using the default parameters for the mapping of short reads with minimap2.17-r941 (Li, 2018). NOTE: Although the parameters are not optimal for RNA-seq reads mapping, by manual comparison with another RNA specific splice aware mapper, I observed no extreme differences, and even the splice sites were nicely defined, so minimap2 was used because of a notably shorter mapping time instead of the mentioned more commonly used option. After the mapping to the genome, Paula calculated the number of reads mapped per kilobase transcript per million reads (RPKM value). The value was calculated by counting the number of reads that map from start to end of each gene, and dividing this number by the length of the gene's coding region (sum of exon lengths), and by number of millions of reads mapped. This scaling enables a relatively fair comparison among libraries of different sizes and is invariant of gene length.

To analyse the potential impact of transposable elements on gene expression, I divided the genes into separate groups, depending on the overlaps with transposable elements. Genes which did not overlap with any transposable element were assigned to the "No overlap with TE" group. For all genes that overlapped at least one transposable element, I summed the width of the intersect between each gene and all transposable elements which overlap the gene. For each gene I summed the intersect widths by groups of TE (DNA, RC,LTR, LINE and Unknown). I assigned the gene to the group which had the largest overlap size with the gene. For every gene, I also determined whether it is a gene encoded by a transposable element or not in the following way: if a gene was completely inside the transposable element, I defined it to be a transposable element (TE) gene. Other genes were defined as non-TE genes. All the analysis was done using a custom R script.

To determine potential difference between the expression of TE encoded genes and non-TE encoded genes, I plotted the values of the gene expression from the dataset obtained on day 1 of primmorph formation, separately as a boxplot for each group of transposable elements (DNA, RC, LTR, LINE, Unknown, No overlap with TE) and type of gene (TE/non-TE gene). The whiskers in the boxplot extend to IQR±1.5*IQR. The value for the expression was logarithmically transformed to reduce variability in the data due to skewness of the distribution of RPKM values, which is normal for any expression data (e.g. RNAseq). All the analysis and visualization was done with a custom script using R version 4.1 and data.table version 1.12.8 (https://cran.r-project.org/web/packages/data.table/index.html) and ggplot2 packages both available on CRAN https://cran.r-project.org/ .

To observe the differences between the expressions of TE encoded genes and non-TE encoded genes, I plotted the RPKM expression values for day 1 and day 10 of the primmorph formation as a scatterplot. I divided the genes into 10 equally sized groups based on their expression values measured in day 1. Groups one and two both had the median expression value of 0, so they were merged together as the group containing 20% least expressed genes. Groups 3 - 8 when ordered by expression values from smallest to largest were named appropriately as [20% - 30%> , … , [80%-90%> denoting that the genes in this group have the expression value in the bottom [A%-B%> of all values. Group 9 was named "10% most expressed". Again, the RPKM expression values were logarithmically transformed. All the analysis and visualization was done with a custom script using R version 4.1 and data.table version 1.12.8 (https://cran.r-project.org/web/packages/data.table/index.html) and ggplot2 packages both available on CRAN https://cran.r-project.org/ .

For the genes which have a transposable element inserted into them, I analysed if there is a difference between the expression of genes in which the insertion was into an exon, and genes for which the insertion was into an intron. I divided the genes which have an overlap with TE into the mentioned groups based on the total length of the overlap. Genes with total length of the intersection between TE and introns larger than between TEs and exons, were assigned into "intron" group, and vice versa for the "exon" group. The expression values are plotted based on the element type which was most prevalent in the gene measured by length of the intersect (DNA, RC, LTR, LINE and unknown), and the intron/exon group.

Finally, I analysed the expression of the genes based on the type of the contribution which the TE has to the gene. In other words, for every transposable element, Paula Štancl determined if the element overlapped with 3' exon, 5' exon, any other internal exon or intron. I grouped the elements based on those groups. The transposable elements which encoded their

genes were grouped separately into “transcript inside” group. I represented each element with the expression value of the gene it overlaps with, measured on the first day of the primmorph formation. I plotted the values as boxplots with the ranges defined as before. I did this analysis again using a custom R script in R4.1 with data.table 1.12.8 (https://cran.r-project.org/web/packages/data.table/index.html) and ggplot2.

### 2.6.5 Identification of the homologs of small RNA machinery

I identified the sponge homologs of components of the piRNA pathway in the following way: I used the homologs from the human genome and used the protocol from (Grimson *et al.*, 2008; Marchler-Bauer *et al.*, 2017) to find homologs in the sponges. In short, I used diamond blastp to search for the identified piRNA pathway proteins in the translated protein sequences found in sponges. I only selected the one top scoring sponge sequence for each human homolog and used them as a query against a non redundant protein set of sequences from the human genome. If the best hit in such reciprocal search matched the expected family member, those query sequences were considered to be potential homologs. In the same way I found all the homologs of those potential genes by all pairwise sponge comparisons. I found the conserved domains in the homologs by conducting a web conserved domain search (Marchler-Bauer *et al.*, 2017). All of the results were further analysed manually by Paula Štancl for the presence of conserved diagnostic domains (Marchler-Bauer *et al.*, 2017): two DEAD-like helicase domains for the DDX4 and MOV10L1 proteins, motor domain superfamily for KIF17, Paz and Piwi domain for Piwi family members, Maelstorm domain for MAEL, PLDc superfamily domain for PLD6, PRMT5TIM, PRMT5 and PRMT5C domains for PRMT, 6 TDR domains for RNF17, zf and multiple TDR domains for TDRD1, HRPa and a TUDOR domain for TDRD9, two KH1 and a TUDOR domain for TDRKH, Hen1 superfamily domain for Hen1 and WD40 superfamily for WDR77. Candidates without any one of the required diagnostic domains were excluded, while candidates containing all domains but some partially conserved were marked partial. The homologs in all species and the representative domains for each protein are shown in the appendix.

I plotted the gene expression values measured on day 1 and day 10 of the primmorph formation as a scatterplot and labeled the homologs on the plot using a custom R script. The genes were colored as explained before, based on their expression group. The RPKM values are logarithmically transformed for the reasons mentioned before.

# 3 Results

## 3.1 Raw reads preprocessing

Very conservative quality trimming and filtering removed 53.2% of all bases from the *E. subterraneus* Illumina data set, leaving 26.9 billion bases (Table 4). All of the 211.9 million reads left after filtering have an average quality higher than 30, meaning that we expect on average less than 1 error in 1000 bases. Using an average genome size of sponges (200Mb) as an approximation of the genome size, the calculated coverage of the genome by those high quality Illumina reads is 135x. After the removal of adapters, trimming of all bases with quality lower than 25 and removing all reads with average quality lower than 30, the *S. domuncula* data set was left with 31.9 billion bases. Assuming the same approximation for the genome size, on average, each location of the *S. domuncula* genome should be covered by at least 160 reads (Table 5).

*Table 4. Results of the quality trimming and filtering of Illumina reads for the E. subterraneus.*

| *Eunapius subterraneus* | | | | |
|---|---|---|---|---|
| | number of reads | number of bases | reads (%) | bases (%) |
| Input: | 381488100 | 57604703100 | 100.00 | 100.00 |
| QTrimmed: | 269680206 | 25323562204 | 70.69 | 43.96 |
| KTrimmed: | 1093372 | 151008885 | 0.29 | 0.26 |
| Low quality discards: | 42847986 | 5138496496 | 11.23 | 8.92 |
| Total Removed: | 169578282 | 30613067585 | 44.45 | 53.14 |
| Result: | 211909818 | 26991635515 | 55.55 | 46.86 |

*Table 5. Results of the quality trimming and filtering of Illumina reads for the S. domuncula.*

| *Suberites domuncula* | | | | |
|---|---|---|---|---|
| | number of reads | number of bases | reads (%) | bases (%) |
| Input: | 431653482 | 65179675782 | 100.00 | 100.00 |
| QTrimmed: | 284701883 | 25654652688 | 65.96 | 39.36 |
| KTrimmed: | 610901 | 70883851 | 0.14 | 0.11 |
| Low quality discards: | 50704634 | 6206555993 | 11.75 | 9.52 |
| Total Removed: | 177253912 | 31932092532 | 41.06 | 48.99 |
| Result: | 254399570 | 33247583250 | 58.94 | 51.01 |

After the removal of adapter sequences from the nanopore data sets and filtering out all reads shorter than 200 bases, the estimated coverage for the *E. subterraneus* data set was 19.7, and 69.4 for the *S. domuncula* (table 6). Since shorter reads do not improve the assembly quality, and can lead to more fragmented assembly, they were removed from the *S. domuncula* data set. After removal of reads shorter than 5000, the estimated coverage decreased to 40.1 and this is the data set I used for the assembly. The number of bases in the *E. subterraneus* data set after the removal of reads shorter than 5000 was 2677086352, which reduced the coverage to 13.4. Since the decrease in coverage would impact the quality of the assembly, all reads longer than 200 bp were used.

*Table 6. Read statistics for nanopore reads after quality filtering.*

| Species | Dataset | Number of reads | Number of bases | Estimated coverage | N50 of read length | Mean read length | Median read length |
|---|---|---|---|---|---|---|---|
| *Eunapius subterraneus* | Nanopores | 1061783 | 3938829765 | 19.7 | 8645 | 3709.637 | 1573 |
| *Suberites domuncula* | Nanopores | 5355451 | 13879366742 | 69.4 | 6116 | 2591.634 | 968 |
| *Suberites domuncula* | Nanopores longer than 5000 | 842015 | 8011339630 | 40.1 | 10045 | 9514.486 | 8039 |

To test if the correction of nanopore reads prior to assembly would improve the assembly, I corrected the nanopore reads with Illumina high quality reads. The results of the correction are shown in table 7. The correction procedure split some of the reads, which is why the overall number of the reads increased slightly. The total number of bases has also increased by 1.8% because some of the indels were corrected.

*Table 7. Statistics for the corrected nanopore dataset of E. subterraneus*

| Species | Dataset | Number of reads | Number of bases | Estimated coverage | N50 of read length | Mean read length | Median read length |
|---|---|---|---|---|---|---|---|
| *Eunapius subterraneus* | Nanopores | 1061783 | 3938829765 | 19.7 | 8645 | 3709.637 | 1573 |
| *Eunapius subterraneus* | Lordec corrected nanopores | 1064507 | 4013267263 | 20.1 | 8791 | 3770.071 | 1598 |

## 3.2 Genome assembly and annotation

I assembled the genomes of *E. subterraneus* and *S. domuncula* into scaffolds from raw nanopores and polished them to correct errors. I removed the scaffolds which matched to bacteria, and determined the regions in the genomes which are potentially problematic in terms of quality. I assessed the qualities of the produced assemblies and compared them to existing publicly available poriferan genome assemblies. I further annotated the genomes and compared them with publicly available annotations.

### 3.2.1 Assembly results

The strategy for the assembly was to assemble the raw nanopores into scaffolds, and correct the remaining errors in the nanopore-only scaffolds. This polishing is done with high quality Illumina reads. Since the mapping of the short reads to the genome is problematic in low complexity regions, I also polished the assembly separately with paths from the assembly graph instead of reads. Finally, I used a combination of paths and high quality Illumina reads to further refine the assembly (see methods 3.3.3.2 for details, Figure 5), and chose the best assembly tactic based on predicted genome completeness.

*E. subterraneus* nanopore dataset was assembled with Flye assembler into 3664 scaffolds longer than 500 bases and had a total length of 202Mb. *S. domuncula* nanopore data set had a higher coverage, so the assembly based only on nanopore reads produced a more contiguous result - it was assembled into 900 scaffolds of which only one was shorter than 500 nucleotides. Statistics for the nanopore-only assemblies, assembly graphs containing edges, and paths produced from assembly graphs are shown in the table 8.

Since the nanopore only assemblies contained errors, I removed them by polishing the assembly in the following way: All assemblies were polished with both high quality Illumina reads (Walker et al., 2014) and paths created from Illumina-only assembly graph. The polishing of the produced nanopore only assemblies was done in the following combinations:

- Using only short reads in one iteration ("Reads only"),
- Using only paths constructed from the assembly graph ("Paths only"),
- Using short reads in the first round and the paths from the assembly graph constructed on the read-polished assembly in the second round ("Reads->Paths"),
- Using the paths from the assembly graph in the first round and short reads in the second round ("Paths->Reads"),
- Using short reads in two rounds consecutively ("Reads->Reads").

I evaluated the polished assemblies using BUSCO and chose the best scoring one as the final one. The BUSCO orthologs are defined as sets of genes from OrthoDB database of orthologs (www.orthodb.org) present in a single copy in more than 90% of the species. In this way a set of orthologs was defined for eukaryotes (eukaryota_odb10) and for metazoans (metazoa_odb10). I chose the best assembly for each species based on BUSCO score and used only this one in further analysis.

For the polishing, I first assembled high quality Illumina reads for *E. subterraneus* and *S. domuncula* into separate assembly graphs using SPAdes assembler. *E. subterraneus* assembly graph had 685.6 thousands edges and the total length of all edges was 315Mb. *S. domuncula* assembly graph was less complicated and contained 319.8 thousand edges.

I constructed two different sets of paths - first paths were constructed using the nanopore-only assembly as a scaffold and are noted as "Paths in round 1" in the table. Second set of paths was constructed based on an assembly corrected with high quality Illumina reads ("Paths in round 2".).

*Table 8. Statistics for the raw nanopore-only assemblies, assembly graphs containing edges, and paths produced from assembly graph*

| | ***Eunapius subterraneus*** | | | | ***Suberites domuncula*** | | | |
|---|---|---|---|---|---|---|---|---|
| **Statistics without reference** | **Edges** | **Nanopore -only** | **Paths in round 1** | **Paths in round 2** | **Edges** | **Nanopore -only** | **Paths in round 1** | **Paths in round 2** |
| # scaffolds | 150424 | 3664 | 34583 | 32637 | 117228 | 899 | 6322 | 6484 |
| # scaffold (>= 50000 bp) | 20 | 1138 | 141 | 53 | 26 | 393 | 544 | 569 |
| Largest scaffold | 77780 | 867181 | 533635 | 123943 | 99248 | 5245982 | 422147 | 271963 |
| Total length/Mb | 315.7 | 202.2 | 179.3 | 161.1 | 218.6 | 123.4 | 106.2 | 106.2 |
| Total length/Mb (>= 50000 bp) | 1.2 | 172.8 | 13.3 | 3.3 | 1.6 | 115.1 | 44.8 | 49.0 |
| N50 | 3723 | 162773 | 9950 | 9241 | 3109 | 514763 | 41485 | 45374 |
| N75 | 1541 | 78596 | 4693 | 4437 | 1205 | 194989 | 19854 | 20804 |
| L50 | 21910 | 364 | 4567 | 4613 | 14474 | 59 | 728 | 656 |
| L75 | 54814 | 804 | 11108 | 10900 | 43943 | 158 | 1667 | 1521 |
| GC (%) | 43.41 | 44.1 | 43.29 | 43.13 | 40.28 | 40.4 | 40.14 | 40.06 |
| # N's | 0 | 2100 | 2374 | 0 | 0 | 1300 | 640 | 0 |
| # N's per 100 kbp | 0 | 1.04 | 1.32 | 0 | 0 | 1.05 | 0.6 | 0 |

The analyses of assemblies annotation completeness are shown in Table 9 and Table 10. *S. domuncula* nanopore-only assembly shows high completeness even with no polishing. From the total 255 BUSCO orthologs we would expect to find in eukaryotes in a single copy, this assembly has 206. 7 BUSCOs were found in assembly but were duplicated, and 20 were not found. Out of 954 orthologs we would expect to find in a single copy in metazoan species, this assembly has 754. The nanopore-only assembly for *E. subterraneus* showed less annotation completeness and I was only able to detect 557 metazoan orthologs. The polishing of assemblies improved on those results. For the assembly of *S. domuncula* the BUSCO results on eukaryotic and metazoan datasets improved as shown in the table 10. In the first round, after polishing with paths the overall completeness found in the eukaryotic data set was improved by 9.8%, while polishing with reads instead of paths improved the completeness by 8.6%. In the second round, both combinations involving paths and reads showed similar improvement on the eukaryota BUSCO data set. In both cases the total completeness was 95.3%, but the polishing with paths first and reads after increased the percentage of duplicated BUSCOs to 4.3%. Of note, when the assembly was polished iteratively with reads alone, the overall completeness was 92.9% on the eukaryotic data set, and 89% on the metazoan data set.

Table *9. Annotation completeness of the differently polished assemblies for the S. domuncula genome. The estimation is performed by BUSCO based on the number of found complete, fragmented and missing single copy orthologs expected to exist in all eukaryotes and all metazoans.*

| ***Suberites domuncula*** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **BUSCO:** | | **Nanopore only assembly** | **Polished with:** | | | | |
| | | | **Paths only** | **Reads only** | **Paths-> Reads** | **Reads -> Paths** | **Reads->Reads** |
| **Eukaryota Odb10** | Complete | 83.5 | 93.3 | 92.1 | 95.3 | **95.3** | 92.9 |
| | Complete, single copy | 80.8 | 89.4 | 88.6 | 91 | **91.8** | 89.4 |
| | Complete, duplicated | 2.7 | 3.9 | 3.5 | 4.3 | **3.5** | 3.5 |
| | Fragmented | 8.6 | 3.9 | 4.3 | 2.4 | **2.4** | 4.3 |
| | Missing | 7.8 | 2.8 | 3.6 | 2.3 | **2.3** | 2.8 |
| **Metazoa Odb10** | Complete | 81.4 | 89.9 | 89.1 | 89.5 | **90.7** | 89 |
| | Complete, single copy | 79 | 86.5 | 85.7 | 86.1 | **87.1** | 85.5 |
| | Complete, duplicated | 2.4 | 3.4 | 3.4 | 3.4 | **3.6** | 3.5 |
| | Fragmented | 4.7 | 1.9 | 2.3 | 1.7 | **1.6** | 2.2 |
| | Missing | 13.9 | 8.2 | 8.6 | 8.3 | **7.7** | 8.8 |

Due to lower initial nanopore coverage, the assembly of *E. subterraneus* genome was more challenging. Only 58% complete BUSCO orthologs were identified in the nanopore-only *E. subterraneus* assembly (58.5% and 58.4% for eukaryotic and metazoan ortholog set, respectively). Polishing with reads alone increased the annotation completeness to 87.9%, compared to 80.4% when the polishing was done by paths alone. However, in the second round of polishing, the results are again similar regardless of the order of polishing with paths and reads. The best BUSCO completeness on the metazoan dataset (82%) is achieved when the assembly was refined with reads in the first round and paths in the second round(table 10).

Of note, when the nanopore data set was corrected prior to the assembly with short reads the initial nanopore-only assembly was superior compared to nanopore-only assembly with no prior correction. Its completeness on the eukaryotic set of orthologs was 86.3%, and was 76.3% complete when the metazoan set of orthologs was used. However, the results of the final polishing did not further vastly improve the pre-corrected nanopore-only assembly, and the

results never outperformed the assemblies with no pre-correction on the larger metazoan set of orthologs. Full results are added to the appendix as table 16.

*Table 10. Annotation completeness of the differently polished assemblies for the E. Subterraneus genome. The estimation is performed by BUSCO based on the number of found complete, fragmented and missing single copy orthologs expected to exist in all eukaryotes and all metazoans.*

| ***Eunapius subterraneus*** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **BUSCO:** | | **Nanopore only assembly** | **Polished with:** | | | | |
| | | | **Paths only** | **Reads only** | **Paths-> Reads** | **Reads -> Paths** | **Reads->Reads** |
| **Eukaryota Odb10** | Complete | 58.5 | 80.4 | 87.9 | 89.5 | 88.7 | 89.8 |
| | Complete, single copy | 56.9 | 79.2 | 85.9 | 87.5 | 86.7 | 87.8 |
| | Complete, duplicated | 1.6 | 1.2 | 2 | 2 | 2 | 2 |
| | Fragmented | 20 | 8.2 | 5.9 | 5.1 | 5.9 | 4.3 |
| | Missing | 21.5 | 11.4 | 6.2 | 5.4 | 5.4 | 5.9 |
| **Metazoa Odb10** | Complete | 58.4 | 74 | 81.6 | 81.4 | **82** | 81.9 |
| | Complete, single copy | 57.9 | 72.7 | 80.2 | 80 | **80.6** | 80.3 |
| | Complete, duplicated | 0.5 | 1.3 | 1.4 | 1.4 | **1.4** | 1.6 |
| | Fragmented | 15.1 | 7.4 | 4.4 | 4.5 | **4.1** | 4.2 |
| | Missing | 26.5 | 18.6 | 14 | 14 | **13.9** | 13.9 |

Based on the BUSCO results, polishing with reads in the first round and paths in the second round consistently showed best results in both *E. subterraneus* and *S. domuncula* datasets, thus those assemblies were chosen as the best ones and used in further analysis.

### 3.2.2 Filtering out bacterial scaffolds and identification of low quality regions

The GC content of assemblies ranges from 35.8 in *A. queenslandica* to 47.0 in *S. ciliatum*. GC content density plot shows a mixture of two different distributions for the genomes of *S. ciliatum* and *E. subterraneus* (figure 6). This indicates a presence of another genome, and was resolved after the removal of bacterial scaffolds in further analysis.

.

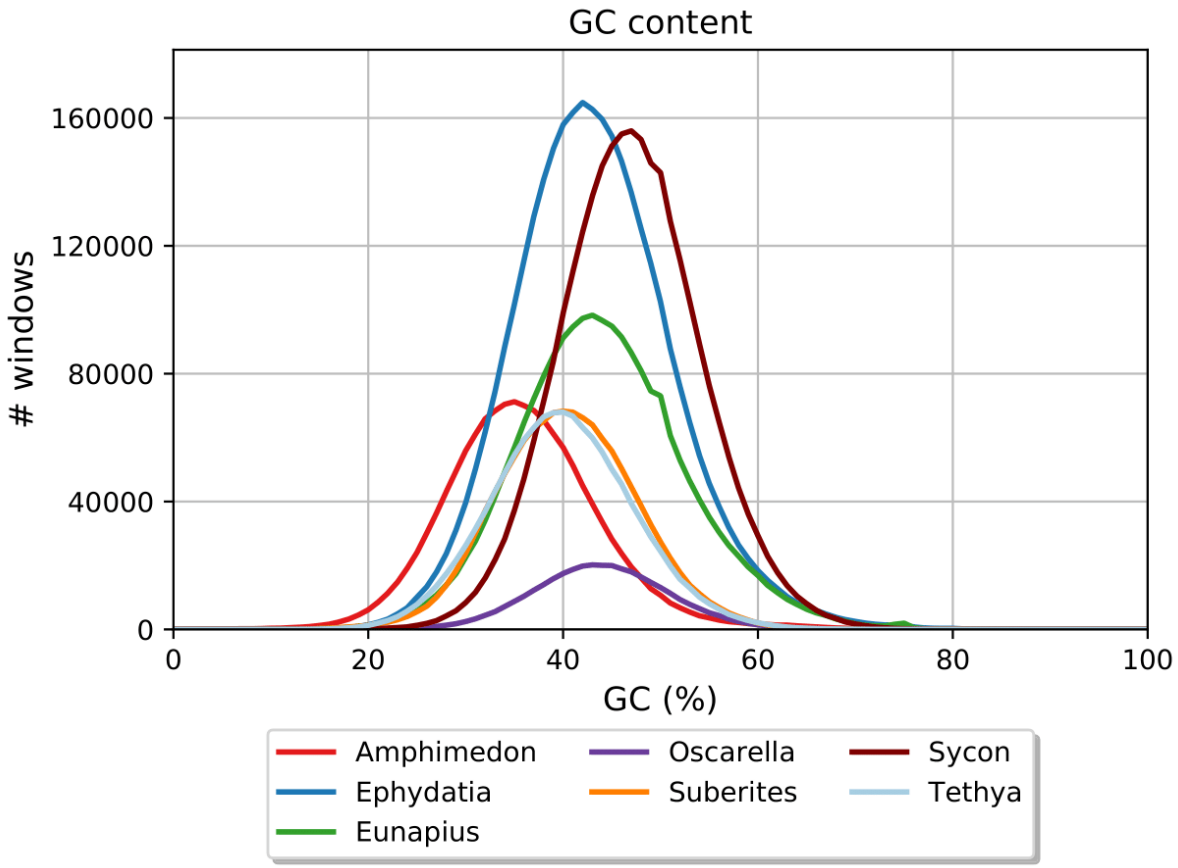



*Figure 6. GC content distribution on the available sponge genomes before the removal of bacterial scaffolds.*

I used MEGAN to detect bacterial scaffolds in the assembled genomes. In total, 355 scaffolds from the *E. subterraneus* assembly were identified as bacterial, summing up to the total length of 17.3Mb. Most of the bacterial scaffolds (14.9Mb) were assigned to proteobacteria, while the rest was unknown. I detected 67 scaffolds assigned to bacteria in the genome of *S.domuncula*. Their total length was 22.5Mb, and they were also mostly assigned to Proteobacteria. Although the authors reported to have removed all bacterial scaffolds from the assembly, I detected seven bacterial scaffolds in the genome of *E. muelleri*. Results for all genomes are shown in the table 11. All detected scaffolds were removed from the assemblies for the remaining analysis.

*Table 11. Bacterial scaffolds in the assembled genomes.*

| Assembly | Amphimedon | Ephydatia | Eunapius | Oscarella | Suberites | Sycon | Tethya |
|---|---|---|---|---|---|---|---|
| # scaffolds | 13397 | 1444 | 3694 | 67767 | 900 | 7780 | 5936 |
| # bacterial scaffolds | 136 | 7 | 355 | 508 | 67 | 284 | 77 |
| Total length before the removal of bacterial scaffolds /Mb | 166.7 | 322.6 | 202.8 | 57.8 | 123.8 | 357.5 | 125.7 |
| Total length of the bacterial scaffolds /Mb | 0.7 | 0.1 | 17.3 | 0.4 | 22.5 | 12.8 | 0.5 |

I defined low quality regions in the assemblies by mapping nanopore and high quality Illumina reads back to the genome. For the genome of *E. subterraneus*, 1.33% of the bases were of low quality when judged by nanopore reads, most of which (1.05% out of 1.33%) was located inside the scaffolds. Those regions indicate wrongly assembled sequences (chimeras). 10.25% bases are low quality with respect to Illumina reads. Those regions are not polished as efficiently as the rest of the genome, and are mostly derived from nanopore reads so are more erroneous. The assembly of *S.domuncula* genome had only 0.1% bases not confirmed by nanopore reads, and 10.8% bases were not confirmed by high quality Illumina reads.

