Jihočeská univerzita v Českých Budějovicích

Přírodovědecká fakulta

Hledání telomerické sekvence u pavouků

Bakalářská práce

Lukáš Martínek

Školitelka: RNDr. Anna Voleníková

České Budějovice 2023

Martínek, L., 2023: **Hledání telomerické sekvence u pavouků.** [Searching for the telomeric sequence in the order Araneae. Bc. Thesis, in Czech] – 46 p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic

Anotace:

Telomeres are highly conserved throughout taxons. Most of Arthropods share the same telomeric sequence $(TTAGG)_n$. There are some deviations from this motif, but in most cases, the telomeric sequence follows the same pattern. However in spiders, the telomeric motif typical for Arthopoda is not present. The aim of this thesis is to explore the spider telomeric sequence by leveraging available sequencing data and find out, if it is a variation of the classic motif.

Finanční podpora:

Práce byla financována grantem Studentské grantové agentury (SGA) Přírodovědecké fakulty Jihočeské univerzity v Českých Budějovicích.

Prohlášení:

Prohlašuji, že jsem autorem této kvalifikační práce a že jsem ji vypracoval pouze s použitím pramenů a literatury uvedených v seznamu použitých zdrojů.

V Českých Budějovicích, 26.03.2023 ……

Lukáš Martínek

Poděkování:

Rád bych poděkoval své školitelce Anně Voleníkové za vedení mé bakalářské práce a za její nekonečnou trpělivost a ochotu. Naučila mě základům bioinformatiky a tím ve mně probudila veliký zájem o tento obor. Rád bych poděkoval vedoucímu Katedry molekulární biologie a genetiky a hlavě naší laboratoře Petru Nguyenovi, který mi toto téma nabídnul a díky komu jsem měl čest být součástí jeho laboratoře a skvělého kolektivu. Dále bych chtěl poděkovat Kseniyi Bobryshavě za poskytnutí kompletního „assembly" motýla *Yponomeuta evonymella* a chromozomálních preparátů, Gabče Krejčové za pomoc s mikroskopováním, Anetě Pilíkové za poskytnutí studijních materiálů, Marcovi Badici za poskytnutý script, Martě Dalíkové za soubor s technickými sekvencemi, Pablovi Mora Ruiz za pomoc s bioinformatikou a Nikolasi Tolarovi za pomoc s formátováním. Taktéž bych chtěl poděkovat svým kolegům Monče Hospodářské, Monče Hrubé, Magdě Vališové, Magdě Zrzavé, a Adamovi Bajgarovi za jejich pomoc, podporu, rozveselení, či prosté povídání si u kafe. V neposlední řadě děkuji své rodině, svému příteli a svým přátelům za podporu během mého bakalářského studia.

Obsah

1. Úvod

1.1. Architektura genomu

Genetická informace je v eukaryotickém genomu uložena ve formě deoxyribonukleové kyseliny (DNA), kterou lze rozdělit na DNA kódující protein a nekódující DNA. DNA kódující protein je tvořena geny, úseky DNA o variabilní velikosti, které se překládají do proteinů. Geny tvoří malé procento celkové genetické informace a skládají se z exonů a intronů. Například u člověka je DNA tvořena pouze z 1.1% exony a z 24% introny (Venter a kol., 2001). Alternativním splicingem vzniká široká variabilita ve výsledném proteinu vzniklého překladem jednoho genu tím, že dojde k vystřižení intronů a ponecháním exonů, které se různě poskládají. Exony se následně překládají do sekvence aminokyselin (Watson, 2014).

Kromě intronů je nekódující DNA tvořena cis-regulačními elementy, mezi které patří sekvence ovlivňující transkripci, jako jsou promotory (sekvence, na které se váže DNA polymeráza), enhancery (sekvence vzdálené od místa exprese, které zesilují efektivitu transkripce), silencery (sekvence vzdálené od místa exprese, které zeslabují efektivitu transkripce), insulátory (sekvence vzdálené od místa exprese, které blokují enhancery a silencery) či sekvence, na které se váží transkripční faktory (Arnone a Davidson, 1997; Pennacchio a Rubin, 2001, shrnuto v Ludwig, 2002).