### 3.2.3 Comparison of the assemblies with publicly available genomes

The final polished assembly of *E. subterraneus* genome consists of 3339 scaffolds and has a total size of 185.5 million bases. The total size of the assembled genome for *S. domuncula* is 101.3 million bases. The N50 value for *E. subterraneus* is 166.8 kb, while *S. domuncula* has an even higher value of N50, 420 kb. The genome of *O. pearsei* is the most fragmented with the total of 15342 scaffolds larger then 500 bases, and the genome of *E. muelleri* was the least fragmented, with most of the genome assembled in the largest 24 scaffolds, and an N50 value of 9.8 million bases. Genomes that I assembled had the least amount of unknown bases (around 1 per 100kbp), while the genome of *S. cilliatum* had over 22%, and the genome of *A. queenslandica* 13%.

*Table 12. QUAST results for publicly available and assembled sponge genomes. Mb=million bases.*

| Assembly | **Amphimedon** | **Ephydatia** | **Eunapius** | **Oscarella** | **Suberites** | **Sycon** | **Tethya** |
|---|---|---|---|---|---|---|---|
| # contigs (>= 0 bp) | 13261 | 1437 | 3339 | 67259 | 833 | 7496 | 5859 |
| # contigs (>500 bp) | 13261 | 1419 | 3309 | 15342 | 832 | 7496 | 5859 |
| # contigs (>= 50000 bp) | 646 | 252 | 1026 | 27 | 360 | 1750 | 727 |
| Total length (>= 0 bp) /Mbp | 166 | 322.5 | 185.5 | 57.4 | 101.3 | 344.7 | 125.2 |
| Total length (>= 50000 bp) /Mbp | 109.9 | 299.2 | 159.2 | 1.7 | 93.6 | 288.3 | 77.9 |
| Largest contig /Mbp | 1.9 | 34.7 | 0.9 | 0.1 | 2.2 | 1 | 0.7 |
| GC (%) | 35.76 | 43.19 | 43.67 | 43.7 | 40.16 | 46.97 | 39.92 |
| N50 | 121800 | 9883643 | 166815 | 8583 | 420097 | 170675 | 73988 |
| N75 | 20221 | 7973121 | 81986 | 2498 | 156862 | 81913 | 25715 |
| L50 | 306 | 11 | 330 | 1266 | 64 | 572 | 480 |
| L75 | 1116 | 20 | 717 | 4004 | 162 | 1289 | 1168 |
| # N's per 100 kbp | 13040 | 577.85 | 1.02 | 3385.51 | 0.71 | 22183 | 1207.73 |

Analysis of the annotation completeness for the assembled genomes is shown on the figures 7 and 8. The genome for *S. domuncula* scored the best among all analysed genomes, with predicted 92.2 % of complete single copy BUSCOs from eukaryotic set of orthologs, and 88.2% from metazoan set. Genome of *E. subterraneus* was predicted to be 80.7% complete, estimated on the metazoan BUSCO set. Genome which is most contiguous in assembly, that of *E. muelleri*, is missing 12.1% of BUSCO eukaryotic orthologs and 21.6% BUSCO metazoan orthologs, and over 7% of the BUSCOs are found duplicated. Genome of *O. pearsei* is the least complete of the analysed genomes, with only 65.4% metazoan orthologs found as complete and in a single copy.

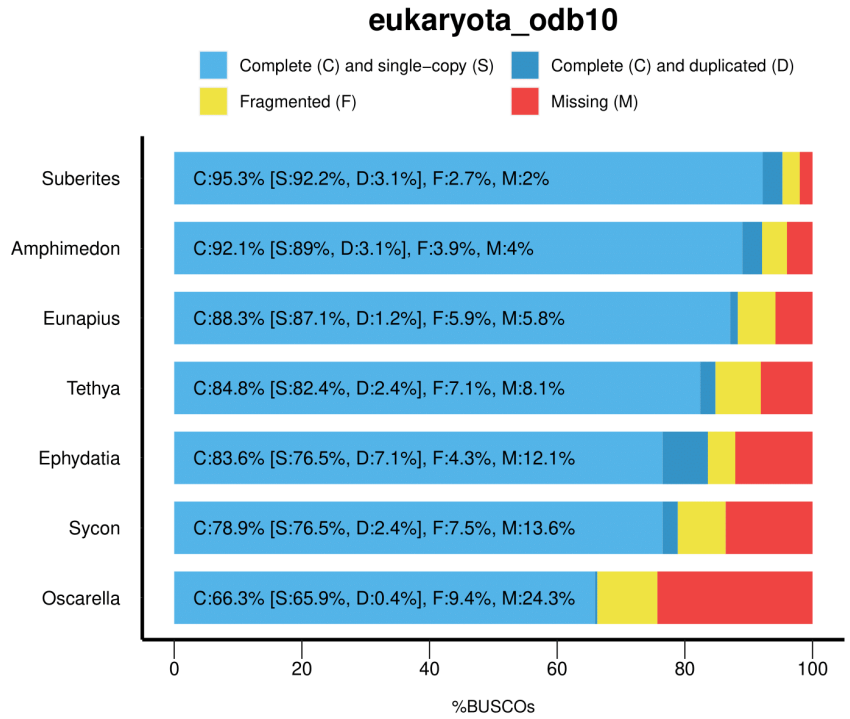



*Figure 7. Estimated annotation completeness based on the BUSCO set of orthologs expected in eukaryotes for all analysed sponge genomes*

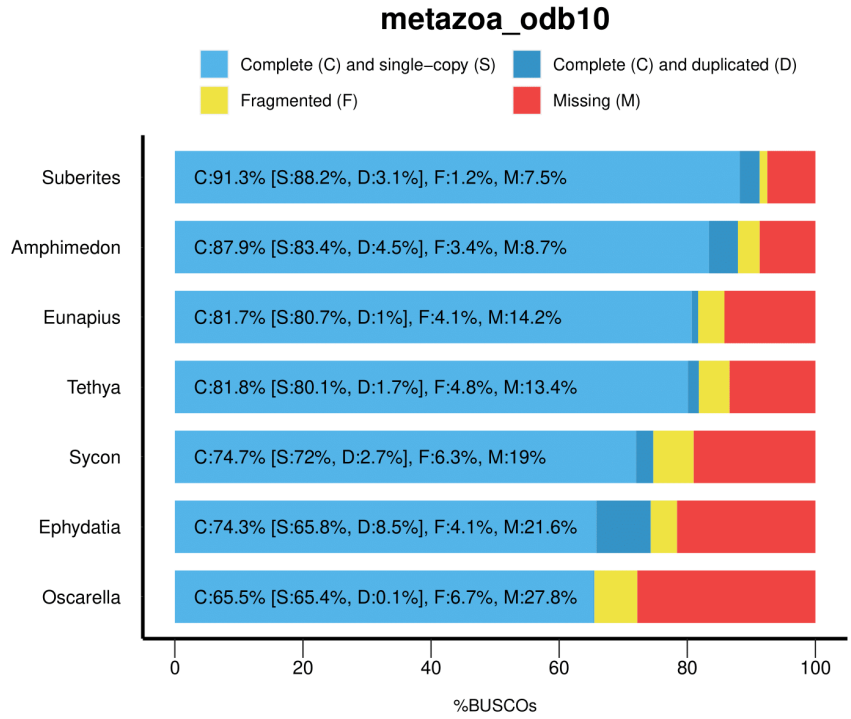



*Figure 8. Estimated annotation completeness based on the BUSCO set of orthologs expected in metazoa for all analysed sponge genome*

## 3.3 General characteristics of the sponge genomes

In this chapter, I present the general characteristics of the sponge genomes. I first compared the number of genes and their lengths among the genomes and analysed the correlation between genome contiguity and gene lengths.

Next, I analysed the total length of exons, introns and intergenic regions in the light of size of the genomes. I also compared the intron, exon and intergenic regions lengths among sponge genomes. Finally, I compared the distribution of lengths of first introns and non-first introns in all sponges.

### 3.3.1 Dinucleotide content

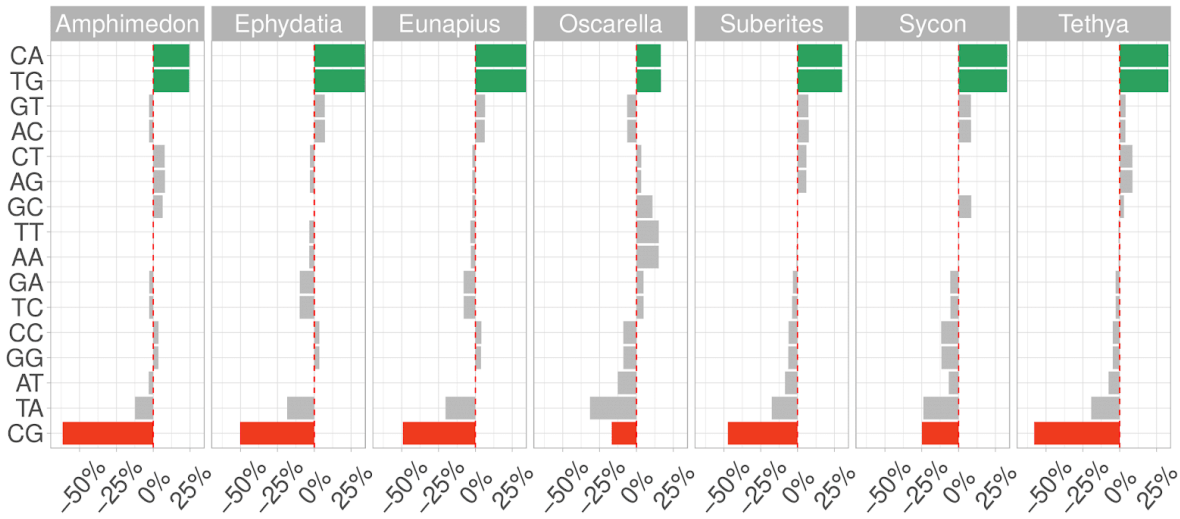



*Figure 9. Enrichment in dinucleotide frequencies for sponge species. The enrichment is calculated as percent of offset for observed value of the dinucleotide frequency from the expected value of the dinucleotide frequency ( (observed-expected) / expected), where the expected frequency of dinucleotide XY is calculated by multiplying the observed frequencies of nucleotides X and Y. The dinucleotides are ordered by median frequency for enrichment in all sponges, in a way that the dinucleotide with highest median enrichment among all sponges is shown on the top and dinucleotide with the highest depletion is shown in the bottom. The red dashed line represents the expected enrichment of dinucleotide frequency. It is visible that CpG dinucleotides are depleted in all sponge species.*

Figure 9 shows enrichment in dinucleotide frequencies for sponge species. The level of depletion in Oscarella and Sycon is the lowest (-16.8% and 24.9% respectively), while the CpG depletion is the most pronounced in Amphimedon queenslandica where CpG dinucleotides appear 61.4% less frequently than expected. Other analysed genomes also show the CpG dinucleotides depletion. All of the genomes also show CpA and TpG excess, which is consistent with the demethylation of CpG dinucleotides theory.

### 3.3.2 Gene content

Number of annotated genes in sponges ranges from 18906 in *O. pearsei* to 47022 in *E. subterraneus* (figure 10a, bottom). Although the number of genes is not significantly correlated with genome length, there is a significant positive correlation between genome length and total length of genes in the genome, measured Spearman's rank correlation is 0.96 (p value=0.003).

Mean length of all genes found in sponges is 3277 bases and the median is 1566. There is no significant correlation between genome size and gene lengths. Due to a large number of genes, all pairwise comparisons of gene lengths between species show a statistically significant difference (measured by Dunn post hoc test, and after the correction of the p values for multiple testing by Benjamini Hochberg method) except for Amphimedon-Oscarella pair. Although all other comparisons are significant, the most notable difference is visible for *S. ciliatum*, the only representative of the Calcarea class. Median gene length for *S. ciliatum* is 3123 bases (figure 10a, top). *O. pearsei* is the only analysed representative of the Homoscleromorpha class and it has the smallest genome size of all analysed genomes. It also has the lowest number of genes (18906) and the smallest median gene length (1057 bases). The Demospongia class is the largest and most diverse group of sponges, so the observed differences in genome sizes and gene lengths in this group are not surprising. While *A. queenslandica* and *E. subterraneus* show similar median gene length as *O. pearsei* (1008 and 1147, respectively), others are larger (the median for *Tethya wilhelma* is 1660 and for *E. muelleri* 2234). *S. domuncula* is the smallest analysed representative of the Demospongiae group. While it shows a similar number of genes (19195) as *O. pearsei*, the median gene length is the largest in this group (2673 bases).

The genome of *O. pearsei* is the most poorly assembled among all analysed genomes when measured by number of unique and complete BUSCO orthologs, which might impact the quality of annotation. As genome contiguity might influence gene lengths, I tested the significance of correlation between number of scaffolds in the genome and mean and median of gene lengths (Figure 10b).

As expected, there is a negative correlation between gene length and number of scaffolds in the assembly. When all sponges are analysed, this correlation is not significant (p value 0.16 for the mean and 0.2 for the median) measured by spearman correlation coefficient. Interestingly, the genome of *S. ciliatum* acts as an outlier in this data set. When it is excluded, the negative correlation becomes significant (p value 0.016 for the mean length and 0.033 for the median).

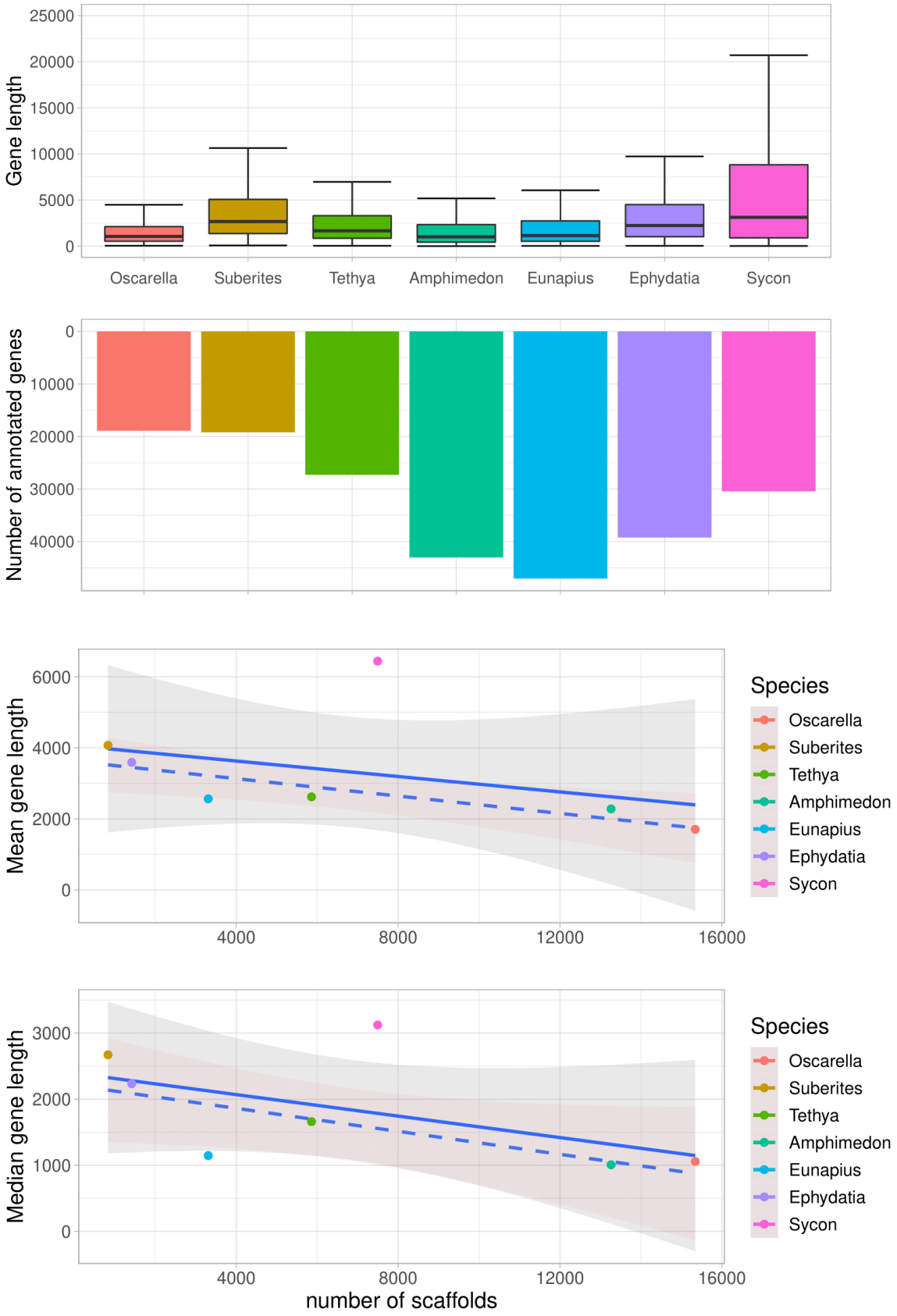



*Figure 10. a) Gene content in the sponge genomes. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. b) Gene length modelled by the number of scaffolds in the assembly. Full blue line shows the linear model when all points are included, while the dashed line shows the model when the outlier Sycon ciliatum is excluded.*

### 3.3.3 Exons, introns and intergenic regions

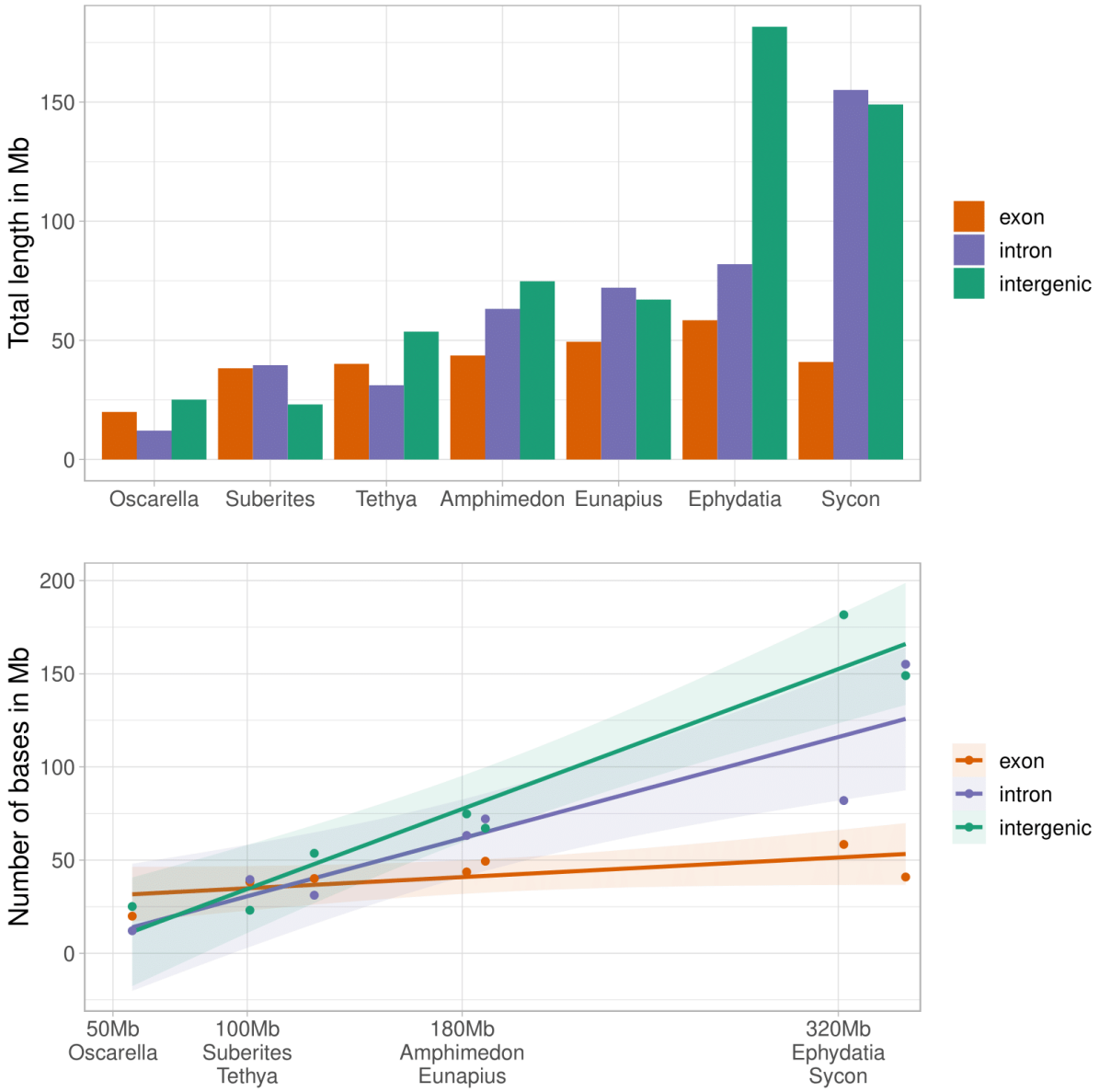



*Figure 11. Abundance of exons, introns and intergenic regions in the genomes. Total length of each region is shown in the top part. Bottom part shows the correlation between genome size and region abundance modelled by a linear model. Pale colored region shows a 95% confidence interval for the coefficient. Mb = mega bases.*

I divided the genomes into exons, introns and intergenic regions and analysed the content of those regions in sponge genomes (Figure 11). The genome of *O. pearsei* shows the lowest number of megabases covered by exons, only 19.9, which is not surprising due to the lowest measured genome completeness. Other genomes show a relatively similar number of bases covered by exons, 38-43Mb in *S. domuncula*, *T. wilhelma*, *A. queenslandica* and *S. ciliatum*, while the most exons are found in the genomes of *E. subterraneus* (49Mb) and *E.*

*muelleri* (58Mb). There is no significant correlation between the number of coding bases (exons) and genome lengths. There is a positive correlation between the total length of both introns (pearson correlation coefficient 0.91, p value 0.005), and intergenic regions (correlation coefficient 0.96, p value 0.0005) with genome size (figure 11, bottom). Notably, the genome of the only analysed Calcareous sponge *S. ciliatum* shows a low intron to exon ratio - only 0.26, while this ratio ranges from 0.7 in *A. queenslandica*, *E. subterraneus* and *E. muelleri*, to 1.6 in *O. pearsei*, (figure XXX, top) .

Next, I analysed the distribution of lengths for exons, introns and intergenic regions (figure 12). Mean exon lengths were in range from 155 bases in *O. pearsei* to 267 bases in *T. wilhelma*. Length of intergenic regions was shortest in *O. pearsei* genome, where it was on average 306 bases long. The largest sponges, *E. muelleri* and *S. ciliatum* not surprisingly have the longest intergenic regions, with the median length of 2.5 Kb and 2.2 Kb, respectively, while the rest have the mean value of around 1.5Kb. Again interestingly, while the median size of introns is smaller than 155 bases in most sponges, *S. ciliatum* has the median of intron lengths of 655 bases. Small rectangle in the figure 12 shows the zoomed-in boxplot for exon and intron sizes.

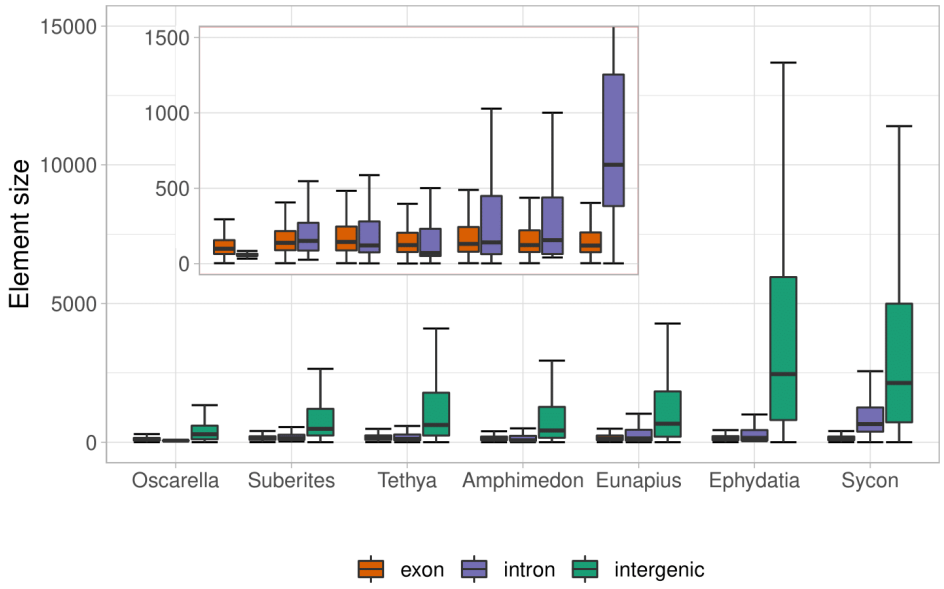



*Figure 12. Distribution of lengths for exons, introns and intergenic regions in the sponge genomes. Exons and introns are shown additionally in the zoomed plot in the top left rectangle. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

Genomes of higher eukaryotes have long first introns compared to other introns. I analysed if this trend is visible in the genomes of sponges by plotting the density of the lengths of first introns and non-first introns (figure 13) on logarithmic scale. The medians of the lengths for the two groups (first intron and non-first intron) were the same in *O. pearsei* (55 bases in

both groups), *S. domuncula* (150 bases), *T. wilhelma* (120 bases), and *A. queenslandica* (70 bases). *E. subteraneus* and *E. muelleri* show a larger median size for non-first introns than first introns (144 vs 133 and 156 vs 151 bases, respectively). Median length for first introns in *S. ciliatum* is 671 bases, which is larger than the median length for non-first introns.

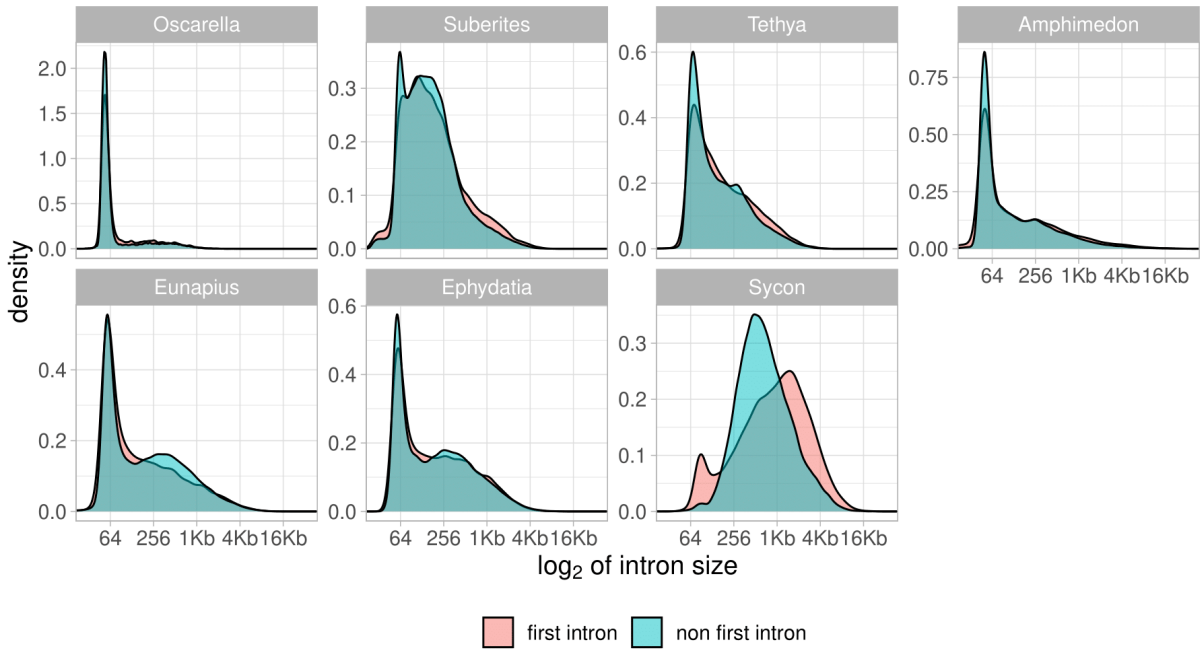



*Figure 13. Distribution of intron lengths separately for first introns versus non first introns shown as a density function.*

## 3.4 Identification of transposable elements

I identified the transposable elements in sponge genomes using two approaches. I first identified them by scanning the genomes against publicly available consensus sequences. To enable the detection of previously unknown sequences, I also identified them using a *de novo* approach. I filtered out the elements which overlap with regions of the genome which were assembled poorly (low quality regions). I annotated the detected elements by comparison with known elements and compared their abundances among sponge genomes.

### 3.4.1 Identification and annotation of potential transposable elements

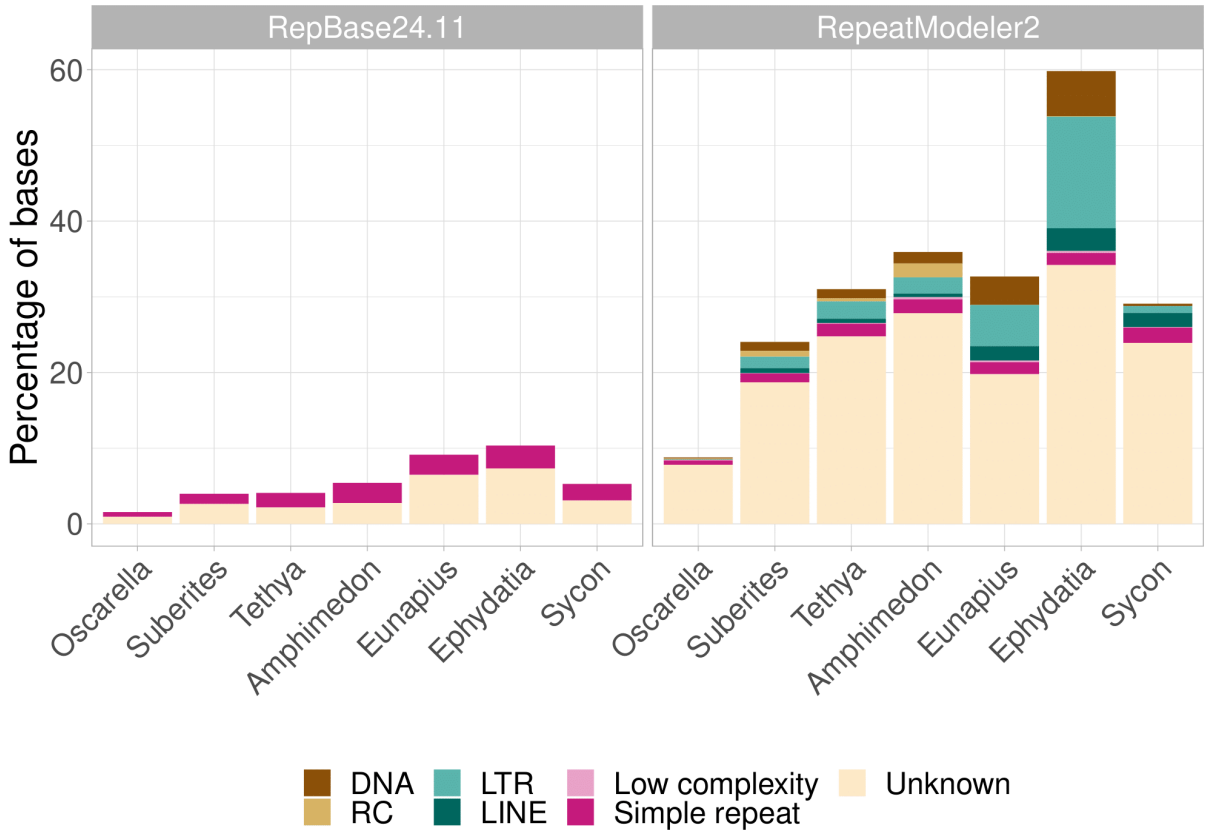



*Figure 14. Distribution of repetitive regions in sponge genomes shown as percentage of total bases in each genome. Left plot shows the identified repetitive regions using RepBase24.11 as a repeat library, while right plot shows the amount of identified repetitive regions when a custom de novo made library is used.*

I identified repetitive regions in the sponge genomes using Repeat Masker. First, I used the repeat consensus library provided in RepBase24.11 to find the repetitive regions (Figure14, left). This approach assigned 6.28% of bases as repeats on average, with the least percent identified in *O. pearsei* (1.28%), and the most in *E. muelleri* (12.06%). I used RepeatModeler2 to define a consensus repeat library de novo for each of the assembled genomes (figure 14, right). Using those libraries instead of RepBase resulted in more identified repetitive regions, with an average of 36.3% over all genomes. I identified the repeats using RepeatClassifier by

comparing them to known repeats in RepBase and Dfam databases. On average, 71% of repeats did not show significant similarities to known repetitive elements and were grouped into "Unknown" group. Number of bases identified in each class is shown in table XXX in the appendix.

### 3.4.2 Filtering of low quality transposable elements

I used the previously defined low quality regions in the assembled genomes to filter out transposable elements of low quality. Top of the figure 15 shows the fraction of bases remaining after filtering in blue. Bottom part of the figure shows percent of all bases in the genome which belong to each repeat group, colored by qualities. Bases marked "low quality" are defined as low quality both by Illumina and nanopore reads mapping to the genome. Bases defined as "ok" are not defined to be low quality by any of the two data sets. In the genome of *E. subterraneus* 81.0% of the transposable elements passed the quality filtering. Out of the transposable elements which did not pass the quality filtering, most failed due to low quality judged by Illumina reads mapping back to the genome (16%). Low assembly quality, judged by Illumina reads mostly affected repetitive elements grouped to low complexity and simple repeats groups, where 49% of low complexity regions and 34.7% of simple repeats were filtered out. The genome of *S. domuncula* shows less low quality transposable elements; 93.7% of elements remained after filtering. As was the case with *E. subterraneus*, most of the filtered elements were filtered due to low quality of assembly judged by Illumina reads, and they account for 5.7% of all transposable elements in this genome. Here, the largest fraction of low quality bases was found in the group of DNA elements, where 8.4% of elements were filtered out. In summary, in the genome of *S. domuncula*, 0.86% of all identified repetitive elements were filtered out of the results due to low quality, and they contributed to 6.3% of the bases defined as repetitive. In the genome of *E. subterraneus* 10.9% of elements contributing to 19% of the total length of repetitive elements, were filtered out from the final list of repetitive elements.

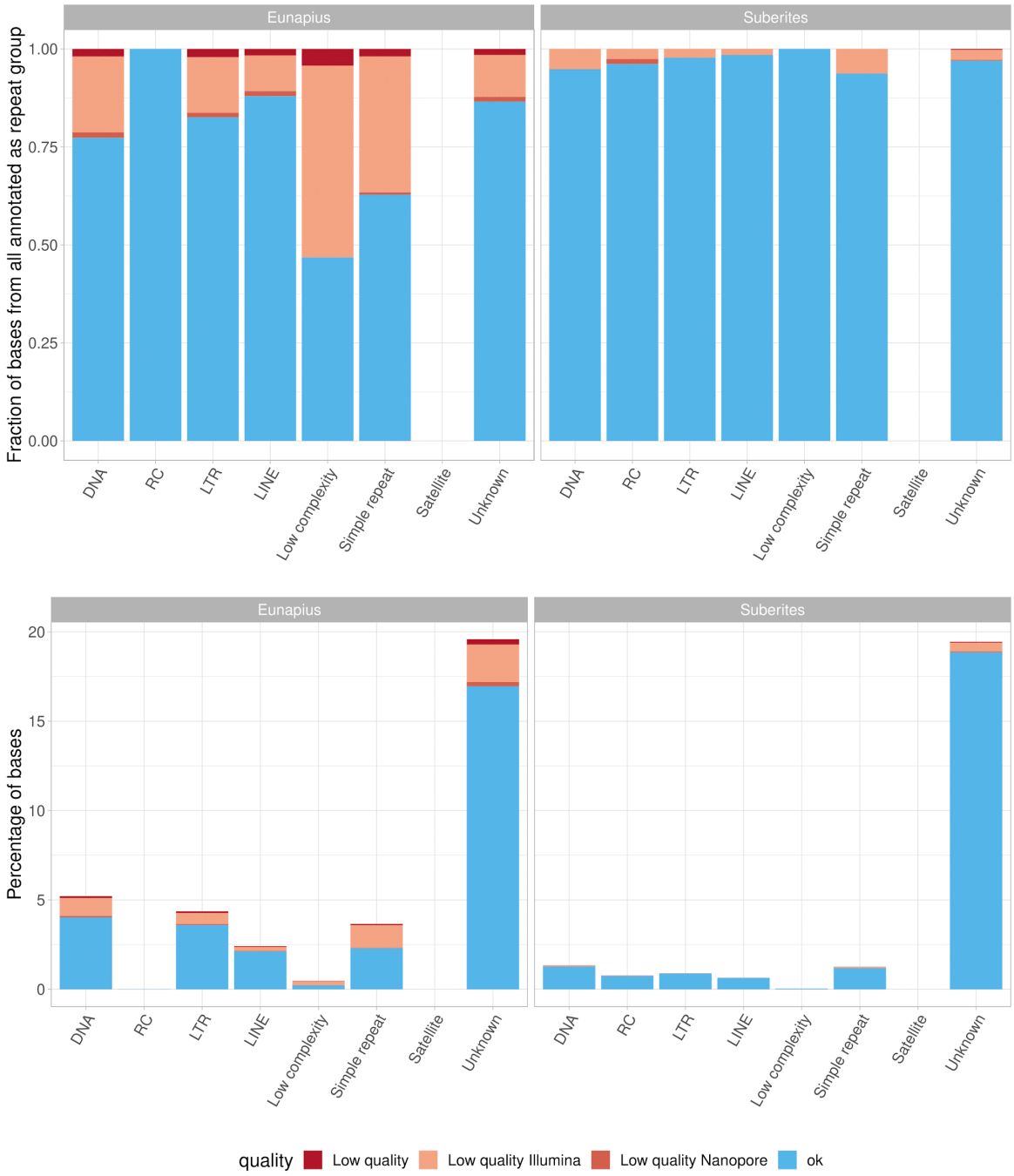



*Figure 15. Results of filtering repetitive regions by quality of assembly for E. subterraneus and S. domuncula. Top plots show a fraction of repetitive bases assigned to each quality category. Bottom plots show percentages of total number of bases in the genomes.*

### 3.4.3 Comparison among sponge genomes

After filtering out low quality transposable elements, the repetitive regions which did show similarities to known repetitive elements were divided into 6 groups. Elements which are grouped as simple repeats and low complexity regions are shown in shades of pink in the figures 16 and 17. Class I elements are grouped into long terminal repeat elements (LTR) and long interspersed nuclear elements (LINE) and are shown in shades of green. Class II elements are

grouped into DNA elements and RC elements which replicate by a rolling-circle mechanism and are shown in shades of brown. Figure 16 shows the relative contribution of different groups of repetitive elements within each sponge species to the total number of bases which belonged to elements with similarities to known repetitive elements. Values are shown in Table 18 in the appendix.

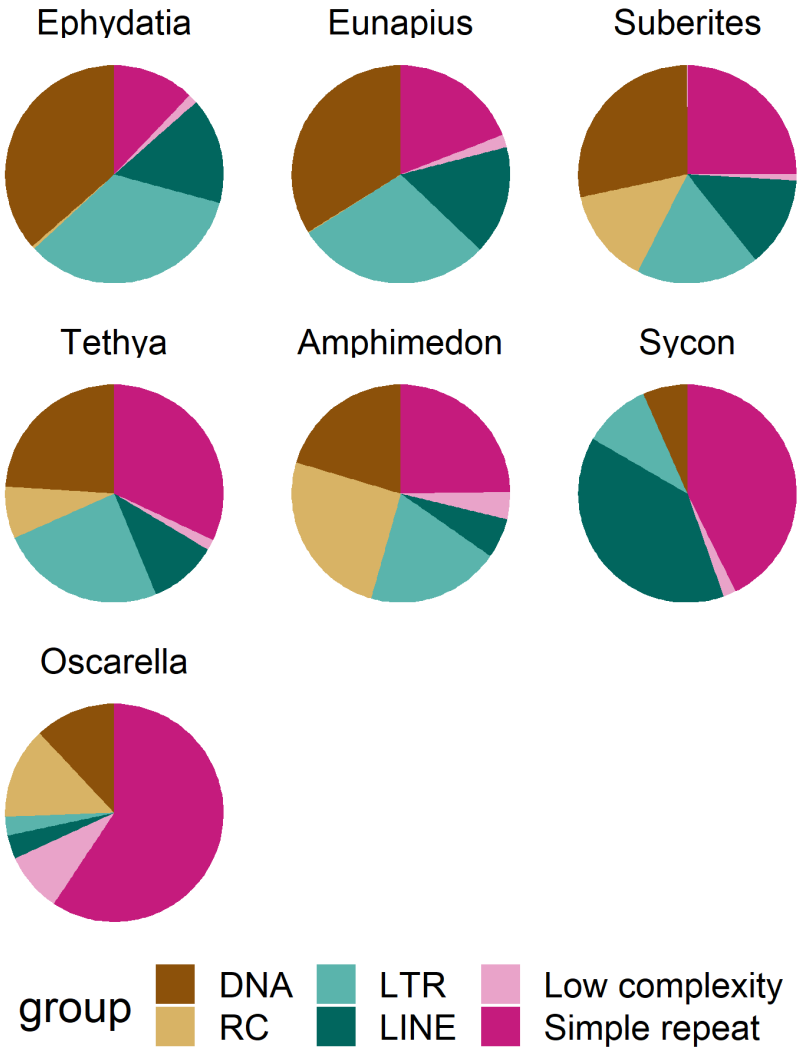



*Figure 16. Proportion of bases from elements with similarities to known elements belonging to different repeat groups.*

Out of elements in the *O. pearsei* genome with similarity to known elements, only 31.9% of all bases are assigned to interspersed repeats, while the majority (59.3%) belong to simple repeats and low complexity regions (8.8%). *S. ciliatum* genome shows a similar trend, where simple repeats make up 42.7% of bases, and the majority of the interspersed elements are LINE elements, contributing to 38.8% of total number of bases. Interspersed elements make up around three quarters of the bases in repeats which could be classified in the genomes of *T. wilhelma* (66.4%), *A. queenslandica* (71.2%), *S. domuncula* (74.1%) and *E. subterraneus*

(79.0%), while in the genome of *E. muelleri* they account for 86.5% of the bases. Interestingly, in the genomes of *E. subterraneus* and *E. muelleri* rolling circle elements were identified but contributed to under one percent of bases, while in *S. ciliatum* they were not found at all. In the remaining genomes they account for up to a quarter of all repetitive bases (*A. queenslandica*, 25.2%). Total number of transposable elements is positively correlated with genome size (95% confidence interval 0.33-0.98, p value=0.01).

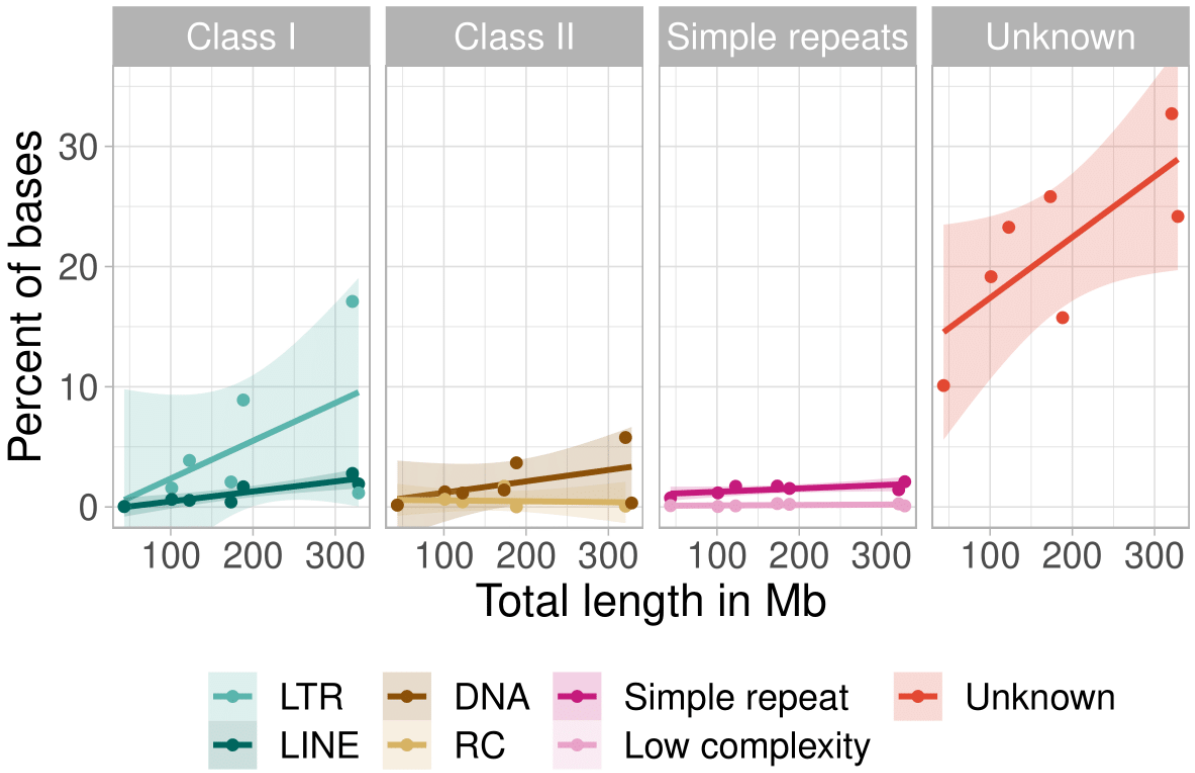



*Figure 17. Percentage of the total number of bases assigned to each of the 6 repeat categories modelled as a linear function of genome length. The colored area around the lines represent 95% confidence interval for the coefficient.*

In general, the number of bases annotated as repetitive elements grows with the size of the genome. Percent of total bases in the genome which is occupied by repetitive elements also grows with increasing genome size. This positive correlation is true for all groups of repetitive elements except for rolling circle elements which comprise 1.65% bases in *A. queenslandica*, but are found in only 0.07% bases in *E. muelleri* and 0.01% bases in *E. subterraneus*, and are not detected in *S. ciliatum* (Figure 17). Although the percentage of other groups of repetitive elements grows with growing genome size, this correlation is only statistically significant for LINE elements (pearson correlation coefficient 0.86, p value 0.01).

## 3.5 Impact of transposable elements on genome evolution

I first analysed the impact of the transposable elements on the evolution of the sponge genomes by assessing the contribution of transposable elements to the genome organization. I divided the genomes into regions containing all exons, all introns and all intergenic regions, and analysed the overlap of the regions and transposable elements.

I assessed the conservation of transposable elements by determining the sequence divergences among elements of the same group of transposable elements compared to the consensus element. I compared the sequence divergences of different groups of transposable elements within each sponge genome, and among all sponges.

Next, I separately analysed the conservation of the LTR elements. This group of LTR was subdivided into "intact", "solo" and "unassigned" subgroups. Intact elements are the elements which show all the structural characteristics expected in a full length LTR element. Solo elements are LTR elements which appear as a solitary intact long terminal repeat (see methods for details). Unassigned elements are defined as LTR elements which are neither intact or solo. I analysed the contribution of the defined subgroups to genome organization, and their sequence divergences across different genomes.

To determine if there is a correlation between the insertions of transposable elements and gene expression, I analysed gene expression in the sponge *E. subterraneus* in the first and tenth day of primmorphs formation. First I analysed the expression of the genes encoded by the transposable elements and compared it with the expression of genes not encoded by transposable elements. This analysis was also done for each group of transposable elements. I further analysed the expression of genes in which there was an insertion of transposable elements by analysing their expression depending on the location of the insertion (if the element was inserted to exons or introns), and the group of transposable elements. Lastly, I analysed the expression of genes which harboured an insertion of LTR elements and compared it based on the annotation of LTR element and location of the insertion.

Finally, I identified the homologs of the proteins involved in the piRNA pathway which serves as a defence against transposable elements, in all sponge genomes. I also analysed the expression of the identified homologs during the formation of primmorphs in *E. subterraneus.*

### 3.5.1 Contribution to genome organisation

I analysed the contribution of transposable elements to introns, exons and intergenic regions. Figure 18 shows the percentage of bases annotated as transposable elements either assigned to class I, class II or annotated as unknown elements, separately for exons, introns and intergenic regions.

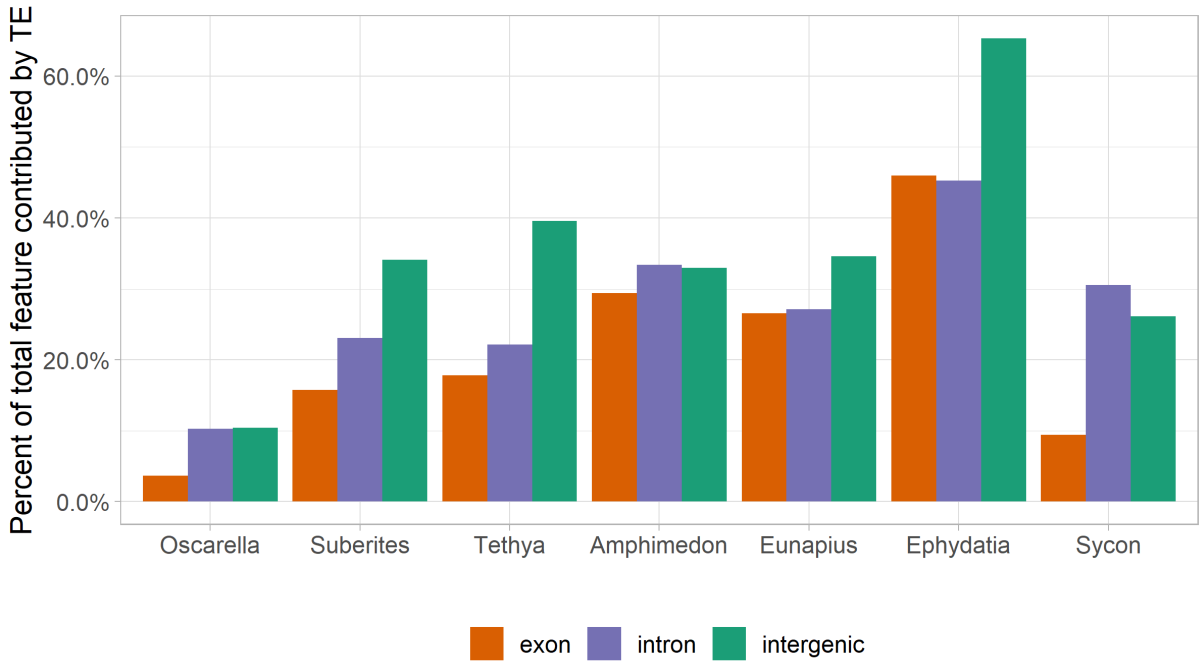



*Figure 18. Percentage of total number of bases assigned to exons, introns and intergenic regions which are covered in transposable elements.*

The contribution of transposable elements to total exon length in genomes ranges from 3.6% in *O. pearsei* and 9.5% in *S. ciliatum* to 46.0% in *E. muelleri*. In other sponges belonging to the Demospongia group, transposable elements contribute to exons from 15%-29%.

Same trend is observed when analysing the percentage of introns annotated as transposable elements. This percentage is the lowest in *O. pearsei* (10.3%), and highest in *E. muelleri* (45.24%). Again other Demospongiae sponges show similar TE content in introns, ranging from 22% - 27%, while 30.5% of total intron length in *S. ciliatum* is annotated as a transposable element. Of note, the Demospongia group of sponges reveals a similar contribution of TE to both exons and introns - ratio of percentage of TE contribution to introns over percentage of TE contribution to exons ranges from 1 to 1.4. This is not the case for the representatives of other groups, where TEs contribute by a higher percentage to introns. Intron to exon contribution ratio in *O. pearsei* is 2.8 and in S. ciliatum 3.2.