Další významnou složkou nekódující DNA jsou repetitivní elementy. Repetice jsou motivy sekvencí, které se opakovaně vyskytují napříč genomem v mnoha kopiích. Ačkoliv jsou eukaryotické genomy na repetice bohaté, jejich fukce není zcela objasněna. Předchozí výzkum ukázal, že množství repetitivní DNA ovlivňuje velikost genomu, ale nesouvisí s jeho komplexitou (Van Straalen a Roelofs, 2006). Repetice se proto dříve považovaly za tzv. junk DNA, což je DNA bez specifické funkce (Ohno, 1972). Dnes již ale víme, že hrají významnou roli například v regulaci genové exprese, epigenetice (Pappalardo, 2021), či v evoluci genomu (Feschotte a Pritham, 2007).

Repetitivní elementy nekódují protein, ale jejich přepisem vznikají nekódující RNA, mezi které řadíme tzv. mikroRNA ("miRNA"), malé interferující RNA ("siRNA"), piwi-interagující RNA ("piRNA") a dlouhé nekódující RNA ("lncRNA"). Tyto RNA hrají důležitou roli v mnoha buněčných funkcích. Například u mouchy *Drosophila melanogaster* se mikroRNA podílí na růstu tkání, hormonální aktivitě, vývoji buněk zárodečné linie a buněk nervové tkáně (Carthew a kol., 2017). siRNA a piRNA plní

důležitou roli při vývoji embrya (Lucchetta a kol., 2009; Belicard a kol., 2018). "piwi-interagující" RNA se také podílí na metabolismu tuků (Jones a kol., 2016). lncRNA se obecně podílí na organizaci genomu a zajišťují i spouštění exprese tím, že k sobě přikládají vzdálené části DNA zodpovědné za efektivní expresi, např. enhacery a promotory (Soibam a Zhamangaraeva, 2021; Kajino a kol., 2019). Dále má funkci v buněčném metabolismu a uplatňují se při opravných reakcích poškozené DNA. U samic se také podílí na inaktivaci chromozomu X (Yang a kol., 2015; McHugh a kol., 2015; shnuto v Shapiro, 2022).

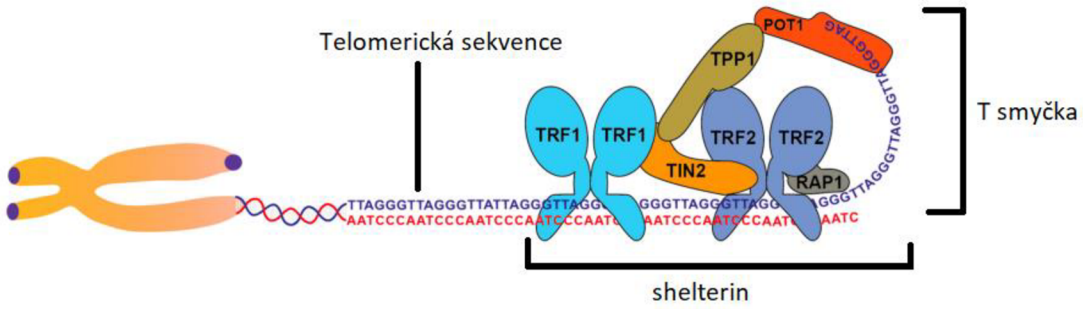
Z organizačního hlediska mohou repetitivní elementy být dvou typů – mobilní elementy a tandemová opakování (Biscotti a kol., 2015). Mobilní elementy neboli transpozony jsou sobecky se pohybující repetitivní sekvence, které mohou mít velikost od 100 páru bazí („bp") do 10 000 páru bazí (Arkhipova a Yushenova, 2019). Obsahují v sobě oblasti, které po přeložení dávají vznik proteinům s biochemickými aktivitami zodpovědnými za stříhání, vkládání, či slepování molekul DNA a nekódující sekvence zodpovědné za jejich pohyb po genomu (Wells a Feschotte, 2020). Existují dva typy mobilních elementů – retrotranspozony a DNA transpozony. Retrotranspozony se po genomu pohybují způsobem nazývaným "copy and paste", což znamená, že se mobilní element z jednoho místa zkopíruje a vloží do jiného místa v genomu. DNA transpozony se po genomu šíří principem tzv. "cut and paste", což znamená, že se mobilní element v jednom místě vystřihne a vloží do jiné části genomu. Při tomto procesu se využívá enzymů transpozáz a integráz (Wells a Feschotte, 2020).