Finally, transposable elements comprise 10.4% of intergenic regions in *O. pearsei* and 26.1% in *S.ciliatum*. In the genomes of Demospongiae *A. queenslandica, S. domuncula, E. subterraneus* and *T. wilhelma* they contribute to 32-39% of the total intergenic region length, while in the genome of *E. muelleri* they comprise 65.3% of all intergenic regions.

To determine if there is a potential preference of transposable elements to certain region type, I analysed abundances of transposable elements in introns, exons and intergenic regions (Figure 19). Since those regions are not distributed uniformly within a genome or between different genomes (for example *E. muelleri* has over two times larger intergenic regions compared to other sponges, see Figure 11), we do not expect a uniform distribution of transposable elements neither within a single genome, nor between different genomes. For this reason I calculated the number of bases that we can expect to be assigned to transposable elements based on the assumption that the likelihood of discovering transposable elements in any group depends only on the size of the group in the genome (in other words, that there is an equal probability that transposable element is found in exon, intron and intergenic region ). Figure 19 shows the distribution of all transposable elements in the mentioned regions in the genomes as the number of bases assigned to transposable elements in each of the groups. Black dotted line represents the expected number of bases.

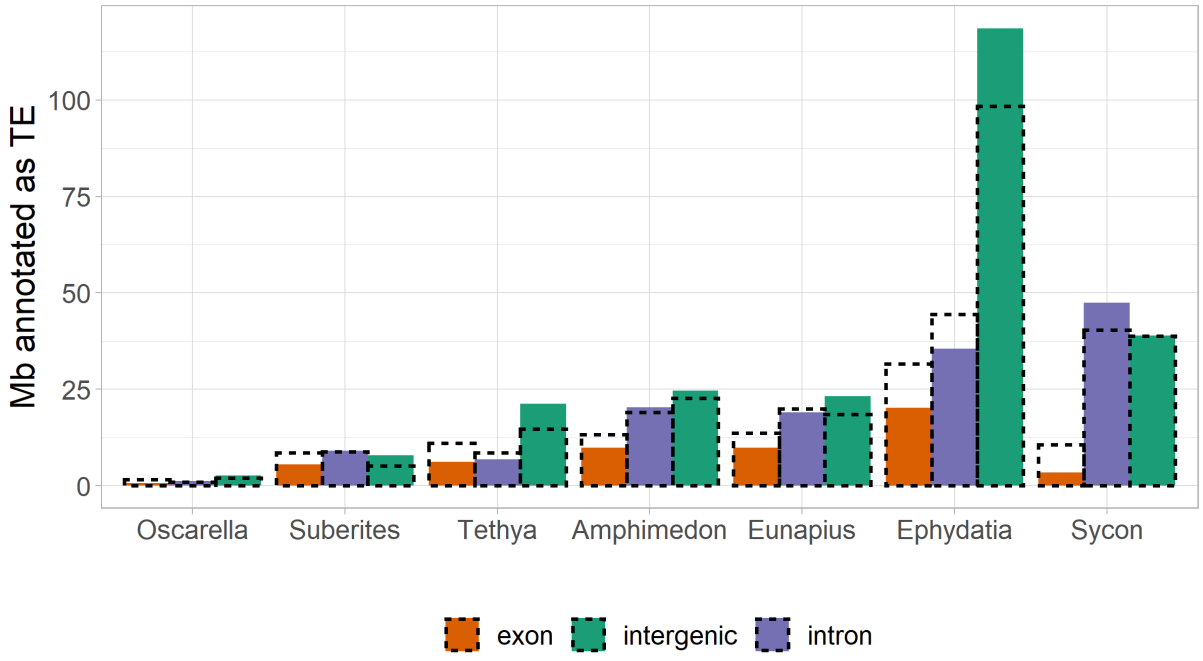



*Figure 19. Number of bases annotated as transposable elements in introns, exons and intergenic regions. Black dashed line represents the expected number of bases. Mb=mega bases.*

The observed and expected values compared by chi-square test differ significantly for species *E. muelleri* (p value=0.006) and *S. ciliatum* (p value=0.047). In the genome of *E. muelleri* the calculated expected number of bases annotated as transposable elements in exons

is 13.6 Mb and observed is 9.8 Mb. Expected number in introns is 19.9 Mb which is similar to observed 19 Mb. By assuming a fair distribution of transposable elements we would expect 18.5 Mb of transposable elements in intergenic regions, and we find 23.2 Mb. For *S. ciliatum* the expected and observed number of bases annotated as TE align for intergenic regions. However, the expected value for exon is 10.6 Mb and the observed is 3.4 Mb, while the trend for introns is the opposite - we would expect 40 Mb of introns to be covered by TEs and in the genome of *S. ciliatum* we find 47 Mb.

I calculated the enrichment of observed value over expected value as percent of the difference between observed and expected value compared to the expected value (Figure 20). Top part of Figure 20 shows the percent enrichment when all transposable elements are analysed. Values under 0 represent depletion and values above 0 represent enrichment. There is a clear and consistent trend of depletion of transposable elements in the coding regions of the genome in all species. Conversely, the trend of enrichment of transposable elements in the intergenic regions is also consistent between species. Perhaps surprisingly, while most species show a small change or depletion of transposable elements in introns, *S. ciliatum* shows enrichment. Note that enrichment and depletion values observed in *O. pearsei* are unreliable due to the low number of bases assigned to each category.

To get a better understanding on which groups of transposable elements might contribute the most to the observed differences in abundance between the regions in the genome, I divided the transposable elements into groups as defined before. DNA transposons, transposable elements which replicate via rolling circle (RC), long terminal repeats containing TEs (LTR) and long interspersed nuclear elements (LINE). All the transposable elements which could not be recognised as any of those groups, and were not recognised as simple repeats or low complexity regions, were grouped together into “Unknown” group. I calculated the enrichment of transposable elements abundances in exons, introns and intergenic regions for the defined groups of transposable elements separately (Figure 20, bottom).

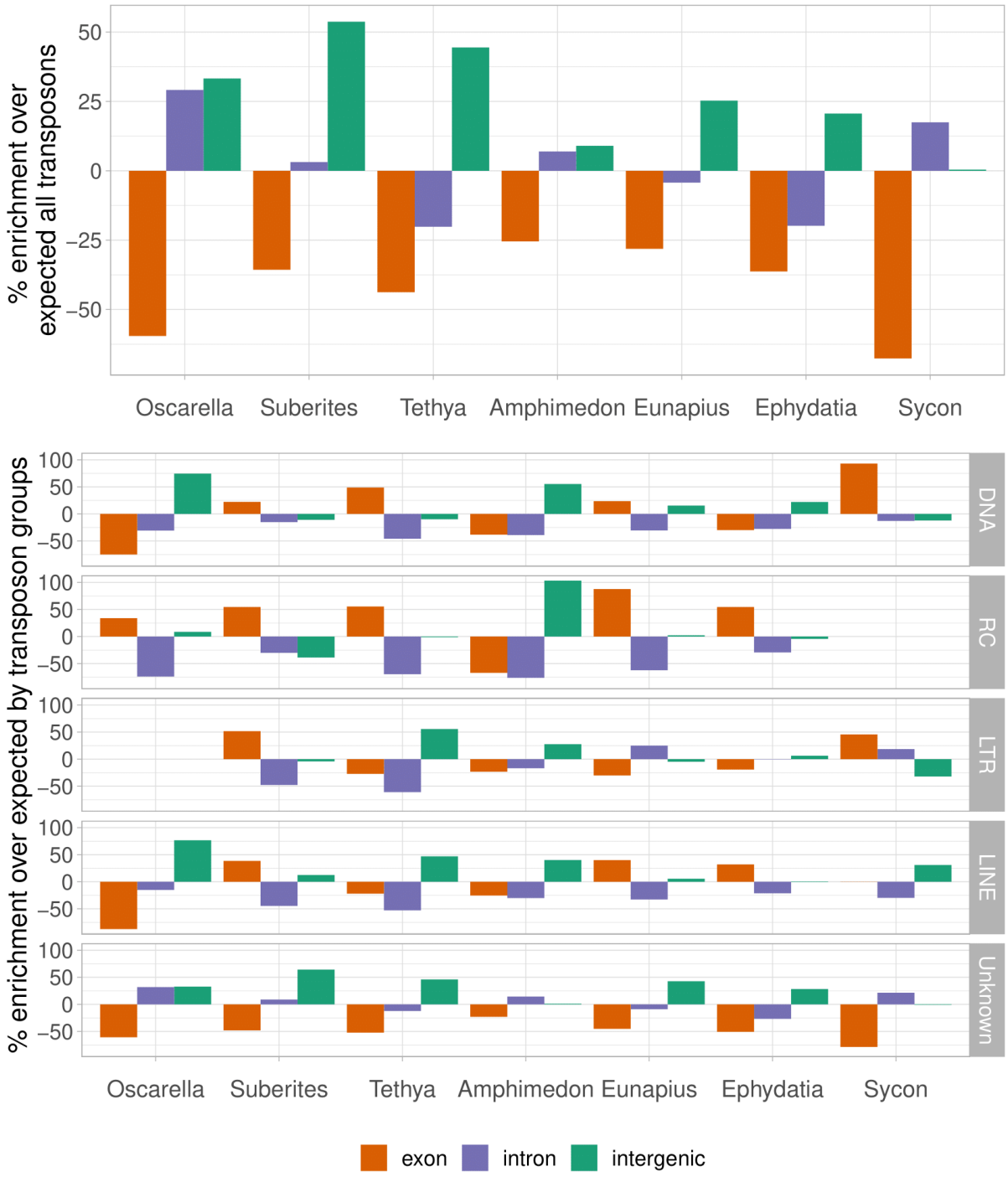
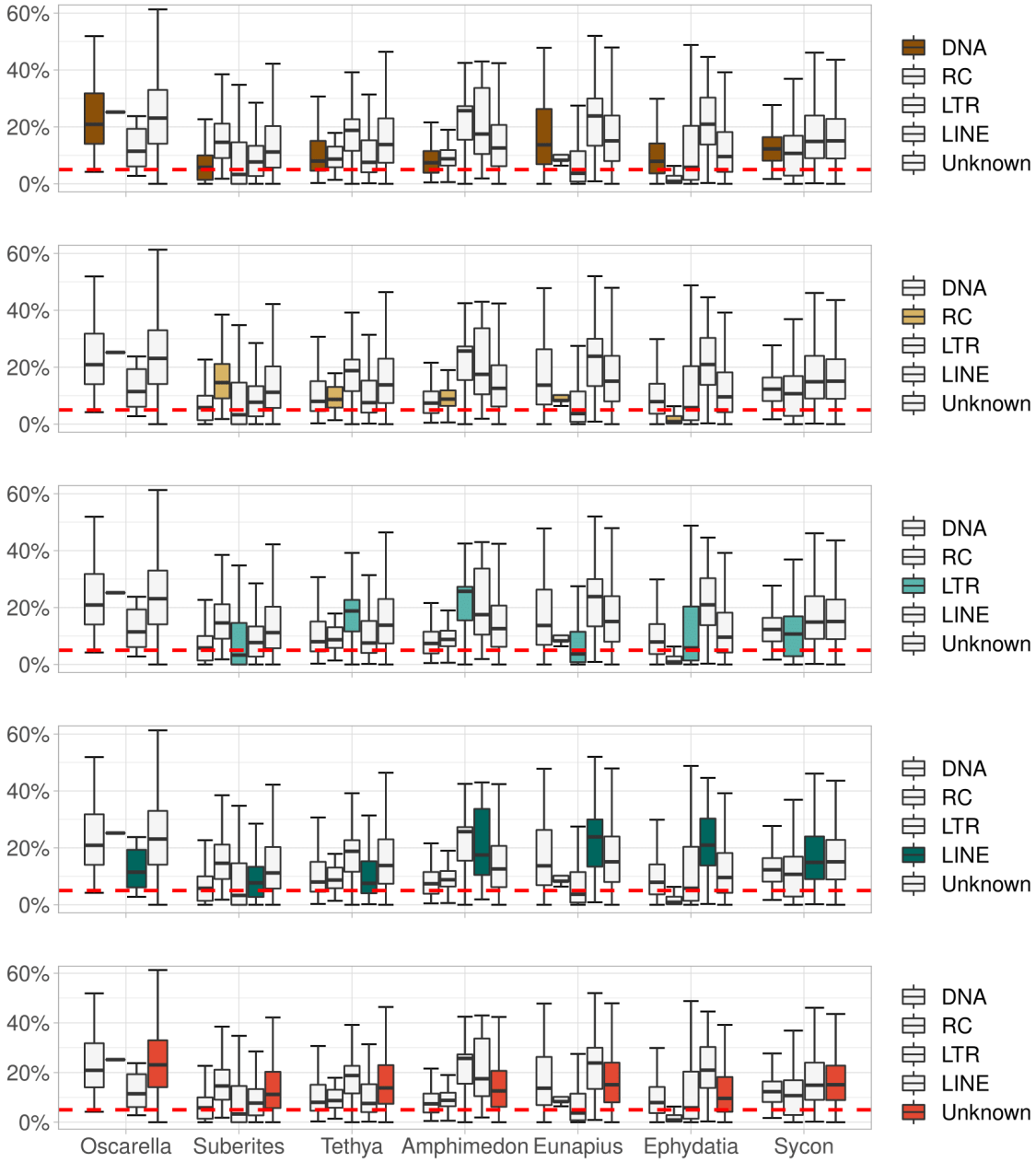



*Figure 20. Percent enrichment of the observed number of bases assigned to transposable elements over the expected number of bases, by exons, introns and intergenic regions. Top figure shows all transposable elements together. Bottom figure shows the enrichment for each group of transposable elements separately.*

There is a consistent depletion of transposable elements in introns for LINEs, RC and DNA elements, while for LTR elements there is a slight enrichment in introns compared to expected values in *E. subterraneus* and *S. ciliatum*. I found a consistent depletion of transposable elements of unknown types in exons across all species. There was no consistency in this depletion for other element groups. While a large majority of transposable element

groups show an enrichment in intergenic regions, and depletion in coding regions, this trend is inverted for LTR elements in *S. ciliatum* and RC elements in *S. domuncula*.

### 3.5.2 Sequence divergences




4

*Figure 21. Total sequence divergences (substitution + deletion + insertion) of the transposable elements identified with as over 90% of the consensus repeat length. Rates are calculated by comparison with the consensus repeat for each repeat type. Red dashed line represents 5%. Middle line in the box plot represents the median, box represent the interquartile range and the whiskers extend to IQR±1.5*IQR.*

I analysed the conservation of transposable elements by observing their sequence divergences compared to the consensus element. The consensus element for each type of transposable elements represents the most likely ancestor element from which all other copies

originated in a genome. Differences from consensus arise from the mutations which accumulate over time, and thus the element type which integrated into the genome most recently will have the least mutations compared to the consensus. Figure 21 shows the total sequence divergences per repeat group and species. DNA elements have the lowest sequence divergence out of all elements in the genome of *A. queenslandica*. Rolling circle replication elements, although sparse, have a low sequence divergence in *E. muelleri*, followed by LTR elements. The genomes of *S. domuncula* and *E. subterraneus* show the largest conservation of LTR elements, where their total sequence divergence is under 5%. LTRs are also the most conserved group of transposons in the genome of *S. ciliatum*. Finally, LINE elements show least deviation from the consensus in the genomes of *O. pearsei* and *T. wilhelma*.

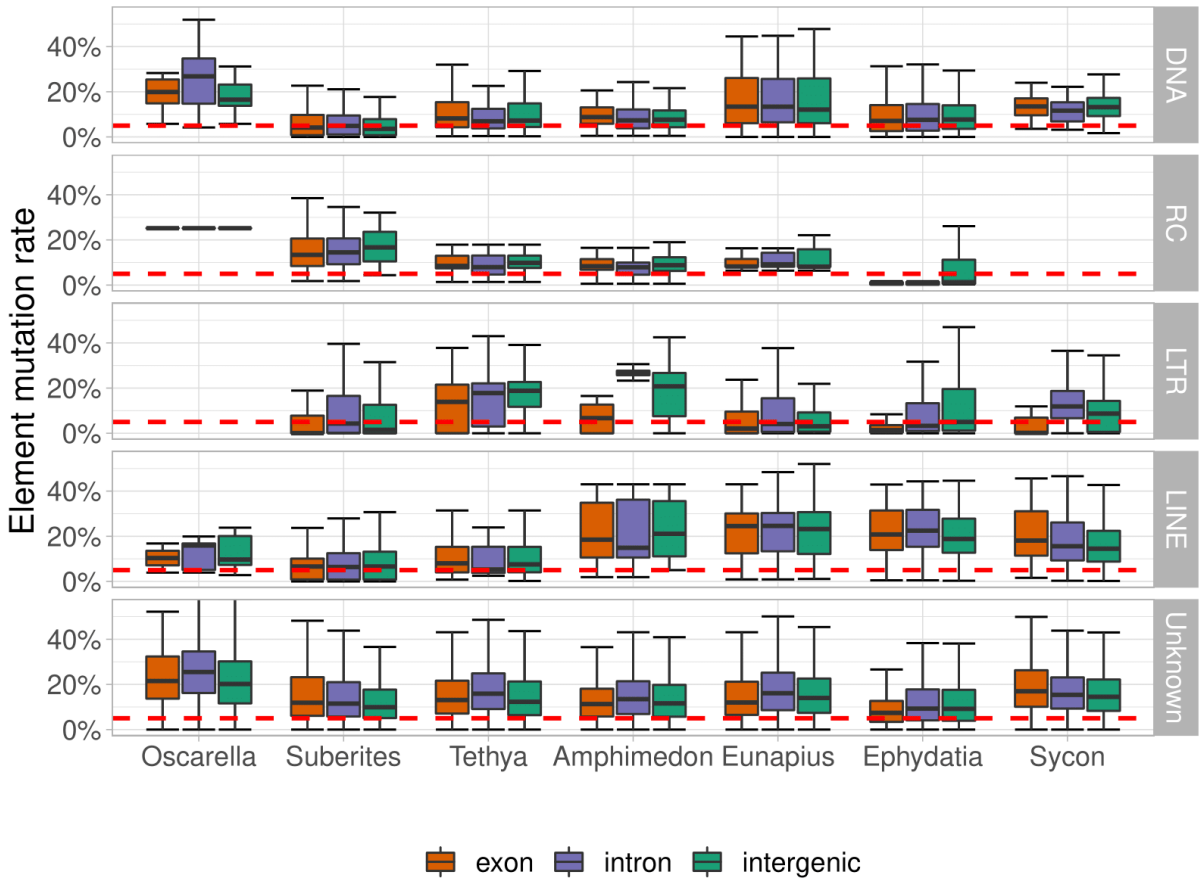



*Figure 22. Sequence divergences of the elements which are at least 90% of the consensus length, grouped by repeat types and region type. Red dashed line represents 5%. Middle line in the box plot represents the median, box represent the inter quartile range and the whiskers extend to IQR±1.5*IQR.*

To determine if there is a difference between conservation of elements with respect to their integration site (exon, intron or intergenic region), I analysed the sequence divergences of elements and compared them across regions (Figure 22). Again, due to a large number of elements, there was a statistically significant difference in sequence divergences between the

regions for the elements of unknown class for all sponges measured by ANOVA test. All sponges in which LTR elements were detected (all except for *O. pearsei*) also showed statistically significant differences in the sequence divergences for LTR elements among the regions. The LTR elements consistently show lower sequence divergences when they overlapped with the coding region. For *S. ciliatum* and *A. queenslandica*, the sequence divergences of LTR elements in the introns were higher than for intergenic LTRs, while the other species showed reverse trend. Interestingly, in *S. ciliatum* LTR elements in exons show the smallest sequence divergences compared to LTR elements in other regions, while in all other groups this trend was not observed and was even reversed - the elements overlapping the exons show even higher sequence divergences than other elements. This anomaly for LTR elements was observed for other genomes as well but it was not so pronounced. There was a significant difference for the sequence divergences of LINE elements in *S. ciliatum* and *E. muelleri*. All p values are shown in a table 19 in the appendix.

### 3.5.3 Conservation of LTR elements

LTR elements are integrated into the genome as full length elements, containing two long terminal repeats and an internal sequence. After the integration, the internal region of many elements is excised due to the homologous recombination of the long terminal repeats, leaving only one full length “solo” long terminal repeat.

I separately analysed the conservation of LTR elements with respect to their structural characteristics, region of integration and species. The results are shown on the figure 23.

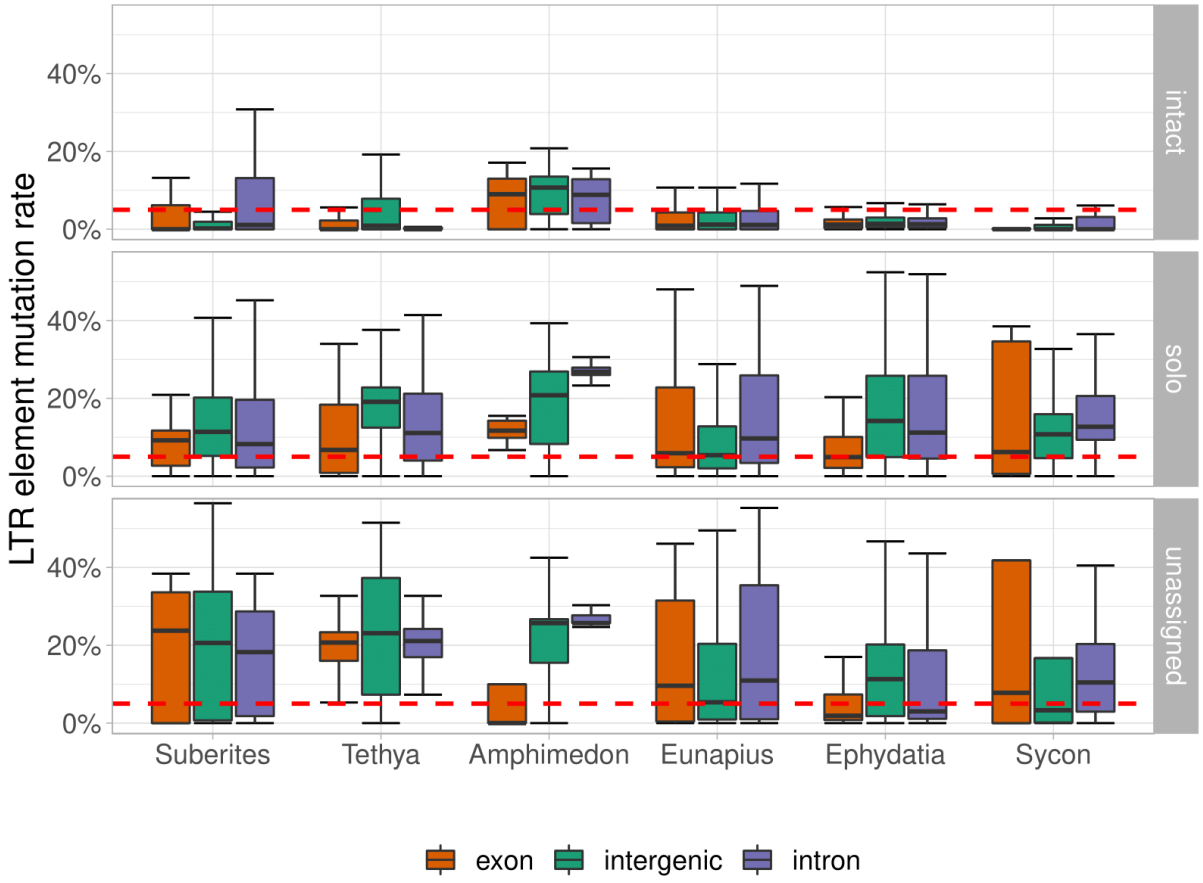



*Figure 23. Sequence divergences of the LTR elements which are at least 90% of the consensus length, grouped by region type and element conservation. Red dashed line represents 5%. Middle line in the box plot represents the median, box represent the inter quartile range and the whiskers extend to IQR±1.5*IQR.*

Intact elements show the lowest sequence divergences in all species, and this sequence divergence is not biased by region. Solitary LTR elements show higher sequence divergences then the intact elements. They are consistently best preserved in exons among most species. The sequence divergences of the elements which were not assigned to intact or solo groups were variable among genomes and regions. Interestingly, unassigned elements in *A. queenslandica* and *E. muelleri* show the highest conservation in exons, while this high conservation (under 5% sequence divergence) is observed in introns of *S. ciliatum*.

### 3.5.4 Analysis of the impact of transposable elements on gene expression

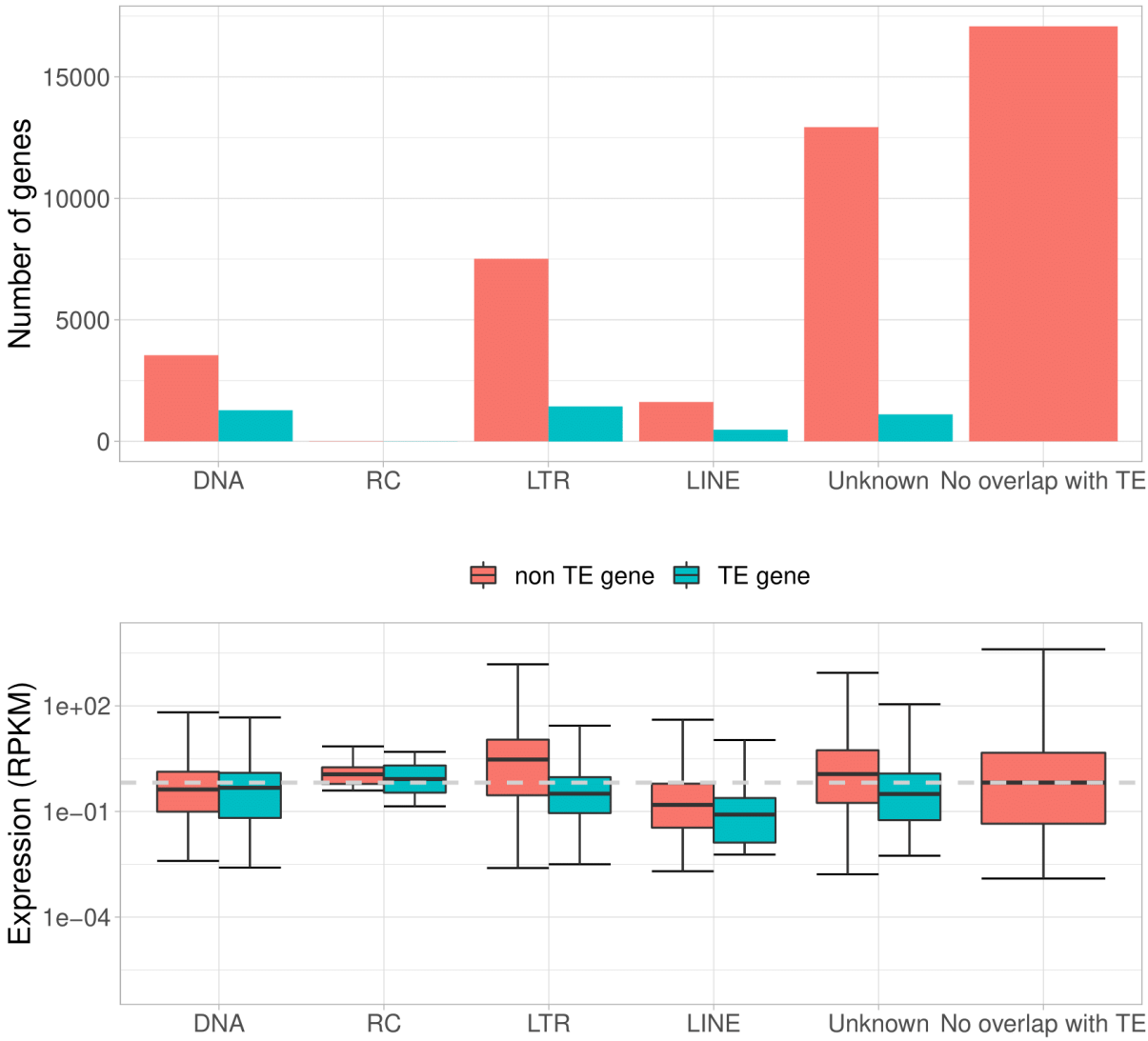



*Figure 24. The amount and expression of the E. subterraneus genes encoded by transposable elements (TE gene), and other genes (non-TE gene) grouped by overlaps with transposable element groups. Top plot shows the number of genes in each category. Bottom plot shows the distribution of the expression levels for each group of genes, measured by RNAseq. The dashed line represents the median value of expression for genes which are not encoded by TE. Hinges of the box plots represent inter quartile range of values, and the notches extend to IQR±1.5*IQR. RPKM = reads per million reads per kilobase of coding sequence length.*

To determine if there is a correlation between the insertions of transposable elements and gene expression I analysed gene expression in the sponge *E. subterraneus* in the first and tenth day of primmorphs formation.

First I analysed the expression of the genes encoded by the transposable elements and compared it with the expression of genes not encoded by transposable elements (Figures 24 and 25). The expression of the genes encoded by transposable elements was consistently lower than

the expression of genes which were not encoded by transposable elements, but overlapped with transposable elements of the observed group, which was statistically significant measured by t-test for all groups except for the RC elements. This difference is consistent over both data sets, shown in Figure 25.

There is an obvious difference between the expression of elements which have no integration of transposable elements and of those elements which have an integration of LINE elements. The expression is always lower for genes which overlap a LINE element, regardless of the origin of the gene (whether it is encoded by the LINE element or no). This difference is statistically significant, measured by two sided t-test on the logarithmically transformed expression values to stabilize the variance of the residuals and remove skewness. The p value was 6.7e-7 and the 95% confidence interval for the mean from 0.16 to 0.36 when comparing the expression of the group of genes which are not encoded by TE but have a LINE integration against the expression of genes which do not have any TE integration.

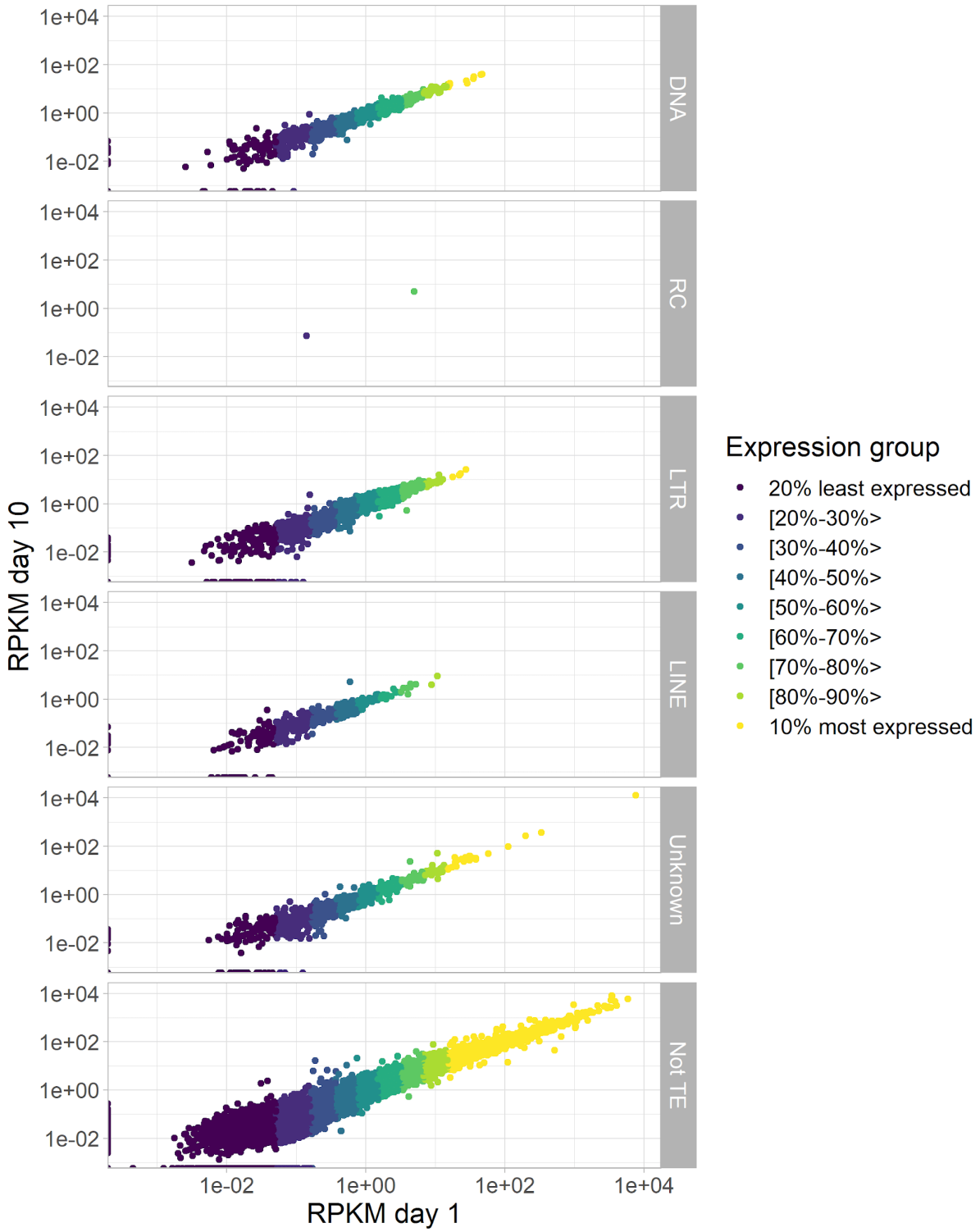



*Figure 25. The expression of the E. subterraneus genes in day 1 and day 10 of primmorphs formation. Points represent genes encoded by transposable elements of different groups, and other genes (non TE). Color represents expression level, darker is less expressed and lighter is more expressed. RPKM = reads per million reads per kilobase of coding sequence length.*

I further analysed the expression of genes in which there was an insertion of transposable elements by analysing their expression depending on the location of the insertion (if the element was inserted to exons or introns), and the group of transposable elements. The results were consistent over all groups of transposable elements and show that the expression of genes for which the integration of elements occurs predominantly in introns is higher than the gene expression for genes in which the transposable elements are predominantly inserted into exons (Figure 26).

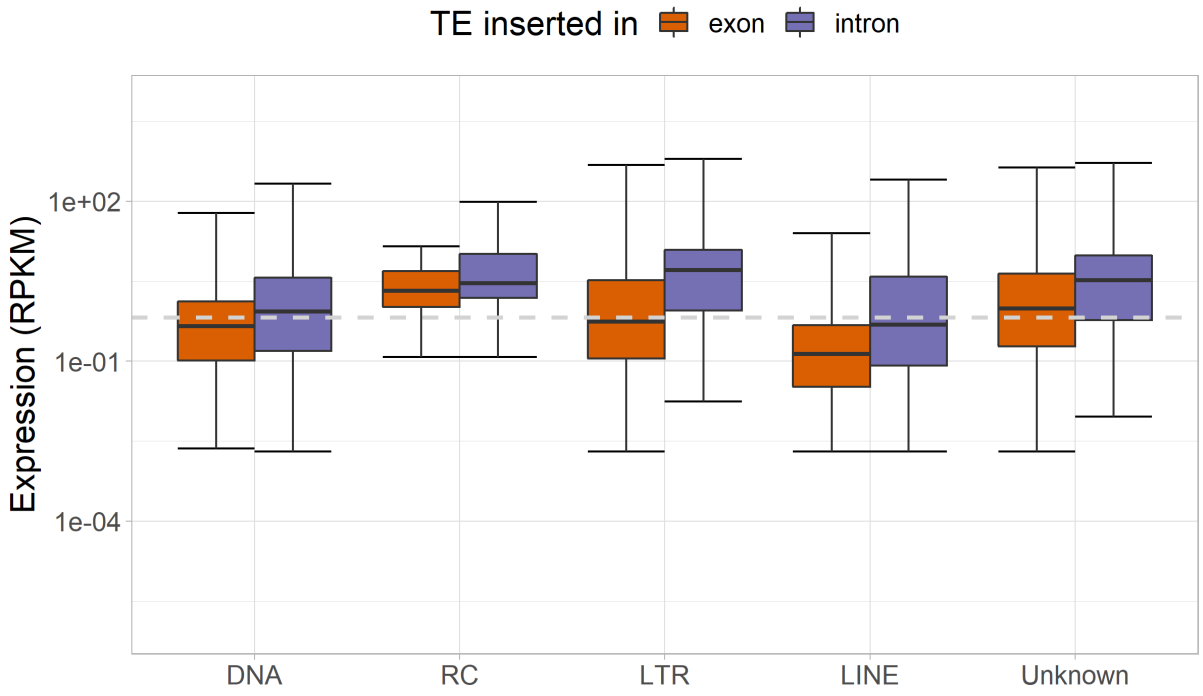



*Figure 26. Distribution of the expression of genes with at least one integration of a transposable element in E. subterraneus and grouped by type of the inserted element. The distribution is shown separately for genes in which the integrations predominantly occur in exons versus genes with predominant intronic integrations. The dashed line represents the median value of expression for genes which are not encoded by TE. Hinges of the box plots represent inter quartile range of values, and the notches extend to IQR±1.5*IQR. RPKM = reads per million reads per kilobase of coding sequence length.*

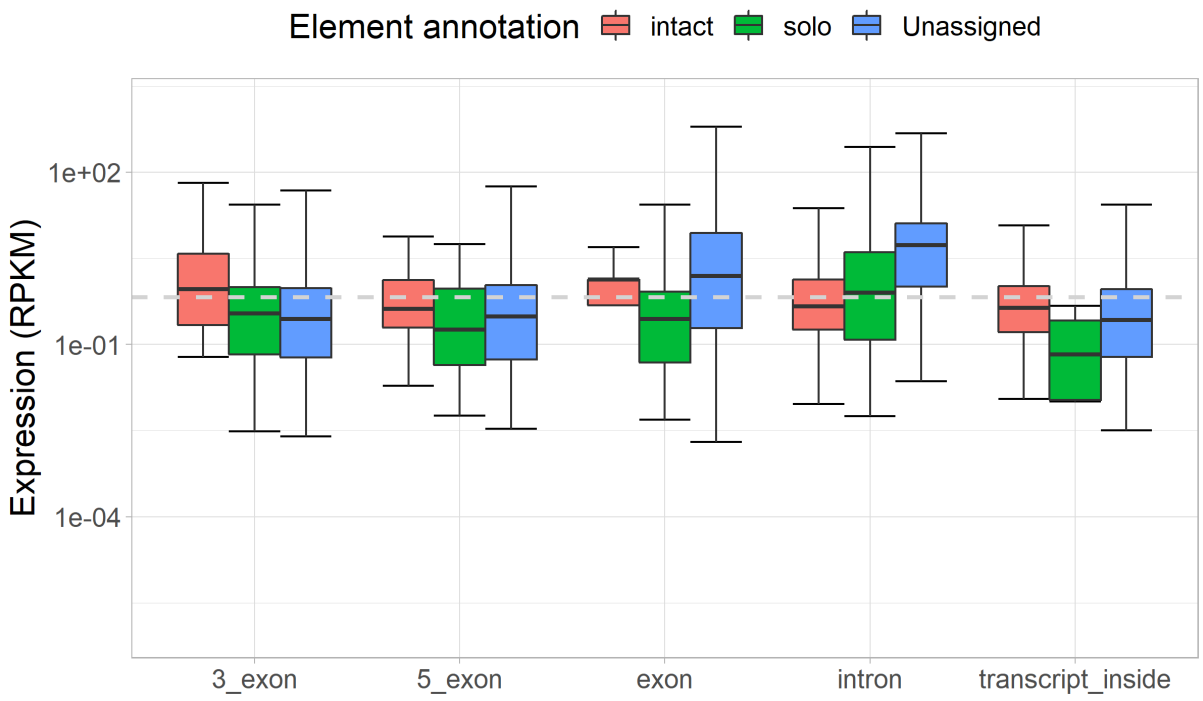



*Figure 27. Distribution of the expression of genes in E. subterraneus which overlap LTR elements. Genes are grouped based on the type of LTR element (intact, solo or unassigned), and its contribution to gene (3' exon, 5' exon, internal exon, intron). Genes which are encoded by the LTR elements are grouped separately in the last group (transcript_inside). The dashed line represents median expression of the genes which are not encoded by TE. Hinges of the box plots represent inter quartile range of values, and the notches extend to IQR±1.5*IQR. RPKM=reads per kilobase of coding sequence length, per million reads.*

Lastly, I analysed the expression of genes which harboured an insertion of LTR elements and compared it based on the annotation of LTR element and location of the insertion (Figure 27). The expression of genes which harboured an intact element was consistently higher if the element overlapped an exon then the genes in which the exon was originating from a solitary or unassigned element in the cases when the exon was first or last exon of the gene. Highest expression rates for genes in which the integration overlapped the intron were observed for unassigned elements. Also, if the exon originating from an LTR element was in the middle of the gene, genes with LTRs of unassigned type had the highest expression compared to genes where the internal exon originated from a solitary or intact LTR element. Finally, the expression of genes in cases when the gene was encoded by the LTR element was highest if the LTR element encoding the gene was an intact element.

### 3.5.5 Catalog of the piRNA pathway components in sponges

I searched for the homologs of the human piRNA pathway components in sponges. The potential homologs were found if the proteins were best reciprocal hit to each other when comparing the full protein set in humans and each sponge. Potential homologs were manually checked for the presence of the conserved domains (see Methods for details). If one of the domains was completely missing, the homolog was discarded. If only one domain was partial and others were found, the homolog was marked as "partial". The characteristic representation of the domains in the homolog for each gene is presented in the appendix. Table 13 shows the number of homologs found in each of the sponge species.

*Table 13. Homologs of the human piRNA pathway found in sponges.*protein is found split in two parts in consecutive genes*

| | | **Amphimedon** | **Ephydatia** | **Eunapius** | **Oscarella** | **Suberites** | **Sycon** | **Tethya** |
|---|---|---|---|---|---|---|---|---|
| **Effectors** | **PIWIL1** | **2** | **2** | **2** | **only piwi** | **2** | **2** | **1** |
| | **PIWIL2** | **0** | **0** | **0** | **1** | **0** | **0** | **0** |
| | **PIWIL3** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| | **PIWIL4** | **0** | **0** | **0** | **0** | **0** | **only paz** | **0** |
| **piRNA biogenesis factors** | **DDX4** | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| | **HENMT1** | 3 | 1 | 1 | 1 | 1 | 1 | 1 |
| | **KIF17** | 1 | 1 | 1 | 1 | 2 | 1 | 1 |
| | **MAEL** | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| | **MOV10L1** | 1 | 1 | 3 | 2 | 1 | 0 | 1 |
| | **PLD6** | 1 | 0 | 1 | 1 | 1 | 2 | 1 |
| | **PRMT5** | 1 | 2 | 1* | 1 | 1* | partial | 0 |
| | **RNF17** | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| | **TDRD1** | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| | **TDRD6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | **TDRD9** | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| | **TDRD12** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | **TDRKH** | 0 | 3 | 1 | 0 | 1 | 1 | 1 |
| | **WDR77** | 1 | 1 | 1 | 0 | 1 | 1 | 0 |

In all but one species, two distinct homologs of the PIWIL1 gene were found. Only exception was *O. pearsei*, where PIWIL1 was found only partially conserved (only piwi domain), but a related homolog PIWIL2 was found in full length. Homolog for gene PIWIL4 was only found as potential hit in *Sycon ciliatum* where only paz domain was found conserved. The other sponges did not show any gene which was identifiable as homolog for neither PIWIL2, PIWIL3 nor PIWIL4.

The homologs of the genes HENMT1, KIF17 are found in all sponge species, while the homologs of DDX4, MAEL, MOV10L1, PRMT5, PLD6 and PIWIL1 are missing only from one of the species. The homolog for DDX4 is found in *E. muelleri*, but only contains a single DEAD-like helicase domain. Homolog of the protein PRMT5 in *S. domuncula* and *E. subterraneus* was split into 2 different genes interrupted by a transposable element encoded gene. The presence of the full length protein is confirmed by the transcriptome data for *E. subterraneu*s, and the split in gene is most likely a problem with the genome annotation, so the genes are included as homologs.

The gene MOV10L1 in the human contains two conserved DEAD like helicase domains. Sponge homologs for this gene were found in all sponge species, with only one exception - *S. ciliatum*, where one domain was only partially conserved. It is worth mentioning that there were potential homologs for this gene found in the species *E. muelleri, E. subterraneus* and *S. domuncula* which were discarded because they contained only one DEAD-like domain. However, all of them also contained another conserved domain, AAA. Those variants were present in multiple copies in the mentioned species, more specifically, seven in *E. muelleri*, four in *S. domuncula* and one full and one partial in *E. subterraneus*. Another interesting difference is that the homologs of the TDRKH gene in sponges seem to have 3 conserved KH-1 domains whereas the human protein only has 2. Also, the TDRD1 and TDRD9 sponge homologs display a various number of conserved TUDOR domains.

To check if the piRNA pathway is active during the formation of primmorphs, I analysed the expression of the identified homologs in *E. subterraneus*. Figure 28 shows the expression of all genes in *E. subterraneus* on the first and tenth day of primmorph formation measured by the amount of mRNA in the sample. Genes are colored from darker to lighter color depending on their level of expression. All the genes involved in piRNA pathway have a level of expression in the top 40% of all expressed genes. There are two different PIWIL1 homologs in *E. subterraneus* and both are very highly expressed. One is in top 10% of all expressed genes, and the other one is in top 20%, in both samples. The expression of DDX4 gene and TDRKH

is also notable, with DDX4 being on the edge of the 90th percentile of expression, and TDRKH being in the top 20%.

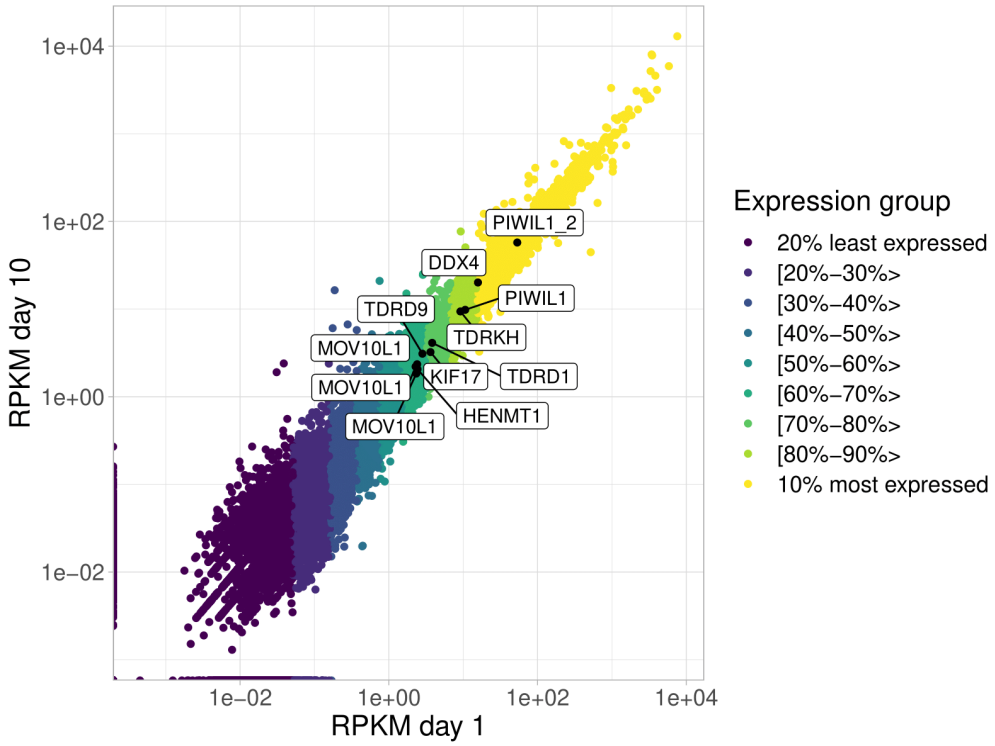



*Figure 28. Expression of the genes in E. subterraneus on the first (x axis) and tenth (y axis) day of primmorphs formation. Homologs involved in the piRNA pathway are labeled. Genes are colored from darker to lighter color depending on their level of expression. RPKM= reads per kilobase of coding sequence per million reads.*

Finally, I was interested to see if there is a difference in the levels of expression of the homologs during the development of the sponge. I analysed RNAseq data available for 4 different embryonic, 4 larval, a juvenile and an adult stage of the development of sponge *A. queenslandica*. Figure 29 shows that the levels of expression (measured by the amount of mRNA) of the genes DDX4 and two PIWIL1 homologs change during the embryonic development and growth of the animal. They are the highest in early embryos, fall before the larva is released into the water, and rise again until the animal is settled. During the further development they fall again, but are always among top expressed genes. Similar pattern but on a smaller scale is observed for WDR77 and PRMT5 homologs as well.

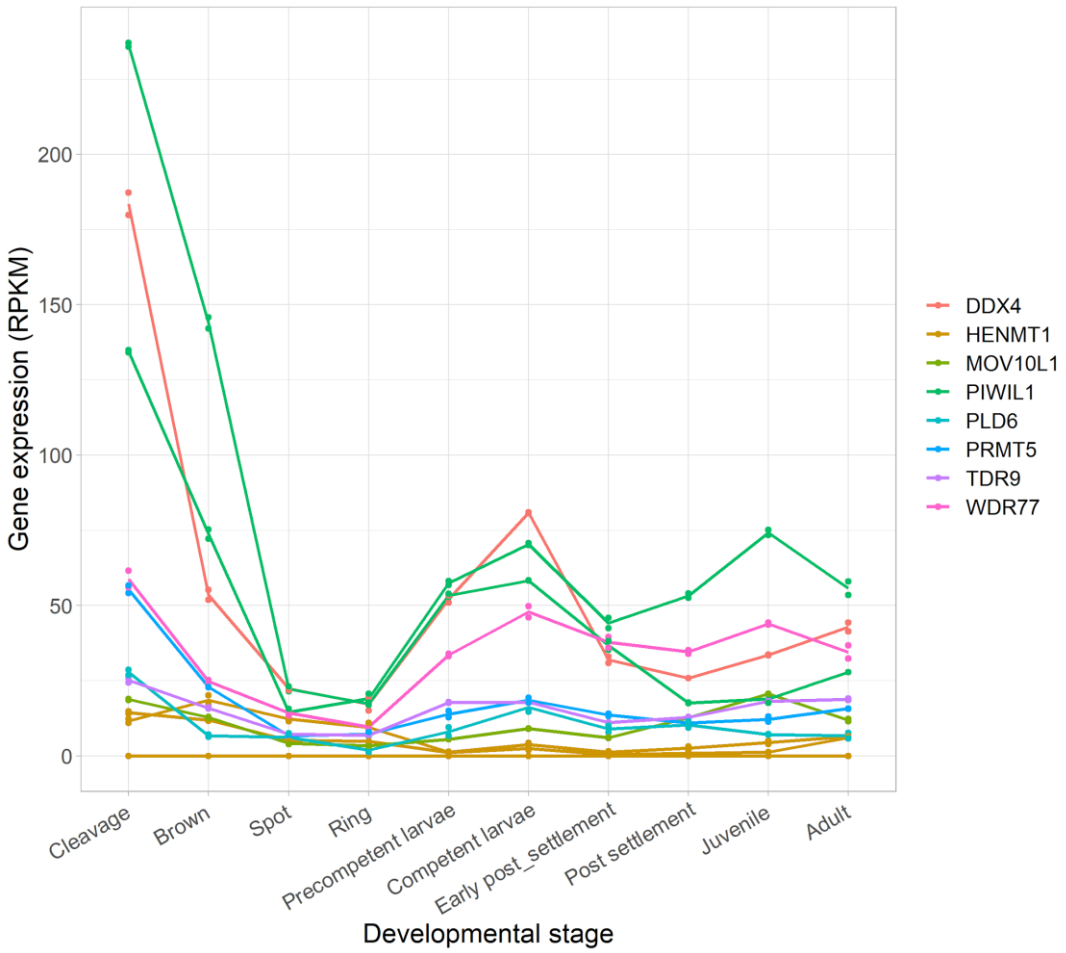



*Figure 29. Expression of homologs involved in the piRNA pathway during different developmental stages of A. queenslandica. RPKM= reads per kilobase of coding sequence length per million reads*

# 4 Discussion

## 4.1 Genome assembly of *Eunapius subterraneus* and *Suberites domuncula*

In the thesis, I present high quality draft assemblies for previously unpublished genomes of *E. subterraneus* and *S. domuncula*. The number of scaffolds in the assembly and their lengths are often referred to as contiguity. The contiguity measured by N50 value for *E. subterraneus* assembly is comparable to the genomes of *A. queenslandica*, *T. wilhelma* and *S. ciliatum*, while *S. domuncula* has an even higher contiguity of 420 kb. Unfortunately, the degree of contiguity varies among different genomes and assemblies, and is not well correlated with genome correctness (Salzberg *et al.*, 2012)

The total size of the assembled genome of *E. subterraneus* is 185.5 million bases, which is close to the average size (200 Mb) of the sponge genomes (Jeffery, Jardine and Gregory, 2013). The only sponge from the genus Eunapius whose genome size was determined experimentally is *E. fragilis*, with genome size estimated to 303 Mb. However, the estimation of the genome size varies across the same genus, for example *Haliclona* genus shows the estimated genome size range from 78 Mb in *H. implexiformis* to 205 Mb in *H. cymaeformis (Jeffery, Jardine and Gregory, 2013)*, so it would be valuable to experimentally determine the genome size for *E. subterraneus* to get a sense of assembly completeness.