Tandemová opakování jsou repetice tvořící oblasti mnoha za sebou jdoucích jednotek, které se dále podle délky monomeru rozdělují na mikrosatelity (<5 bazí), minisatelity (<25 bazí) a satelity (>150 bp) (shrnuto v práci Jenner a kol., 2022). Satelitní DNA může tvořit i více než 50% DNA v genomu (Mora a kol., 2023). Speciálním typem tandemových repetic jsou strukturální elementy, mezi které patří centromery a telomery (Biscotti a kol., 2015). Centromery jsou tvořené repetitivní sekvencí s doménami značícími místa vytvoření kinetochoru (Willard, 1990) a hrají důležitý význam při replikaci, jelikož se na ně vážou mikrotubuly dělícího vřeténka (Ludwig, 2002). Podle místa výskytu centromery dělíme chromozomy na metacentrické (centromera se nachází uprostřed chromozomu), submetacentrické (centromera je slabě posunutá k jednomu konci chromozomu, takže velikost ramének na jedné straně neodpovídá velikosti ramének na druhé straně chromozomu), akrocentrické (centromera je silně posunutá k jednomu konci chromozomu), telocentrické (centromera se nachází na konci chromozomu). Posledním typem jsou chromozomy holocentrické kde centromera není lokalizovaná a mikrotubuly

se vážou podél větší části chromozomu (Schroeder-Reiter a Wanner, 2009). Navzdory konzervované funkci, centromerická DNA podléhá rychlé evoluci na úrovni sekvence (Logsdon a Eichler, 2022). Telomerická DNA je podrobně popsaná v následujících kapitolách.

1.2 Základní informace o telomerách

Telomery na koncích chromozomů plní ochrannou funkci. Bez telomer by byl chromozom nestabilní a mohl by degradovat (Blackburn, 1991). Samotné konce telomer jsou tvořené repetitivní sekvencí, která formuje na konci smyčku s proteinovým komplexem. Tento proteinový komplex byl podrobně zkoumán u savců a je tvořený proteiny TRF1 a TRF2, které se vyskytují ve formě dimeru, poté RAP1, POT1, TPP1 a TIN2 (Obrázek 1). Na samotném konci telomery se těchto komplexů vyskytuje mnoho (Doksani, 2019). Tato struktura zabraňuje tomu, aby chromozomy navzájem fúzovaly a buňka jejich konce vnímala jako dvojvláknový zlom (Greider, 1999). Přijde-li chromozom o své telomery, vede to ke genomové nestabilitě a tvorbě chromozomových aberací (McClintock, 1940).



Obrázek 1: Schéma chromozomálního konce u savců. Proteinový komplex shelterin se váže na mikrosatelitní telomerickou sekvenci formující na konci chromozomu tzv. T smyčku. Hlavní proteiny této struktury TRF1 a TRF2 se vyskytují ve formě dimeru a vážou se na dvouvláknovou DNA. Jednořetězcový převis na 3´ konci, tzv. G převis, představuje vazebné místo pro POT1 a vytvoření smyčky je zprostředkováno vzájemnou interakcí s dalšími proteiny - RAP1, TPP1 a TIN2 (Doksani, 2019).