The total size of the assembled genome for *S. domuncula* is 101.3 Mb. This size is larger than the experimentally determined genome size for other members of Suberites genus, *S. aurantiacus* (68 Mb) and *Suberites sp.*(88 Mb) (Jeffery et al. 2013). Although the genome size for *S. domuncula* was previously experimentally determined to be 1564.8 Mb based on Feulgen densitometry (Imsiecke et al. 1995), those values are considered highly unreliable due to methodological issues producing estimates of up to double the highest determined sponge genome (Jeffery et al. 2013).

A common computational approach to estimate genome completeness is to calculate the percentage of identified conserved single copy Metazoan genes. The estimates are 91.3% for *S. domuncula* and 81.7% for *E. subterraneus*, suggesting a very good overall quality of assembled genomes. Compared to the available published genomes, assemblies presented here are ranked first (*S. domuncula*) and third (*E. subterraneus*). The gene completeness based metrics do not necessarily correlate well with genome quality estimates based on transposable element analysis (Ou, Chen and Jiang, 2018; Wierzbicki *et al.*, 2020), since repetitive regions

are usually more problematic to assemble than gene rich regions. Unfortunately, transposon-based quality estimates demand at least 5% of the genome to be annotated as LTR transposons (Ou, Chen and Jiang, 2018; Wierzbicki *et al.*, 2020), or the locations of piRNA clusters to be known in the genomes (Ou, Chen and Jiang, 2018; Wierzbicki *et al.*, 2020)), none of which was available for all of the genomes I analysed.

The assemblies I present are based on long reads produced by nanopore sequencing technology which outperform the assembly of the repetitive regions of the genome compared to short reads (Amarasinghe *et al.*, 2020). Since long read assembly is erroneous, I used the short reads to correct the errors in the assembled long reads. Using the short reads alone to correct the assembly improved the genome completeness for both *E. subteraneus* and *S. domuncula* assemblies. The completeness was further improved by a pipeline I designed that assembles short reads into an assembly graph and corrects the nanopore based assembly with the assembled graph. The pipeline is designed to improve assembly parts containing any type of sequence (gene rich regions or repetitive).

Moreover, the approach presented here does not require previous correction of the nanopore dataset, and is demonstrated to work equally well with uncorrected nanopore reads as well as when nanopore reads were corrected in three iterative rounds with short reads prior to assembly, even in a setting where the total coverage of the nanopore reads used is under 20x. Higher nanopore coverage is beneficial in improving both assembly contiguity and completeness, as is the case with S. domuncula where only nanopore reads longer than 5000 bases were used, with coverage 40x.

Most contiguous genome assemblies are produced when using a combination of short and long reads together with HiC data, as demonstrated by the chromosome level assembly of the sponge *E. muelleri* (Kenny *et al.*, 2020), so there is great potential to use such data to further improve the contiguity of the assemblies here presented. However, as apparent from the *E. muelleri* assembly, the high contiguity does not imply best assembly, at least when measured by BUSCO annotation completeness score. Although the contiguity of *E. muelleri* genome is high, with more than 90% of the total length contained in first 24 scaffolds, 21.6% of metazoa specific genes are missing from its assembly, whereas only 7.5% are missing in the assembly of *S. domuncula* and 14.2% in the assembly for *E. subterraneus*.

Furthermore, finishing the genome assembly with HiC data includes incorporating gaps between the contigs. In the case of *E. muelleri*, there were on average 577 unknown nucleotides per 100Kb of sequence length, whereas the genomes I present have on average 1.02 and 0.71 unknown nucleotides per 100Kb of sequence. Those gaps are introduced by software design,

are all of unique length and do not represent true nucleotide distances in the genome. The pipeline I present here could be used to transcend the gaps, which would serve to confirm the joins of the neighbouring contigs as well as to decipher the sequence of the gap.

In summary, there currently exist programs such as LoRDEC (Salmela and Rivals, 2014) which use the de Bruijn graph for correction of long reads, as well as programs which use short reads directly to correct the assemblies (Walker *et al.*, 2014). To my knowledge no protocol exists that uses de Bruijn graph for correction of the assembled genome, and the pipeline presented here outperformed both mentioned approaches.

## 4.2 General characteristics of the genomes

Sponges in general have a high number of genes, and the number of predicted genes in *E. subterraneus* is the highest among all sponges reaching 47022. This number should not come as a surprise as the annotation includes transposon-derived genes because the gene prediction for the presented sponges was done on the unmasked genomes. Genes are shorter than typical eukaryotic genes, and more similar to an average gene length of yeasts (Wagner, 2005). However, the gene lengths are negatively correlated with the number of scaffolds in the genomes, so better assemblies might in future shift the distributions of gene lengths.

Consistent with previous findings (Francis and Wörheide, 2017) all sponges show a ratio of introns to intergenic regions close to one, except *E. muelleri* which show unexpectedly shorter introns/ larger intergenic regions. I have also confirmed the positive correlation between total intron sizes and genome sizes (Deutsch and Long, 1999). Interestingly, the genome of the only Calcareous sponge, *S. ciliatum* shows a drastic difference in lengths of introns compared to exons. Furthermore, although longer first introns are a general property of eukaryotic gene structure which could be connected with their functional properties (Bradnam and Korf, 2008), this feature is only observed in *S. ciliatum* and not in other sponge genomes. Longer introns are connected with (alternative) exon inclusion in mammals (Epstein, 2003). Given the difference in sizes of introns among *Sycon* and other sponges, it will be interesting to compare the homologs between the sponges to determine if such an effect is present in *Sycon*.

It was shown previously that DNA methylation in A. queenslandica is comparable to that in vertebrates (de Mendoza et al. 2019). Same paper shows that the level of methylation in S. ciliatum is also high, but not “vertebrate level high”. A recently published analysis of methylome for the chromosome-level assembled genome of the sponge Ephydatia muelleri (see

figure above and Kenny et al. 2020) shows that the levels of methylation in Ephydatia (37%) are higher than for most invertebrates, but not as high as previously reported for A. queenslandica. The observed CpG depletion shown in this thesis for all sponge species taken together with the exess of CpA and TpG dinucleotides as well as relatively high DNA methylation in A. queenslandica, S. ciliatum and E. muelleri indicate that DNA methylation machinery in all observed sponge species is active and higher than in most invertebrates.

## 4.3 Transposable elements in the phylum Porifera

Transposable elements occupy a large fraction of the sponge genomes and the amount of transposable elements is positively correlated with genome size, consistent with observations for other eukaryotes (Kidwell, 2002). Another paper previously identified repetitive sequences in the sponges *E. muelleri*, *A. queenslandica*, *T. wilhelma* and *S. ciliatum* (Kenny *et al.*, 2020). There is a discrepancy between identified repetitive sequences differs between this thesis and the mentioned paper due to use of a newer version of the software in my thesis which incorporates structural information and results in the annotation of more transposable sequences. It is important to mention that the program used relies on random sampling of the genome and will not give exactly the same results every time. Despite this fact, the general conclusions from the mentioned paper are confirmed here. Most transposable elements in the sponge species could not be classified into either of the known classes, reflecting the fact that they have not been represented in repeat libraries as known transposon consensuses. Thus, transposon consensus libraries I produced de novo for each one of the sponge species will be a valuable resource for future identification and exploration of transposable elements in this phylum.

While in most sponge genomes the proportion of classifiable transposable elements differs among the genomes, *E. muelleri* and *E. subterraneus* show similar distribution of major groups. Most of them belong to LTR elements, followed by DNA transposons and LINEs. This fact is not surprising given the recent split of the two species, predicted by high homology of their mitochondrial DNA (Pleše *et al.*, 2011). It will be interesting to determine the potentially active transposable elements by comparing syntenic regions of those two closely related species. *S. ciliatum* and *O. pearsei* show most differences compared to other sponges. This might be a reflection of their different evolutionary paths, as they belong to Calcarea and Hexactinellida groups whereas other analysed species belong to Demosponges. However, they

are currently the only representatives with sequenced genomes from their class, and it will be interesting to compare those findings with new species once they are available.

## 4.4 Contribution of transposable elements to evolution of sponges

Given that transposable elements constitute up to 60% of the sponge genomes, it is not surprising that they comprise up to almost half of the total exon and intron lengths and up to 65% of all intergenic regions. In general transposable elements are depleted from coding regions of the genome and enriched in intergenic regions. Most species also show a weak depletion in introns, except for *S. ciliatum* where I found transposable elements to be enriched. It will be interesting to explore if the transposons in this sponge are at least partially responsible for the intron lengths (Roy, 2004).

I have shown that LTR elements are the most conserved elements from all transposable elements in the genomes of *E. subterraneus*, *S. domuncula* and *S. ciliatum*. The sequence divergence of LTR elements is lowest in exons, compared to introns and intergenic elements. LTR elements are integrated into the genome as full length elements, containing two long terminal repeats and an internal sequence. After the integration, the internal region of many elements is excised due to the homologous recombination of the long terminal repeats, leaving only one full length "solo" long terminal repeat (Chuong, Elde and Feschotte, 2017). I have shown that in sponges the rate of sequence divergence of intact LTR elements is the lowest compared to solo and unassigned elements. This fact is not surprising since the LTR elements gather mutations upon insertion, and the intact elements are the ones which are inserted in the genomes most recently. Interestingly, I have shown that the rate of sequence divergence of solo LTR elements is lower in exons for most of the species compared to introns or intergenic regions. Same is observed for LTR elements of unassigned type (neither intact nor solo) in the species *A. queenslandica* and *E. muelleri*. This observation might be interesting to further investigate, as their conservation might potentially indicate their co-option as promoters or coding exons. Such co-options are widely reported for LTR elements since some of the types contain preserved regulatory elements (Peaston et al., 2004; Lamprecht et al., 2010, Friedli and Trono, 2015; Franke *et al.*, 2017).

Although the genes encoded by transposable elements were expressed, their expression was consistently lower than the expression of genes which were not encoded by transposable elements during the formation of the primmorphs in the species *E. subterraneus*. Genes overlapping LINE elements had lower expression than genes which did not overlap with

transposable elements, even if they were not encoded by the LINE element. On the other hand, the expression of genes which overlapped LTR elements but were not encoded by them was higher than the expression of genes which did not overlap with any transposable element. Expression was higher for genes in which the transposable elements were predominantly integrated into introns compared to exons. Those findings are not surprising having in mind that primmorphs are aggregates of dissociated sponge cells comprising proliferating and differentiating cells which can be cultured for several months (Custodio et al. 1998). From an evolutionary perspective through the viewpoint of a transposable element, an ideal scenario would be to be expressed in the germline, and not the somatic cells (Haig, 2016). Transposable elements which are expressed in the germline would be propagated to the next generation, and the deleterious ones will have been selected against (Calvi and Gelbart, 1994; Kano *et al.*, 2009).

Finally, the lower expression of transposable elements compared to other genes is also expected. Hosts have developed multiple mechanisms to restrict transposable elements expression (Goodier, 2016; Liu et al., 2018; Molaro and Malik, 2016), which include small RNA, chromatin and DNA modification pathways. Piwi-interacting RNA (piRNA) is the most diverse class of small non-coding RNA molecules (Calcino et al. 2018) and the piRNA pathway operates predominantly in germ cells where it targets transposable elements (Hartig, Tomari and Forstemann, 2007; Watanabe *et al.*, 2015; Tóth *et al.*, 2016; Meseure and Alsibai, 2020).

I investigated the expression of the piRNA pathway components during the formation of primmorphs in *E. subterraneus* and the development of *A. queenslandica*. Eight out of fifteen human piRNA pathway components are found in all or most sponges. During the formation of the primmorphs, all the genes involved in piRNA pathway are among the top 40% of all expressed genes, and TDRKH, DDX4 and both PIWIL1 homologs were in the top 20%. By analysing the expression levels of piRNA homologs during the development of the sponge *A. queenslandica*, I found that DDX4 and both PIWIL1 homologs show the highest expression during the early development of the sponge, drop in the spot and ring phase, but grow again in the larval phase and are also active in the adult sponge. The activity of piRNA pathway associated genes could be explained by the change in the ratio of the number of somatic/ germline / stem cells in sponges, since PIWI proteins are shown to be expressed in stem cell analogues in sponges (choanocytes and archeocytes)(Funayama et al. 2010).

It was previously reported (Kenny et al. 2020) that the methylation level of repetitive sequences of *Ephydatia muelleri* positively correlates with the age of repeat, and that repeats located within gene bodies have higher levels of methylation than those located outside of

genes. This was not true for LTR retrotransposons which were more likely to be targeted by DNA methylation irrespective of their position. This selective targeting by methylation together with the expression of PIWI proteins suggests that DNA methylation might be involved in regulation of transposable sequences in sponges, which will be an interesting topic to research in the future.

# 5 Conclusion

In this thesis I used the methods of computational genomics to assemble the genomes of two sponge species and identify and characterize transposable elements in the phylum Porifera.

I present quality draft assemblies for the sponge species *Eunapius subterraneus* and *Suberites domuncula* whose genomes were previously unpublished. To assemble the genomes I designed a pipeline which uses short reads assembled into a de Bruijn graph to correct the errors in the nanopore-only assembly. The pipeline outperformed protocols that use de Bruijn graphs to correct the long reads prior to assembly, as well as protocols in which unassembled short reads are used to correct the errors in the assembled long reads.

Furthermore, I used de novo and repository based methods to identify potential transposons in my assemblies and all publicly available sponge genome assemblies. I produced libraries containing consensuses for transposons found in each sponge species. Since no such libraries are currently available for sponges, they will be useful for further exploration of transposons in this phylum.

I annotated the identified transposable elements and compared their distributions between sponge genomes. While the members of the Spongilidae family – *E. muelleri* and *E. subterraneus* showed the most similar proportion of bases belonging to different repeat groups, the other distributions did not follow the phylogeny of the analysed sponge species.

I characterised their impacts on genome evolution of sponges by assessing the contribution to genome organization and analysing their conservation and correlation with gene expression. The percentage of bases assigned to transposable elements generally positively correlated with genome size, most notable for LTR elements. In all sponge species transposable elements are enriched in intergenic regions and depleted in exons. The exceptions are rolling circle type transposable elements in all but *A. queenslandica,* and LTR elements in *S. ciiatum* and *S. domuncula* which are enriched in exons. Transposable elements show varying levels of sequence divergence from the consensus, most higher than 5% which indicates that most groups are present in all species relatively long. Exceptions are LTR elements in *E. subterraneus* and *S. domuncula* and rolling circle elements in *E. muelleri* with median sequence divergences under 5%. Intact LTR elements generally show very low levels of sequence divergence from the consensus regardless of the insertion site – intron, exon or intergenic, whereas solitary LTR elements seem to mutate the least when located in exons while those located in introns and intergenic regions seem to be less resistant to decay. By analyzing gene expression in *E.*

*subterraneus* I conclude that the genes encoded by transposable elements are expressed less than other genes. The same analysis perhaps surprisingly, showed that genes with the integration of LTR or elements of unknown type have higher levels of transcription than the genome average measured by level of mRNA.

Finally, since piRNA pathway in general guardes the genome against transposable elements, I explored this pathway in sponges. I present a manually curated catalog of the homologs of the piRNA pathway in all sponges and show that all but one sponge species observed have two homologs of the PIWIL1 gene. I analysed the expression of identified homologs during the formation of the primmorphs in the species *E. subterraneus* and in ten different developmental stages of the sponge *A. queenslandica* and showed that both PIWIL1 homologs are among top 20% of all genes in the adult form of *E. subterraneus*. Both homologs are also very highly expressed during the development of *A. queenslandica*, although show varying levels which could potentially indicate the existence of and their involvement in different phases of targeting of transposable elements.

# 6 Literature


Adamska, M. *et al.* (2007) 'Wnt and TGF-β Expression in the Sponge Amphimedon queenslandica and the Origin of Metazoan Embryonic Patterning', *PLoS ONE*, p. e1031. doi: 10.1371/journal.pone.0001031.

Amarasinghe, S. L. *et al.* (2020) 'Opportunities and challenges in long-read sequencing data analysis', *Genome biology*, 21(1), p. 30.

*Amphimedon queenslandica Annotation Report* (no date). Available at: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Amphimedon_queenslandica/101/#MaskingPercentagesReport (Accessed: 4 March 2020).

Arkhipova, I. R. and Meselson, M. (2005) 'Diverse DNA transposons in rotifers of the class Bdelloidea', *Proceedings of the National Academy of Sciences*, pp. 11781–11786. doi: 10.1073/pnas.0505333102.

Ax, P. (2012) *Multicellular Animals: A new Approach to the Phylogenetic Order in Nature*. Springer Science & Business Media.

Bankevich, A. *et al.* (2012) 'SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing', *Journal of computational biology: a journal of computational molecular cell biology*, 19(5), pp. 455–477.

Bao, W., Kojima, K. K. and Kohany, O. (2015) 'Repbase Update, a database of repetitive elements in eukaryotic genomes', *Mobile DNA*, 6, p. 11.

Bao, Z. (2002) 'Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes', *Genome Research*, pp. 1269–1276. doi: 10.1101/gr.88502.

*BBMap* (no date) *SourceForge*. Available at: https://sourceforge.net/projects/bbmap/ (Accessed: 30 April 2020).

Bourque, G. *et al.* (2018) 'Ten things you should know about transposable elements', *Genome biology*, 19(1), p. 199.

Bowden, R. *et al.* (2019) 'Sequencing of human genomes with nanopore technology', *Nature communications*, 10(1), p. 1869.

Bradnam, K. R. and Korf, I. (2008) 'Longer first introns are a general property of eukaryotic gene structure', *PloS one*, 3(8), p. e3093.

Britten, R. J. and Kohne, D. E. (1968) 'Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms', *Science*, 161(3841), pp. 529–540.

Brouha, B. *et al.* (2003) 'Hot L1s account for the bulk of retrotransposition in the human population', *Proceedings of the National Academy of Sciences of the United States of America*, 100(9), pp. 5280–5285.

Brůna, T., Lomsadze, A. and Borodovsky, M. (no date) 'GeneMark-EP and -EP : eukaryotic gene prediction with self-training in the space of genes and proteins'. doi: 10.1101/2019.12.31.891218.

Buchfink, B., Xie, C. and Huson, D. H. (2015) 'Fast and sensitive protein alignment using DIAMOND', *Nature methods*, 12(1), pp. 59–60.

Bushati, N. and Cohen, S. M. (2007) 'microRNA Functions', *Annual Review of Cell and Developmental Biology*, pp. 175–205. doi: 10.1146/annurev.cellbio.23.090506.123406.

Bushmanova, E. *et al.* (2019) 'rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data', *GigaScience*. doi: 10.1093/gigascience/giz100.

Butler, J. *et al.* (2008) 'ALLPATHS: De novo assembly of whole-genome shotgun microreads', *Genome Research*, pp. 810–820. doi: 10.1101/gr.7337908.

Calvi, B. R. and Gelbart, W. M. (1994) 'The basis for germline specificity of the hobo transposable element in Drosophila melanogaster', *The EMBO journal*, 13(7), pp. 1636–1644.

Casse, N. *et al.* (2006) 'Species sympatry and horizontal transfers of Mariner transposons in marine crustacean genomes', *Molecular phylogenetics and evolution*, 40(2), pp. 609–619.

Chaisson, M. J. and Pevzner, P. A. (2008) 'Short read fragment assembly of bacterial genomes', *Genome Research*, pp. 324–330. doi: 10.1101/gr.7088808.

de la Chaux, N. and Wagner, A. (2011) 'BEL/Pao retrotransposons in metazoan genomes', *BMC evolutionary biology*, 11, p. 154.

Chen, Q. *et al.* (2017) 'Recent advances in sequence assembly: principles and applications', *Briefings in functional genomics*, 16(6), pp. 361–378.

Chikhi, R. and Medvedev, P. (2014) 'Informed and automated k-mer size selection for genome assembly', *Bioinformatics* , 30(1), pp. 31–37.

Chu, C., Nielsen, R. and Wu, Y. (2016) 'REPdenovo: Inferring De Novo Repeat Motifs from Short Sequence Reads', *PloS one*, 11(3), p. e0150719.

Chuong, E. B. *et al.* (2013) 'Endogenous retroviruses function as species-specific enhancer elements in the placenta', *Nature genetics*, 45(3), pp. 325–329.

Chuong, E. B., Elde, N. C. and Feschotte, C. (2017) 'Regulatory activities of transposable elements: from conflicts to benefits', *Nature reviews. Genetics*, 18(2), pp. 71–86.

Claeys Bouuaert, C. *et al.* (2013) 'The autoregulation of a eukaryotic DNA transposon', *eLife*, 2, p. e00668.

Clark, S. C. *et al.* (2013) 'ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies', *Bioinformatics*, pp. 435–443. doi: 10.1093/bioinformatics/bts723.

Custodio, M. R. *et al.* (1998) 'Primmorphs generated from dissociated cells of the sponge Suberites domuncula: a model system for studies of cell proliferation and cell death', *Mechanisms of Ageing and Development*, pp. 45–59. doi: 10.1016/s0047-6374(98)00078-5.

Deutsch, M. and Long, M. (1999) 'Intron-exon structures of eukaryotic model organisms', *Nucleic acids research*, 27(15), pp. 3219–3228.

Dohm, J. C. *et al.* (2008) 'Substantial biases in ultra-short read data sets from high-throughput DNA sequencing', *Nucleic acids research*, 36(16), p. e105.

Dressman, D. *et al.* (2003) 'Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations', *Proceedings of the National Academy of Sciences of the United States of America*, 100(15), pp. 8817–8822.

Eid, J. *et al.* (2009) 'Real-time DNA sequencing from single polymerase molecules', *Science*, 323(5910), pp. 133–138.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) 'LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons', *BMC bioinformatics*, 9, p. 18.

Epstein, R. J. (2003) *Human Molecular Biology: An Introduction to the Molecular Basis of Health and Disease*. Cambridge University Press.

Erpenbeck, D. *et al.* (2011) 'First evidence of miniature transposable elements in sponges (Porifera)', *Ancient Animals, New Challenges*, pp. 43–47. doi: 10.1007/978-94-007-4688-6_5.

Fedurco, M. *et al.* (2006) 'BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies', *Nucleic acids research*, 34(3), p. e22.

Feschotte, C. and Pritham, E. J. (2007) 'DNA transposons and the evolution of eukaryotic genomes', *Annual review of genetics*, 41, pp. 331–368.

Feuda, R. *et al.* (2017) 'Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals', *Current biology: CB*, 27(24), pp. 3864–3870.e4.

Finnegan, D. J. (1989) 'Eukaryotic transposable elements and genome evolution', *Trends in Genetics*, pp. 103–107. doi: 10.1016/0168-9525(89)90039-5.

Flynn, J. M. *et al.* (2020) 'RepeatModeler2 for automated genomic discovery of transposable element families', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1921046117.

Fortunato, S. A. V. *et al.* (2014) 'Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes', *Nature*, 514(7524), pp. 620–623.

Francis, W. R. *et al.* (2017) 'The genome of the contractile demosponge Tethya wilhelma and the evolution of metazoan neural signalling pathways', *Genomics*. bioRxiv.

Francis, W. R. and Wörheide, G. (2017) 'Similar Ratios of Introns to Intergenic Sequence across Animal Genomes', *Genome biology and evolution*, 9(6), pp. 1582–1598.

Franke, V. *et al.* (2017) 'Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes', *Genome research*, 27(8), pp. 1384–1394.

Friedli, M. and Trono, D. (2015) 'The developmental control of transposable elements and the evolution of higher species', *Annual review of cell and developmental biology*, 31, pp. 429–451.

Funayama, N. *et al.* (2010) 'Piwi expression in archeocytes and choanocytes in demosponges: insights into the stem cell system in demosponges', *Evolution & development*, 12(3), pp. 275–287.

Fu, S., Wang, A. and Au, K. F. (2019) 'A comparative evaluation of hybrid error correction methods for error-prone long reads', *Genome biology*, 20(1), p. 26.

Goerner-Potvin, P. and Bourque, G. (2018) 'Computational tools to unmask transposable elements', *Nature reviews. Genetics*, 19(11), pp. 688–704.

Goodier, J. L. (2016) 'Restricting retrotransposons: a review', *Mobile DNA*. doi: 10.1186/s13100-016-0070-z.

Goubert, C. *et al.* (2015) 'De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti)', *Genome Biology and Evolution*, pp. 1192–1205. doi: 10.1093/gbe/evv050.

Grimson, A. *et al.* (2008) 'Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals', *Nature*, 455(7217), pp. 1193–1197.

Haas, B. J. *et al.* (2013) 'De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis', *Nature protocols*, 8(8), pp. 1494–1512.

Haig, D. (2016) 'Transposable elements: Self-seekers of the germline, team-players of the soma', *BioEssays: news and reviews in molecular, cellular and developmental biology*, 38(11), pp. 1158–1166.

Harcet, M. *et al.* (2010) 'Demosponge EST sequencing reveals a complex genetic toolkit of the simplest metazoans', *Molecular biology and evolution*, 27(12), pp. 2747–2756.

Hartig, J. V., Tomari, Y. and Forstemann, K. (2007) 'piRNAs--the ancient hunters of genome invaders', *Genes & Development*, pp. 1707–1713. doi: 10.1101/gad.1567007.

Hartl, D. L., Lohe, A. R. and Lozovskaya, E. R. (1997) 'Modern thoughts on an ancyent marinere: function, evolution, regulation', *Annual review of genetics*, 31, pp. 337–358.

Hebert, P. D. N. *et al.* (2018) 'A Sequel to Sanger: amplicon sequencing that scales', *BMC genomics*, 19(1), p. 219.

Heydari, M. *et al.* (2017) 'Evaluation of the impact of Illumina error correction tools on de novo genome assembly', *BMC bioinformatics*, 18(1), p. 374.

Higgie, S. (no date) 'Horizontal gene transfer in the sponge Amphimedon queenslandica'. doi: 10.14264/uql.2018.258.

Hoff, K. J. *et al.* (2019) 'Whole-Genome Annotation with BRAKER', *Methods in molecular biology* , 1962, pp. 65–95.

Hooper, J. N. A. and van Soest, R. W. M. (2012) *Systema Porifera: A Guide to the Classification of Sponges*. Springer Science & Business Media.

Houck, M. A. *et al.* (1991) 'Possible horizontal transfer of Drosophila genes by the mite Proctolaelaps regalis', *Science*, 253(5024), pp. 1125–1128.

Howe, K. L. *et al.* (2020) 'Ensembl Genomes 2020-enabling non-vertebrate genomic research', *Nucleic acids research*, 48(D1), pp. D689–D695.

Hubley, R. *et al.* (2016) 'The Dfam database of repetitive DNA families', *Nucleic acids research*, 44(D1), pp. D81–9.

Hunt, M. *et al.* (2013) 'REAPR: a universal tool for genome assembly evaluation', *Genome biology*, 14(5), p. R47.

Huson, D. H. *et al.* (2018) 'MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs', *Biology direct*, 13(1), p. 6.

Ivics, Z. *et al.* (2009) 'Transposon-mediated genome manipulation in vertebrates', *Nature methods*, 6(6), pp. 415–422.

Jain, M. *et al.* (2015) 'Improved data analysis for the MinION nanopore sequencer', *Nature methods*, 12(4), pp. 351–356.

Jain, M. *et al.* (2016) 'Erratum to: The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community', *Genome biology*, 17(1), p. 256.

Jain, M. *et al.* (2018) 'Nanopore sequencing and assembly of a human genome with ultra-long reads', *Nature biotechnology*, 36(4), pp. 338–345.

Jeffery, N. W., Jardine, C. B. and Gregory, T. R. (2013) 'A first exploration of genome size diversity in sponges', *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada*, 56(8), pp. 451–456.

Kannan, S. *et al.* (2015) 'Transposable Element Insertions in Long Intergenic Non-Coding RNA Genes', *Frontiers in bioengineering and biotechnology*, 3, p. 71.

Kano, H. *et al.* (2009) 'L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism', *Genes & development*, 23(11), pp. 1303–1312.

Kazazian, H. H. (2004) 'Mobile Elements: Drivers of Genome Evolution', *Science*, pp. 1626–1632. doi: 10.1126/science.1089670.

Kchouk, M., Gibrat, J. F. and Elloumi, M. (2017) 'Generations of Sequencing Technologies: From First to Next Generation', *Biology and Medicine*. doi: 10.4172/0974-8369.1000395.

Keller, O. *et al.* (2011) 'A novel hybrid gene prediction method employing protein multiple sequence alignments', *Bioinformatics* , 27(6), pp. 757–763.

Kelley, D. R., Schatz, M. C. and Salzberg, S. L. (2010) 'Quake: quality-aware detection and correction of sequencing errors', *Genome biology*, 11(11), p. R116.

Kenny, N. J. *et al.* (no date a) 'The genomic basis of animal origins: a chromosomal perspective from the sponge Ephydatia muelleri'. doi: 10.1101/2020.02.18.954784.

Kidwell, M. G. (2002) 'Transposable elements and the evolution of genome size in eukaryotes', *Genetica*, 115(1), pp. 49–63.

Kolmogorov, M. *et al.* (2019) 'Assembly of long, error-prone reads using repeat graphs', *Nature biotechnology*, 37(5), pp. 540–546.

Koning, A. P. J. de *et al.* (2011) 'Repetitive Elements May Comprise Over Two-Thirds of the Human Genome', *PLoS Genetics*, p. e1002384. doi: 10.1371/journal.pgen.1002384.

Laha, T. *et al.* (2007) 'The bandit, a new DNA transposon from a hookworm-possible horizontal genetic transfer between host and parasite', *PLoS neglected tropical diseases*, 1(1), p. e35.

Lampe, D. J. *et al.* (2003) 'Recent horizontal transfer of mellifera subfamily mariner transposons into insect lineages representing four different orders shows that selection acts only during horizontal transfer', *Molecular biology and evolution*, 20(4), pp. 554–562.

Lamprecht, B. *et al.* (2010) 'Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma', *Nature medicine*, 16(5), pp. 571–9, 1p following 579.

Lander, E. S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921.

Lander, E. S. (2011) 'Initial impact of the sequencing of the human genome', *Nature*, 470(7333), pp. 187–197.

Le Pennec, G. *et al.* (2003) 'Cultivation of primmorphs from the marine sponge Suberites domuncula: morphogenetic potential of silicon and iron', *Journal of biotechnology*, 100(2), pp. 93–108.

Lerat, E., Rizzon, C. and Biémont, C. (2003) 'Sequence divergence within transposable element families in the Drosophila melanogaster genome', *Genome research*, 13(8), pp. 1889–1896.

Leys, S. P. and Degnan, B. M. (2005) 'Embryogenesis and metamorphosis in a haplosclerid demosponge: gastrulation and transdifferentiation of larval ciliated cells to choanocytes', *Invertebrate Biology*, pp. 171–189. doi: 10.1111/j.1744-7410.2002.tb00058.x.

Li, H. (2018) 'Minimap2: pairwise alignment for nucleotide sequences', *Bioinformatics* , 34(18), pp. 3094–3100.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics* , 25(14), pp. 1754–1760.

Lima, L. *et al.* (2019) 'Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data', *Briefings in bioinformatics*. doi: 10.1093/bib/bbz058.

Lin, Y. *et al.* (no date) 'Assembly of Long Error-Prone Reads Using de Bruijn Graphs'. doi: 10.1101/048413.

Liu, N. *et al.* (2018) 'Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators', *Nature*, 553(7687), pp. 228–232.

Li, W.-H. *et al.* (2001) 'Evolutionary analyses of the human genome', *Nature*, pp. 847–849. doi: 10.1038/35057039.

Lohe, A. R., De Aguiar, D. and Hartl, D. L. (1997) 'Mutations in the mariner transposase: The D,D(35)E consensus sequence is nonfunctional', *Proceedings of the National Academy of Sciences*, pp. 1293–1297. doi: 10.1073/pnas.94.4.1293.

Lohe, A. R. and Hartl, D. L. (1996) 'Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation', *Molecular biology and evolution*, 13(4), pp. 549–555.

Lomsadze, A., Burns, P. D. and Borodovsky, M. (2014) 'Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm', *Nucleic acids research*, 42(15), p. e119.

Lucic, B., Chen, H., Kuzman, M. et al. *et al.* (2019) 'Spatially clustered loci with multiple enhancers are frequent targets of HIV-1 integration', *Nature communications*, 10(1), p. 4059.

Lu, X. *et al.* (2014) 'The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity', *Nature structural & molecular biology*, 21(4), pp. 423–425.

Lyon, M. F. (2006) 'Do LINEs Have a Role in X-Chromosome Inactivation?', *Journal of Biomedicine and Biotechnology*, pp. 1–6. doi: 10.1155/jbb/2006/59746.

Mahmoud, M. *et al.* (2019) 'Efficiency of PacBio long read correction by 2nd generation Illumina sequencing', *Genomics*, 111(1), pp. 43–49.

Marchler-Bauer, A. *et al.* (2017) 'CDD/SPARCLE: functional classification of proteins via subfamily domain architectures', *Nucleic acids research*, 45(D1), pp. D200–D203.

Meseure, D. and Alsibai, K. D. (2020) 'Part 1: The PIWI-piRNA Pathway Is an Immune-Like Surveillance Process That Controls Genome Integrity by Silencing Transposable Elements', *Chromatin and Epigenetics*. doi: 10.5772/intechopen.79974.

Metzker, M. L. (2010) 'Sequencing technologies - the next generation', *Nature reviews. Genetics*, 11(1), pp. 31–46.

Miga, K. H. *et al.* (2019) 'Telomere-to-telomere assembly of a complete human X chromosome', *bioRxiv*. doi: 10.1101/735928.

Mikheenko, A. *et al.* (2018) 'Versatile genome assembly evaluation with QUAST-LG', *Bioinformatics* , 34(13), pp. i142–i150.

Miller, D. E. *et al.* (2018) 'Highly Contiguous Genome Assemblies of 15 Species Generated Using Nanopore Sequencing', *G3* , 8(10), pp. 3131–3141.

Molaro, A. and Malik, H. S. (2016) 'Hide and seek: how chromatin-based pathways silence retroelements in the mammalian germline', *Current opinion in genetics & development*, 37, pp. 51–58.

Morgulis, A. *et al.* (2006) 'WindowMasker: window-based masker for sequenced genomes', *Bioinformatics* , 22(2), pp. 134–141.

Muggli, M. D. *et al.* (2015) 'Misassembly detection using paired-end sequence reads and optical mapping data', *Bioinformatics* , 31(12), pp. i80–8.

Mukherjee, S. *et al.* (2019) 'Genomes OnLine database (GOLD) v.7: updates and new features', *Nucleic acids research*, 47(D1), pp. D649–D659.

Muñoz-López, M. and García-Pérez, J. L. (2010) 'DNA transposons: nature and applications in genomics', *Current genomics*, 11(2), pp. 115–128.

Myers, E. W. (2005) 'The fragment assembly string graph', *Bioinformatics* , 21 Suppl 2, pp. ii79–85.

Nichols, S. A. *et al.* (2012) 'Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/β-catenin complex', *Proceedings of the National Academy of Sciences of the United States of America*, 109(32), pp. 13046–13051.

Nigumann, P. *et al.* (2002) 'Many human genes are transcribed from the antisense promoter of L1 retrotransposon', *Genomics*, 79(5), pp. 628–634.

Novák, P., Neumann, P. and Macas, J. (2010) 'Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data', *BMC bioinformatics*, 11, p. 378.

Obbard, D. J. *et al.* (2009) 'The evolution of RNAi as a defence against viruses and transposable elements', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1513), pp. 99–115.

Ostertag, E. M. *et al.* (2002) 'A mouse model of human L1 retrotransposition', *Nature genetics*, 32(4), pp. 655–660.

Ostertag, E. M. and Kazazian, H. H., Jr (2001) 'Biology of mammalian L1 retrotransposons', *Annual review of genetics*, 35, pp. 501–538.

Ou, S. *et al.* (2019) 'Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline', *Genome biology*, 20(1), p. 275.

Ou, S., Chen, J. and Jiang, N. (2018) 'Assessing genome assembly quality using the LTR Assembly Index (LAI)', *Nucleic acids research*, 46(21), p. e126.

Ou, S. and Jiang, N. (2018) 'LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons', *Plant physiology*, 176(2), pp. 1410–1422.

Parra, G., Bradnam, K. and Korf, I. (2007) 'CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes', *Bioinformatics* , 23(9), pp. 1061–1067.

Peaston, A. E. *et al.* (2004) 'Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos', *Developmental cell*, 7(4), pp. 597–606.

Peona, V. *et al.* (no date) 'Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise'. doi: 10.1101/2019.12.19.882399.

Pevzner, P. A., Tang, H. and Waterman, M. S. (2001) 'An Eulerian path approach to DNA fragment assembly', *Proceedings of the National Academy of Sciences of the United States of America*, 98(17), pp. 9748–9753.

Phillippy, A. M., Schatz, M. C. and Pop, M. (2008) 'Genome assembly forensics: finding the elusive mis-assembly', *Genome Biology*, p. R55. doi: 10.1186/gb-2008-9-3-r55.

Piskurek, O. and Jackson, D. J. (2012) 'Transposable elements: from DNA parasites to architects of metazoan evolution', *Genes*, 3(3), pp. 409–422.

Plasterk, R. H. A., Izsvák, Z. and Ivics, Z. (1999) 'Resident aliens: the Tc1/ mariner superfamily of transposable elements', *Trends in Genetics*, pp. 326–332. doi: 10.1016/s0168-9525(99)01777-1.

Pleše, B. *et al.* (2011) 'The mitochondrial genome of stygobitic sponge Eunapius subterraneus: mtDNA is highly conserved in freshwater sponges', *Ancient Animals, New Challenges*, pp. 49–59. doi: 10.1007/978-94-007-4688-6_6.

Pop, M. and Salzberg, S. L. (2008) 'Bioinformatics challenges of new sequencing technology', *Trends in genetics: TIG*, 24(3), pp. 142–149.

Price, A. L., Jones, N. C. and Pevzner, P. A. (2005) 'De novo identification of repeat families in large genomes', *Bioinformatics*, pp. i351–i358. doi: 10.1093/bioinformatics/bti1018.

Rahman, A. and Pachter, L. (2013) 'CGAL: computing genome assembly likelihoods', *Genome biology*, 14(1), p. R8.

Rang, F. J., Kloosterman, W. P. and de Ridder, J. (2018) 'From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy', *Genome biology*, 19(1), p. 90.

*RepeatMasker Home Page* (no date). Available at: http://www.repeatmasker.org (Accessed: 18 May 2020).

Reuter, J. A., Spacek, D. V. and Snyder, M. P. (2015) 'High-Throughput Sequencing Technologies', *Molecular Cell*, pp. 586–597. doi: 10.1016/j.molcel.2015.05.004.

Riesgo, A. *et al.* (2014) 'The analysis of eight transcriptomes from all poriferan classes reveals surprising genetic complexity in sponges', *Molecular biology and evolution*, 31(5), pp. 1102–1120.

Robertson, H. M. (1993) 'The mariner transposable element is widespread in insects', *Nature*, 362(6417), pp. 241–245.

Rodriguez-Martin, B. *et al.* (2020) 'Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition', *Nature genetics*. doi: 10.1038/s41588-019-0562-0.

Roy, S. W. (2004) 'The origin of recent introns: transposons?', *Genome biology*, 5(12), p. 251.

rrwick (no date) *rrwick/Porechop, GitHub*. Available at: https://github.com/rrwick/Porechop (Accessed: 29 April 2020).

Ryu, T. *et al.* (2016) 'Hologenome analysis of two marine sponges with different microbiomes', *BMC genomics*, 17, p. 158.

Salmela, L. and Rivals, E. (2014) 'LoRDEC: accurate and efficient long read error correction', *Bioinformatics* , 30(24), pp. 3506–3514.

Salzberg, S. L. *et al.* (2012) 'GAGE: A critical evaluation of genome assemblies and assembly algorithms', *Genome Research*, pp. 557–567. doi: 10.1101/gr.131383.111.

Sarkar, A. *et al.* (2003) 'Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences', *Molecular genetics and genomics: MGG*, 270(2), pp. 173–180.

Schnable, P. S. *et al.* (2009) 'The B73 maize genome: complexity, diversity, and dynamics', *Science*, 326(5956), pp. 1112–1115.

Schuster, A. *et al.* (2018) 'Divergence times in demosponges (Porifera): first insights from new mitogenomes and the inclusion of fossils in a birth-death clock model', *BMC Evolutionary Biology*. doi: 10.1186/s12862-018-1230-1.

Seppey, M., Manni, M. and Zdobnov, E. M. (2019) 'BUSCO: Assessing Genome Assembly and Annotation Completeness', *Methods in Molecular Biology*, pp. 227–245. doi: 10.1007/978-1-4939-9173-0_14.

Shafin, K. *et al.* (2020) 'Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes', *Nature Biotechnology*. doi: 10.1038/s41587-020-0503-6.

Simão, F. A. *et al.* (2015) 'BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics*, pp. 3210–3212. doi: 10.1093/bioinformatics/btv351.

Simpson, J. T. and Durbin, R. (2010) 'Efficient construction of an assembly string graph using the FM-index', *Bioinformatics* , 26(12), pp. i367–73.

Sultana, T. *et al.* (2017) 'Integration site selection by retroviruses and transposable elements in eukaryotes', *Nature reviews. Genetics*, 18(5), pp. 292–308.

Sundaram, V. *et al.* (2014) 'Widespread contribution of transposable elements to the innovation of gene regulatory networks', *Genome research*, 24(12), pp. 1963–1976.

Su, X. Z. *et al.* (1996) 'Reduced extension temperatures required for PCR amplification of extremely A+T-rich DNA', *Nucleic acids research*, 24(8), pp. 1574–1575.

Tarasov, A. *et al.* (2015) 'Sambamba: fast processing of NGS alignment formats', *Bioinformatics* , 31(12), pp. 2032–2034.

Telford, M. J., Moroz, L. L. and Halanych, K. M. (2016) 'Evolution: A sisterly dispute', *Nature*, 529(7586), pp. 286–287.

*The Cost of Sequencing a Human Genome* (no date) *Genome.gov*. Available at: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (Accessed: 11 April 2020).

Thompson, P. J., Macfarlan, T. S. and Lorincz, M. C. (2016) 'Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire', *Molecular cell*, 62(5), pp. 766–776.

Tóth, K. F. *et al.* (2016) 'The piRNA Pathway Guards the Germline Genome Against Transposable Elements', *Advances in experimental medicine and biology*, 886, pp. 51–77.

Treangen, T. J. and Salzberg, S. L. (2011) 'Repetitive DNA and next-generation sequencing: computational challenges and solutions', *Nature reviews. Genetics*, 13(1), pp. 36–46.

Van Soest, R. W. M. *et al.* (2012) 'Global diversity of sponges (Porifera)', *PloS one*, 7(4), p. e35105.

Wagner, A. (2005) 'Energy Constraints on the Evolution of Gene Expression', *Molecular Biology and Evolution*, pp. 1365–1374. doi: 10.1093/molbev/msi126.

Walker, B. J. *et al.* (2014) 'Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement', *PloS one*, 9(11), p. e112963.

Wang, M. and Kong, L. (2019) 'pblat: a multithread blat algorithm speeding up aligning sequences to genomes', *BMC Bioinformatics*. doi: 10.1186/s12859-019-2597-8.

Wang, T. *et al.* (2007) 'Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53', *Proceedings of the National Academy of Sciences of the United States of America*, 104(47), pp. 18613–18618.

Watanabe, T. *et al.* (2015) 'Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline', *Genome Research*, pp. 368–380. doi: 10.1101/gr.180802.114.

*Website* (no date). Available at: http://www.repeatmasker.org (Accessed: 29 May 2020).

Wei, W. *et al.* (2001) 'Human L1 retrotransposition: cis preference versus trans complementation', *Molecular and cellular biology*, 21(4), pp. 1429–1439.

Wenger, A. M. *et al.* (2019) 'Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome', *Nature biotechnology*, 37(10), pp. 1155–1162.

Wicker, T. *et al.* (2007) 'A unified classification system for eukaryotic transposable elements', *Nature reviews. Genetics*, 8(12), pp. 973–982.

Wickham, H. (2011) 'ggplot2', *Wiley Interdisciplinary Reviews: Computational Statistics*, pp. 180–185. doi: 10.1002/wics.147.

Wiens, M. *et al.* (2009) 'Identification and isolation of a retrotransposon from the freshwater sponge Lubomirskia baicalensis: implication in rapid evolution of endemic sponges', *Progress in molecular and subcellular biology*, 47, pp. 207–234.

Wierzbicki, F. *et al.* (2020) 'Generating high quality assemblies for genomic analysis of transposable elements', *Genomics*. bioRxiv.

Wörheide, G. *et al.* (2012) 'Deep Phylogeny and Evolution of Sponges (Phylum Porifera)', *Advances in Sponge Science: Phylogeny, Systematics, Ecology*, pp. 1–78. doi: 10.1016/b978-0-12-387787-1.00007-6.

Wu, B. *et al.* (2017) 'MEC: Misassembly error correction in contigs using a combination of paired-end reads and GC-contents', *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. doi: 10.1109/bibm.2017.8217652.

Xie, W., Donohue, R. C. and Birchler, J. A. (2013) 'Quantitatively increased somatic transposition of transposable elements in Drosophila strains compromised for RNAi', *PloS one*, 8(8), p. e72163.

Yang, L.-A. *et al.* (2019) 'SQUAT: a Sequencing Quality Assessment Tool for data quality assessments of genome assemblies', *BMC genomics*, 19(Suppl 9), p. 238.

Young, H. (2016) 'Faculty of 1000 evaluation for Regulatory evolution of innate immunity through co-option of endogenous retroviruses', *F1000 - Post-publication peer review of the biomedical literature*. doi: 10.3410/f.726187696.793518075.

Yusa, K. *et al.* (2009) 'Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon', *Nature methods*, 6(5), pp. 363–369.

Zeng, L. *et al.* (2018) 'Genome-Wide Analysis of the Association of Transposable Elements with Gene Regulation Suggests that Alu Elements Have the Largest Overall Regulatory Impact', *Journal of computational biology: a journal of computational molecular cell biology*, 25(6), pp. 551–562.

Zerbino, D. R. and Birney, E. (2008) 'Velvet: Algorithms for de novo short read assembly using de Bruijn graphs', *Genome Research*, pp. 821–829. doi: 10.1101/gr.074492.107.

Zhang, H., Jain, C. and Aluru, S. (no date) 'A comprehensive evaluation of long read error correction methods'. doi: 10.1101/519330.

# 7 Appendix

## Results for testing the accuracy of paths on the part of the human genome:

Total length of the 10 selected canu scaffolds was 29160764, and it mapped to 22 chunks of the human genome with the total length of 29845346 bases. There were in total 13095 paths constructed from the SPAdes graph on those scaffolds and they had the length of 15906170, which account for 54.5% of the total saffold lengths. 2365 paths were larger than 1000 bases and their lengths in summation accounted for 46.16% of the total scaffold length

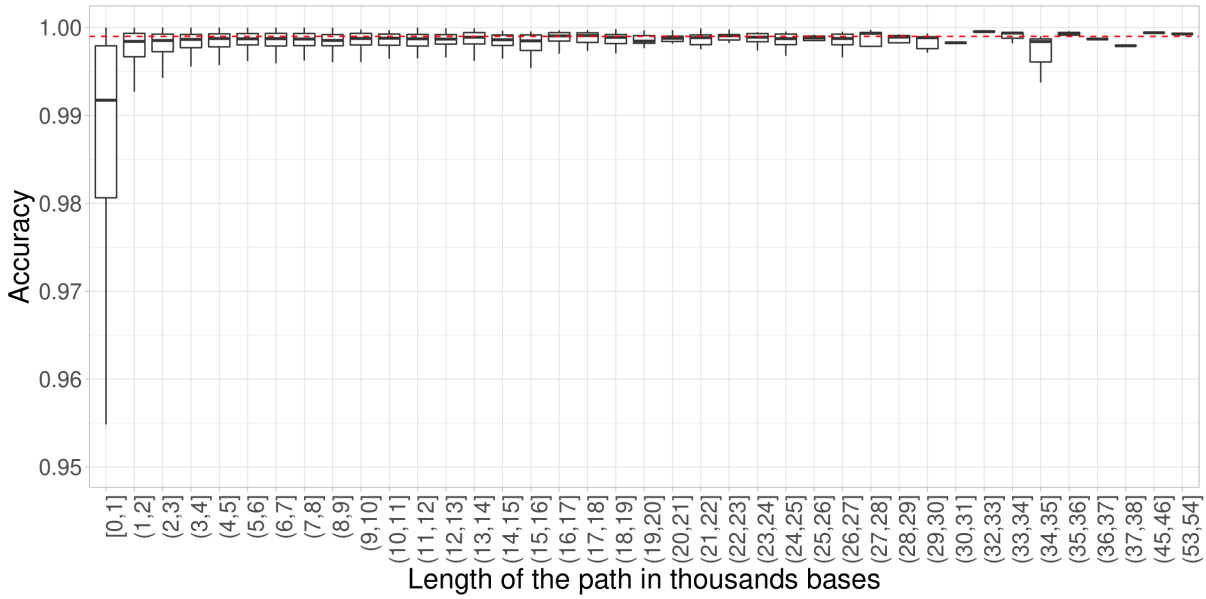



*Figure 30. Paths accuracy on the part of the Human genome. The red line represents 99.9% identity.*

. Figure 31shows the accuracies of paths compared to the human reference genome, depending on the length of the produced paths. Median identity of all paths which were longer than 1000 bases was 99.86%. Those paths were used to correct the nanopore only assembly. The corrected assembly was most obviously improved in the number of indels, which reduced from 1199 indels per 100kbp to 742 indels per 100 kbp (Table 14), but has also improved in total aligned, from 27.4 to 27.6 million aligned bases.