Telomery se při každém dělení buňky zkracují, což je způsobeno neúplnou replikací tzv. opožďujícího se řetězce („lagging strand“) DNA polymerázou (Olovnikov,

1973). Délka telomer ma úzkou spojitost se stářím buňky (Harley, 1991). Cooke a Smith (1986) a další prokázali, že telomery spermie byly delší, než telomery somatických buněk. V tomto experimentu byla dokázána přítomnost telomerázy, enzymu, který má na svědomí prodlužování telomer a který je přirozeně aktivní pouze u buněk zárodečné linie. Enzym telomeráza je druh reverzní transkriptázy, RNA-dependentní DNA polymerázy a jeho součástí je RNA sekvence komplementární k DNA vláknu telomery. Dle tohoto vlákna se reverzní transkripcí přepíše RNA sekvence telomerického motivu do DNA, čímž dojde k prodloužení jednoho vlákna telomery. Komplementární vlákno se poté doplní DNA polymerázou (Chan a Blackburn, 2004). Enzym telomeráza je taktéž aktivní v nádorových buňkách, tudíž výzkum telomerázy má úzkou spojitost s výzkumem potenciálních léčiv rakoviny (Cong, a kol., 2002). Pokud by telomeráza u gametických buněk neexistovala, dědily by se zkrácené chromozomy (Harley, 1991).

Prodlužování telomer je ale možné i bez aktivity telomerázy. Místo toho se na konce chromozomů mohou transpozicí přesouvat mobilní elementy z jiných částí genomu, jak bylo pozorováno například u mouchy *Drosophila melanogaster*, nebo motýla *Bombyx mori* (Mason a Biessman, 1995; Fujiwara a kol., 2005). U některých rakovinných buněk, či transformovaných živočišných buněk, dochází k prodlužování telomer homologní rekombinací (Conomos a kol., 2013). Tyto mechanismy jsou označovány jako alternativní prodlužování telomer. Servant a Deininger (2016) formulovali hypotézu, že telomeráza i telomerické retrotranspozony se vyvinuly ze společného ancestrálního retroelementu a tudíž výše zmíněné způsoby prodlužování telomer v eukaryotických genomech koexistují.

1.3 Stavba telomer

Z hlediska sekvence se u eukaryotických organismů vyskytují dva druhy telomer – telomery tvořené tandemovým opakováním mikrosatelitů a mobilními elementy. Sekvence tvořící mikrosatelity bývá v rámci vývojových linií poměrně konzervovaná a lze ji obecně charakterizovat vzorcem $(T_nA_nG_n)_n$. Mikrosatelitní telomerický motiv $(TTAGGG)_n$ je přítomný u obratlovců a lze ho nalézt i u všech bazálních skupin říše Metazoa jako jsou např. houbovci (Porifera), žahavci (Cnidaria), žabernatky (Ctenophora) a vločkovci (Placozoa). Pravděpodobně se tedy jedná o ancestrální telomerický motiv. U rostlin se obdobně vyskytuje $(TTAGGC)_n$, u hlístic lze nalézt $(TTTAGGG)_n$ a u členovců $(TTAGG)_n$ (shrnuto v práci Fulnečková a kol., 2013). U některých skupin členovců však motiv $(TTAGG)_n$ typický pro Arthropoda nenalézáme. Například u čmeláků z řádu blanokřídlých

(Hymenoptera) nalézáme motiv (TTAGGTTGGGG)n (Fajkus a kol., 2022) a u brouků (Coleoptera) pozorujeme telomerickou sekvenci $(TCAGG)_n$ či $(TTAGGG)_n$ (Prušáková a kol. 2021).