*Table 14. QUAST-LG results for comparison of the published nanopore-only assembly polished with nanopores and the same assembly after polishing with a part of the assembly graph. QUAST results shown calculated based on comparison with the human reference genome.*

| **Genome statistics** | **Nanopore-only, polished with Nanopores** | **Polished with paths** |
|---|---|---|
| Genome fraction (%) | 93.202 | 93.503 |
| Duplication ratio | 0.987 | 0.99 |
| Largest alignment | 2921077 | 2933371 |
| Total aligned length | **27431100** | **27613086** |
| NGA50 | 2812476 | 2820953 |
| LGA50 | 6 | 6 |
| **Misassemblies** | | |
| # misassemblies | 2 | 2 |
| Misassembled contigs length | 5755034 | 5770730 |
| **Mismatches** | | |
| # mismatches per 100 kbp | 648.01 | 633.12 |
| # indels per 100 kbp | **1199.89** | **742.12** |
| # N's per 100 kbp | 0 | 0 |
| **Statistics without reference** | | |
| # contigs | 10 | 10 |
| Largest contig | 2951369 | 2956880 |
| Total length | 29160764 | 29255594 |
| Total length (>= 1000 bp) | 29160764 | 29255594 |
| Total length (>= 10000 bp) | 29160764 | 29255594 |
| Total length (>= 50000 bp) | 29160764 | 29255594 |

## Assembly results for *Eunapius subterraneus* using the LoRDEC corrected nanopore data set:

Correction of *E. subterraneus* nanopore data set with high quality Illumina reads with LoRDEC prior to assembly was performed in 3 rounds. Table 15 shows the resulting nanopore data set after the final round of correction. The correction procedure split some nanopore reads which led to an increase in total number of reads. Total number of bases also increased due to correction of indels.

*Table 15. General statistics on the nanopore only assembly with (Lordec corrected nanopores) and without pre-polishing with Illumina reads*

| Species | Dataset | Number of reads | Number of bases | Estimated coverage | N50 of read length | Mean read length |
|---|---|---|---|---|---|---|
| *Eunapius subterraneus* | Nanopores | 1061783 | 3938829765 | 19.7 | 8645 | 3709.637 |
| *Eunapius subterraneus* | Lordec corrected nanopores | 1064507 | 4013267263 | 20.1 | 8791 | 3770.071 |

Corrected nanopores were used in assembly with Flye, and the results are reported in the Table 16 below.

*Table 16. Genome annotation completeness results for the genome assembled from pre-corrected nanopore reads and polished by reads/paths only, or a combination of paths and reads, measured by BUSCO*

| ***Eunapius subterraneus 3xLoRDEC pre-corrected*** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **BUSCO:** | | **Corrected nanopore only assembly** | **Polished with:** | | | | |
| | | | **Paths only** | **Reads only** | **Paths-> Reads** | **Reads -> Paths** | **Reads->Reads** |
| **Eukaryota Odb10** | Complete | 86.3 | 89.8 | 90.9 | 90.9 | 91.3 | 90.2 |
| | Complete, single copy | 84.3 | 87.8 | 87.8 | 88.2 | 88.2 | 87.8 |
| | Complete, duplicated | 2 | 2 | 3.1 | 2.7 | 3.1 | 3.1 |
| | Fragmented | 6.3 | 5.1 | 3.9 | 3.5 | 3.1 | 3.9 |
| | Missing | 7.4 | 5.1 | 5.2 | 5.6 | 5.6 | 5.2 |
| **Metazoa Odb10** | Complete | 74.6 | 78.8 | 80.1 | 81.4 | 81.1 | 80.1 |
| | Complete, single copy | 71.5 | 75.8 | 77 | 78.4 | 77.6 | 76.9 |
| | Complete, duplicated | 3.1 | 3 | 3.1 | 3 | 3.5 | 3.2 |
| | Fragmented | 7.3 | 5.6 | 4.8 | 4.1 | 4.4 | 5.1 |
| | Missing | 18.1 | 15.6 | 15.1 | 14.5 | 14.5 | 14.8 |

*Table 17. number of bases annotated as repeat by RepeatMasker, when using Repbase24.11 and RepeatModeler2 build consensus as libraries*

| | Amphimedon | Ephydatia | Eunapius | Oscarella | Suberites | Sycon | Tethya | Library used |
|---|---|---|---|---|---|---|---|---|
| Total bases | 165982919 | 322527479 | 185452404 | 57428014 | 101281019 | 344749536 | 125220784 | RepBase 24.11 |
| Bases masked | 9561040 | 34009316 | 22358927 | 931356 | 4299250 | 18663330 | 5433147 | RepBase 24.11 |
| Unclassified | 6072414 | 34429857 | 20563665 | 640622 | 3819247 | 15726061 | 3717565 | RepBase 24.11 |
| Total interspersed repeats | 6072414 | 34429857 | 20563665 | 640622 | 3819247 | 15726061 | 3717565 | RepBase 24.11 |
| Simple repeats | 4717635 | 10028570 | 8096155 | 377215 | 1496723 | 7784959 | 2501765 | RepBase 24.11 |
| LINEs | 789317 | 10477021 | 4361447 | 20288 | 702256 | 7093077 | 725067 | Repeat Modeler2 |
| LTR elements | 2532729 | 22955842 | 8516218 | 15110 | 923072 | 1735726 | 1753904 | Repeat Modeler2 |
| DNA elements | 2654997 | 24809358 | 10341304 | 67624 | 1567474 | 1173350 | 1686144 | Repeat Modeler2 |
| Unclassified | 48515984 | 118977358 | 39414724 | 4549105 | 23117059 | 87256579 | 31539751 | Repeat Modeler2 |
| Total interspersed repeats | 54493027 | 177219579 | 62633693 | 4652127 | 26309861 | 97258732 | 35704866 | Repeat Modeler2 |
| Satellites | 0 | 0 | 198 | 0 | 0 | 0 | 0 | Repeat Modeler2 |
| Simple repeats | 3111174 | 7440893 | 6757601 | 334324 | 1294199 | 7185383 | 2180945 | Repeat Modeler2 |
| Low complexity | 501286 | 850556 | 885585 | 49441 | 44125 | 313569 | 109240 | Repeat Modeler2 |

*Table 18. Relative contributions of different repeat groups to total bases in repeats with similarities to known repeats. Values represent percentages.*

| group | Oscarella | Sycon | Tethya | Amphimedon | Suberites | Eunapius | Ephydatia |
|---|---|---|---|---|---|---|---|
| DNA | 12.0 | 6.6 | 24.0 | 20.4 | 28.4 | 33.9 | 36.5 |
| RC | 13.6 | 0.0 | 7.7 | 25.2 | 14.1 | 0.1 | 0.5 |
| LTR | 2.8 | 10.0 | 24.6 | 19.7 | 18.2 | 28.9 | 33.9 |
| LINE | 3.5 | 38.8 | 10.2 | 5.9 | 13.4 | 16.1 | 15.7 |
| Low complexity | 8.8 | 1.9 | 1.6 | 4.0 | 0.9 | 1.9 | 1.4 |
| Simple repeat | 59.3 | 42.7 | 32.0 | 24.8 | 25.0 | 19.1 | 12.1 |

Identification of repetitive sequences from raw Illumina reads

When using high quality Illumina reads alone, the number of identified repeat consensuses was significantly lower. I identified 40 consensuses of repeats de novo from Illumina reads in the

genome of *E. subterraneus* and 80 in the genome of *S. domuncula*. 37/40 were represented in the consensuses identified de novo from the assembled genome for *E. subterraneus* and 61/80 for the *S. domuncula* genome.

Code

All the code is publicly available at https://github.com/MaKuzman/SpongesTransposons

https://bit.ly/ggplotVsBaseRusers

*Table 19.*

| Species | group | p value |
|---|---|---|
| Ephydatia | LTR | 5.40E-117 |
| Eunapius | Unknown | 1.04E-75 |
| Ephydatia | Unknown | 9.65E-69 |
| Tethya | Unknown | 8.75E-66 |
| Sycon | Unknown | 4.48E-58 |
| Suberites | Unknown | 4.59E-35 |
| Amphimedon | LTR | 1.21E-27 |
| Amphimedon | Unknown | 1.31E-26 |
| Oscarella | Unknown | 1.38E-23 |
| Tethya | LTR | 6.18E-22 |
| Eunapius | LTR | 7.66E-10 |
| Sycon | LINE | 2.51E-07 |
| Ephydatia | LINE | 4.17E-05 |
| Sycon | LTR | 6.85E-04 |
| Suberites | LTR | 8.88E-03 |

*Table 20. Homologs of the piRNA pathway identified in sponge genomes:*

| **Gene name (human)** | **Species** | **Gene name in assembly** |
|---|---|---|
| HENMT1 | Amphimedon queenslandica | Aqu2.1.02440 |
| HENMT1 | Amphimedon queenslandica | Aqu2.1.14913 |
| HENMT1 | Amphimedon queenslandica | Aqu2.1.14915 |
| DDX4 | Amphimedon queenslandica | Aqu2.1.25894 |
| MOV10L1 | Amphimedon queenslandica | Aqu2.1.28281 |
| WDR77 | Amphimedon queenslandica | Aqu2.1.29377 |
| TDRD9 | Amphimedon queenslandica | Aqu2.1.29542 |
| PRMT5 | Amphimedon queenslandica | Aqu2.1.34145 |
| MAEL | Amphimedon queenslandica | Aqu2.1.40655 |
| PIWIL1 | Amphimedon queenslandica | Aqu2.1.42064 |
| PLD6 | Amphimedon queenslandica | Aqu2.1.43344 |
| PIWIL1 | Amphimedon queenslandica | Aqu2.1.43883 |
| KIF17 | Amphimedon queenslandica | Aqu2.1.44010 |
| TDRKH | Ephydatia muelleri | Em0010g41a |
| TDRKH | Ephydatia muelleri | Em0010g79a |
| MOV10L1 | Ephydatia muelleri | Em0010g9a |
| TDRD1 | Ephydatia muelleri | Em0013g844a |
| PIWIL1 | Ephydatia muelleri | Em0015g964a |
| HENMT1 | Ephydatia muelleri | Em0016g986a |
| PIWIL1 | Ephydatia muelleri | Em0017g775a |
| KIF17 | Ephydatia muelleri | Em0017g86a |
| WDR77 | Ephydatia muelleri | Em0021g393a |
| PRMT5 | Ephydatia muelleri | Em0028g34a |
| PRMT5 | Ephydatia muelleri | Em0028g45a |
| TDRKH | Ephydatia muelleri | Em0648g9a |
| MAEL | Eunapius subterraneus | jg11467 |
| PIWIL1 | Eunapius subterraneus | jg14479 |
| DDX4 | Eunapius subterraneus | jg29641 |
| TDRD1 | Eunapius subterraneus | jg31477 |

| | | |
|---|---|---|
| MOV10L1 | Eunapius subterraneus | jg31914 |
| MOV10L1 | Eunapius subterraneus | jg31926 |
| WDR77 | Eunapius subterraneus | jg32590 |
| TDRKH | Eunapius subterraneus | jg39141 |
| PIWIL1 | Eunapius subterraneus | jg44175 |
| PLD6 | Eunapius subterraneus | jg44465 |
| MOV10L1 | Eunapius subterraneus | jg48475 |
| TDRD9 | Eunapius subterraneus | jg52992 |
| KIF17 | Eunapius subterraneus | jg53125 |
| HENMT1 | Eunapius subterraneus | jg56607 |
| HENMT1 | Oscarella paersei | jg10210 |
| MOV10L1 | Oscarella paersei | jg10815 |
| PLD6 | Oscarella paersei | jg15067 |
| KIF17 | Oscarella paersei | jg18258 |
| MAEL | Oscarella paersei | jg3001 |
| PRMT5 | Oscarella paersei | jg4423 |
| MOV10L1 | Oscarella paersei | jg7619 |
| TDRD9 | Oscarella paersei | jg8238 |
| DDX4 | Oscarella paersei | jg8976 |
| HENMT1 | Suberites domuncula | g12049 |
| PIWIL1 | Suberites domuncula | g19725 |
| PIWIL1 | Suberites domuncula | g19828 |
| TDRD1 | Suberites domuncula | g21813 |
| TDRKH | Suberites domuncula | g22673 |
| KIF17 | Suberites domuncula | g2343 |
| DDX4 | Suberites domuncula | g237 |
| KIF17 | Suberites domuncula | g25371 |
| MOV10L1 | Suberites domuncula | g2557 |
| WDR77 | Suberites domuncula | g26874 |
| PLD6 | Suberites domuncula | g667 |
| MAEL | Suberites domuncula | g7886 |

| | | |
|---|---|---|
| PLD6 | Sycon ciliatum | g13582 |
| PLD6 | Sycon ciliatum | g14935 |
| TDRD9 | Sycon ciliatum | g15435 |
| HENMT1 | Sycon ciliatum | g17146 |
| PIWIL1 | Sycon ciliatum | g18502 |
| PIWIL1 | Sycon ciliatum | g21371 |
| TDRD1 | Sycon ciliatum | g23265 |
| KIF17 | Sycon ciliatum | g26519 |
| RNF17 | Sycon ciliatum | g33668 |
| DDX4 | Sycon ciliatum | g33928 |
| TDRKH | Sycon ciliatum | g34848 |
| WDR77 | Sycon ciliatum | g7663 |
| MAEL | Tethya wilhelma | Twilhelma_g14239 |
| KIF17 | Tethya wilhelma | Twilhelma_g16843 |
| PLD6 | Tethya wilhelma | Twilhelma_g2071 |
| DDX4 | Tethya wilhelma | Twilhelma_g25797 |
| MOV10L1 | Tethya wilhelma | Twilhelma_g3408 |
| HENMT1 | Tethya wilhelma | Twilhelma_g4131 |
| PIWIL1 | Tethya wilhelma | Twilhelma_g5354 |
| TDRKH | Tethya wilhelma | Twilhelma_g7153 |

Conserved domains in the sponge homologs of the proteins involved in the (human) piRNA pathway:

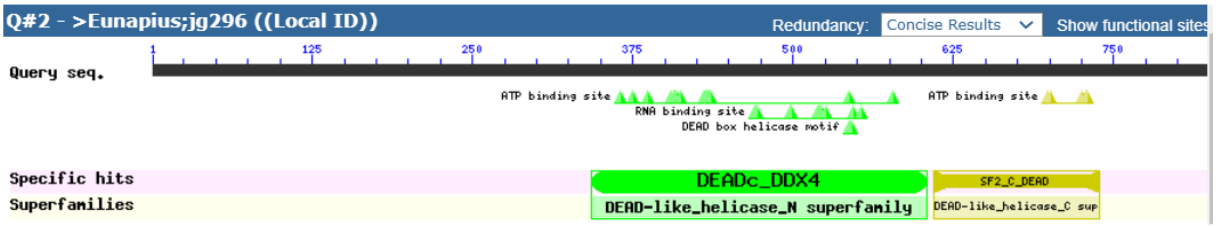



*Figure 31. Conserved domains in the DDX4 E. subterraneus homolog*

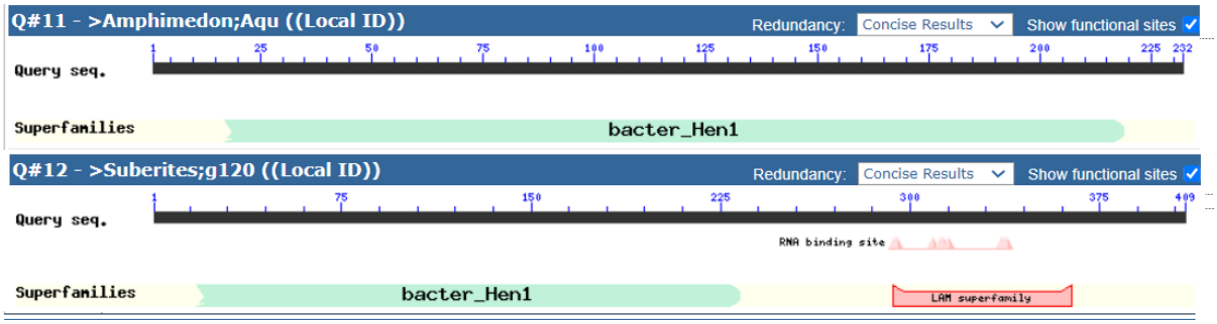



*Figure 32. Conserved domains in the homolog of the human HENMT1 gene in A. queenslandica*

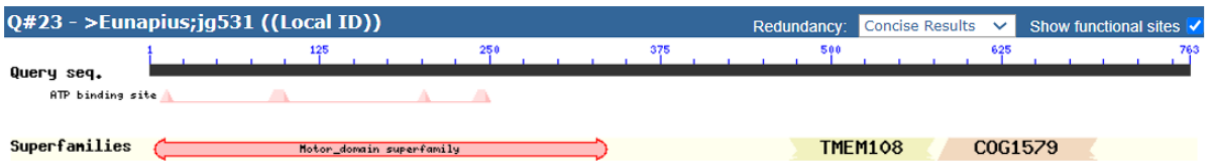



*Figure 33. Conserved domains in the homolog of the human KIF17 gene in E. subterraneus*

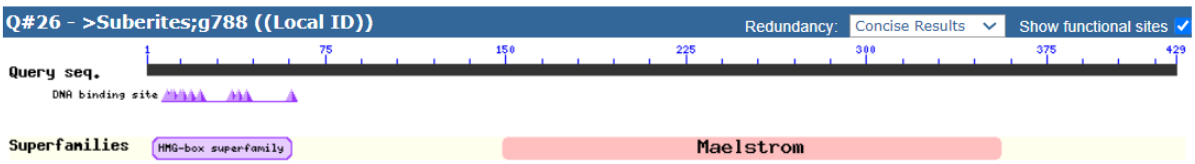



*Figure 34. Conserved domains in the MAEL S. domuncula homolog*

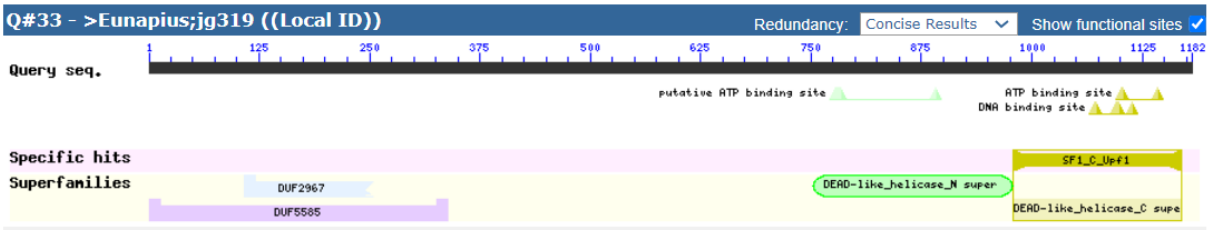



*Figure 35. Conserved domains in the MOV10L1 E. subterraneus homolog*

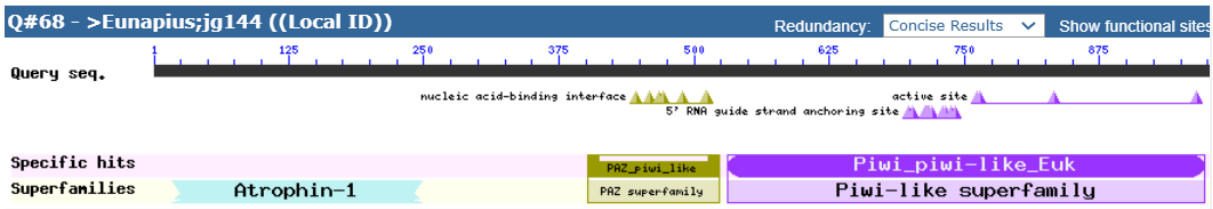



*Figure 36. Conserved domains in the PIWIL1 E. subterraneus homolog*

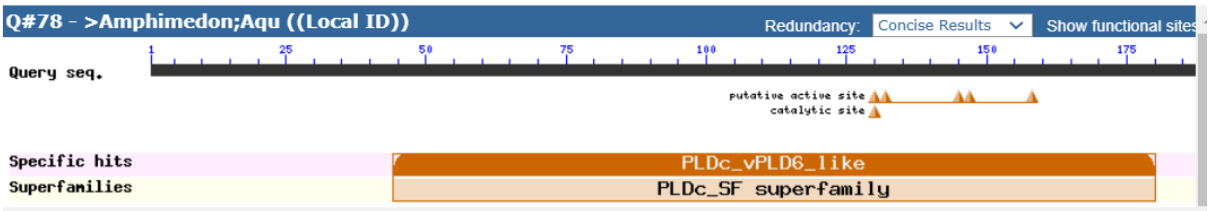



*Figure 37. Conserved domains in the PLD6 A. queenslandica homolog*

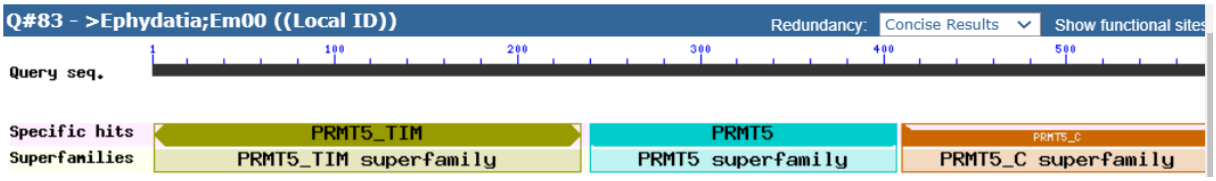



*Figure 38. Conserved domains in the PRMT5 E. muelleri homolog*

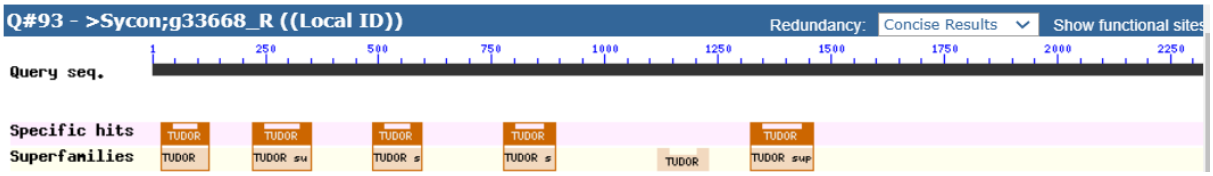



*Figure 39. Conserved domains in the RNF1 S. ciliatum homolog*

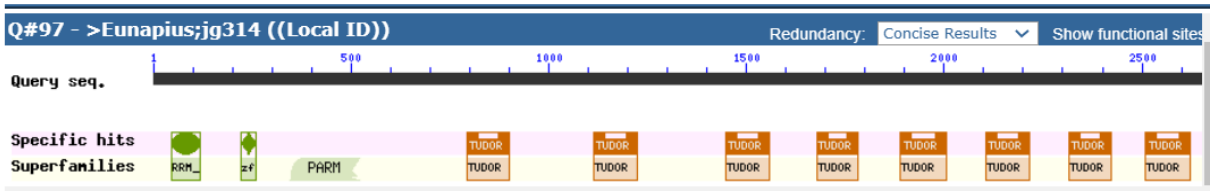



*Figure 40. Conserved domains in the TDRD1 E. subterraneus homolog*

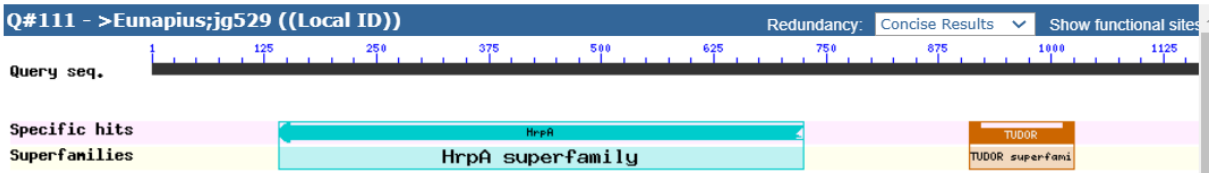



*Figure 41. Conserved domains in the TDRD9E. subterraneus homolog*

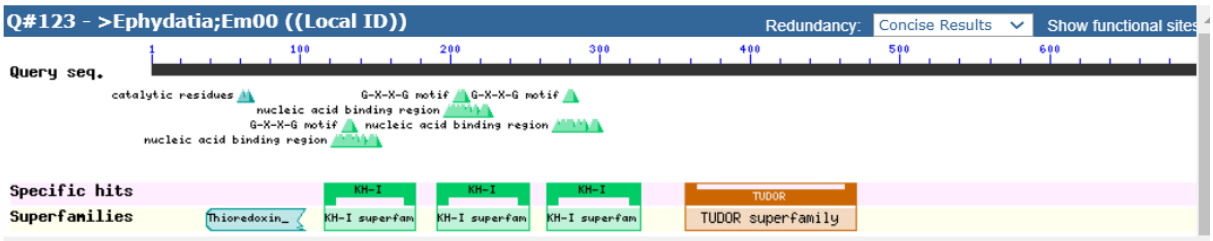



*Figure 42. Conserved domains in the TDRKH E. muelleri homolog*

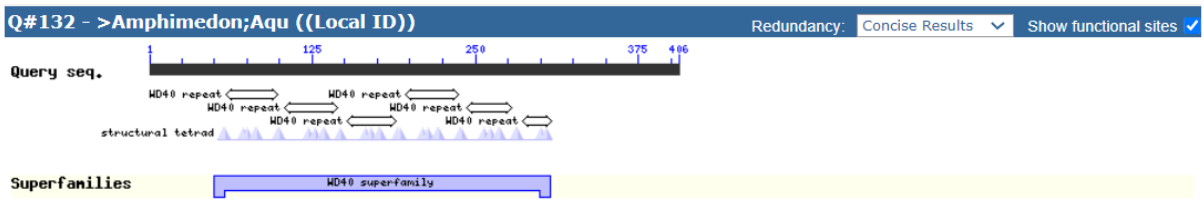



*Figure 43. Conserved domains in the WDR77 A. queenslandica homolog*

# 8 Curriculum vitae

Maja started her education in 2007 in the Department of Mathematics in the University of Zagreb. Three years later she transferred to the Department of Molecular biology determined to become a bioinformatician. After completing bachelors and masters study in molecular biology, she started her doctoral study in biology in the bioinformatics group.

During the entire length of her PhD she has been involved in teaching the courses Bioinformatics, Algorithms and programming, Machine learning and statistics and Computational genomics, and has co-supervised two master thesis.

She gained additional scientific training through collaborations in France and Germany. She attended seven summer schools and conferences where she held three workshops. She has co-organized three international summer schools, two of which in Poland and one in Croatia.

She is a coauthor of one paper and one R package, and the first author of one book chapter. Finally, she is a first author of a paper published in the prestigious journal Nature communications for which she was awarded the Annual award by The society of university teachers, scholars and other scientists of Zagreb.