Zvláštním případem jsou telomery tvořené mobilními elementy. Tento typ telomer je známý pouze u zástupců řádu Diptera, konkrétně například u komárů rodu *Anopheles* a *Chironomus*, či u octomilky *Drosophila melanogaster* (Biessman a kol., 1998). U octomilky byly telomery podrobně zkoumané a zjistilo se, že obsahují 3 typy retrotranspozibilních elementů, označených jako HeT-A, TART a TAHRE, které se vyskytují v telomerické oblasti náhodně, ale vždy v stejné orientaci (detailně popsáno v Pardue a DeBaryshe, 2008).

K telomerám následně přiléhají subtelomerické oblasti, které jsou taktéž tvořené repeticemi a hrají důležitou roli při replikaci a v genomové stabilitě. U člověka je délka těchto sekvencí v rozpětí 2-15 kilobází a obsahují jak sekvence podobné telomerám, tak repetice specifické pouze pro subtelomery, duplikované segmenty DNA a unikátní (nerepetitivní) sekvence (Riethman a kol., 2005). Často v nich můžeme najít také klastry genů pro ribozomální RNA (např. Roa a Guerra, 2012; Provazníková a kol. 2021).

1.4 Identifikace telomerické sekvence

Ke zkoumání telomer se tradičně využívají laboratorní metody, jako jsou například Southernův přenos (tzv. „Southern blot“) a fluorescenční in-situ hybridizace (FISH). Metoda Southernova přenosu umožňuje detekovat cílovou sekvenci na membráně s navázanou DNA pomocí značené próby. Vyizolovaná genomová DNA zkoumaného organismu se nafragmentuje pomocí štípacích enzymů a rozdělí elektroforézou dle velikosti na agarózovém gelu. DNA z gelu se pomocí působení kladného a záporného náboje přenese na membránu, na kterou se hybridizuje fluorescenčně značená próba. Fluorescenční signál poté prozrazuje přítomnost našeho motivu v genomu zkoumaného organismu (Southern, 2006). Pro zjištění zda se cílová sekvence vyskytuje na koncích chromozomů je možné část vyizolované genomové DNA vystavit působení exonukleázy BAL31, která degraduje konce chromozomů. Pokud je následně próba ve standardním vzorku úspěšně detekována, avšak ve vzorku po inkubaci s enzymem BAL31 ne, zkoumaný motiv se s nejvyšší pravděpodobností vyskytuje v telomerické oblasti (McEachern a kol., 2000). Fluorescenčně značené próby využívá i metoda FISH. Sonda je hybridizovaná na chromozomální preparáty a výsledná fluorescence je detekovaná

fluorescenčním mikroskopem (Levsky a Singer, 2003). Nevýhodou Southernova přenosu i metody FISH je však nutná znalost kandidátního motivu.

Současný rozvoj sekvenačních metod umožňuje k objevení nových kandidátních motivů využít bioinformatických přístupů. Základní metoda využívá sekvenování vzorku kontrolní genomové DNA zkoumaného organismu a vzorku DNA vystaveného působení výše zmíněného enzymu BAL31. Následnou komparativní analýzou repetic v obou vzorcích lze zjistit u kterých sekvencí došlo ke změně frekvence výskytu, a tedy které se potenciálně vyskytují na chromozomálních koncích (Peška a kol., 2017). Tato metoda se využila k identifikaci telomerického motivu např. u *Allium ursinum* (česnek medvědí), jehož telomerický motiv $(CTCGGTTATGGG)_n$ byl dlouho neznámý (Peška a kol., 2017).

K identifikaci nových telomerických motivů se využívají jak programy určené k obecné analýze repetitivních sekvencí, tak nástroje specializované na hledání telomer. Mezi nejčastěji využívané programy patří např. RepeatExplorer2, který analyzuje repetice tak, že vyhledává v jednotlivých sekvenačních čteních podobnosti a následně je spojuje do klastrů (Novák a kol., 2020). Ačkoliv se tímto způsobem podařilo úspěšně identifikovat telomerickou sekvenci u mnoha organismů včetně druhů s velkými genomy (např. Peška a kol. 2015), je známé že RepeatExplorer2 není příliš efektivní při hledání mikrosatelitů. Pro analýzu satelitů s krátkým monomerem je naopak vhodný nástroj Tandem Repeats Finder, který hledá DNA motivy opakující se v určité vzdálenosti od sebe, čímž je zajištěna úspěšnost při detekci krátkých tandemových repetic (Benson, 1999). Pro specifickou identifikaci telomerických sekvencí byla vyvinuta celá řada nástrojů, zpracovávajících jak neposkládaná sekvenační čtení, tak zrekonstruované genomové sekvence (tzv. assembly). Příkladem je program BioSERF, který při prohledávání sekvenačních dat využívá nejen informaci o již známých telomerických motivech, ale i délku monomeru satelitní sekvence a pozici motivu ve fragmentu DNA (Somanathan a Baysdorfer 2018). Telomerické sekvence v genomových assembly či dlouhých čtení sekvenování třetí generace dokáže vyhledat např. Telomere Indentification toolKit (https://github.com/tolkit/telomeric-identifier).

1.5 Pavouci a jejich genomy

Pavouci představují atraktivní skupinu pro studium kvůli mnoha vlastnostem, jako je např. struktura a tvorba hedvábí, které představuje užitečný materiál na poli biomedicínského

průmyslu (Salehi a kol., 2020). Další fascinující aspekty hodné bližšího studia představují toxiny pavoučích jedů (Saez a kol., 2019), nebo chování pavouků, jako jsou třeba zásnubní tanečky u skákavek (Girard a kol., 2011).

V současné době je na světě známo 50 800 druhů pavouků ze 129 čeledí (World Spider Catalog, 2023). Řád Araneae se dělí na 2 velké podřády – Mesothelae a Opisthothelae. Pavouky podřádu Mesothelae (sklípkoši) tvoří jediná čeleď čítající 85 druhů (Coddington, 2005) a představují bazální skupinu pavouků. Primitivnost této skupiny je dána segmentací abdomenu, která u Opisthotelae již přítomná není. Podřád Opisthotelae lze rozdělit na infrařád Mygalomorphae, kam patří dvě čeledi Atipoidea (sklípkankovití) a Avicularioidea (sklípkani), a infrařád Araneomorphae (dvouplicní), kteří představují zbytek pavouků (Wheeler a kol., 2016).

Co se týče karyotypů, pavouci disponují velmi variabilním počtem chromozomů, který sahá od 7 do 94 chromozomů v diploidním stádiu (Suzuki, a kol., 1954). Všechny chromozomy ancestrálního pavoučího karyotypu jsou akrocentrické (Král, 2006). U pavouků nacházíme určení pohlaví odvozené od typu Drosophila, které je charakteristické přítomností chromozomů X a Y, kdy samec je určen heterogametickým uspořádáním pohlavních chromozomů (Bull, 1983). Pavouci mají mnohočetné pohlavní chromozomy X, kdy ancestrální systém je X_20 pro samce (0 symbolizuje chybějící chromozom Y) a $X_1X_2X_2$ pro samice. Tento způsob určení pohlaví je zároveň nejčastější. Např. u pavouků ze skupiny Mygalomorphae však nacházíme až 13 chromozomů X (Sember a kol., 2020). U pavouků z čeledi Paratropididae lze pozorovat chromozom Y. který ale hraje roli nepohlavního chromozomu, vzniklého fúzí chromozomu X s autozómem. V karyotypu některých pavouků jako například u čeledi Sparassidae můžeme pozorovat vícečetné chromozomy Y (Rowell, 1991; Král a kol., 2013). U zástupců čeledi Pholcidae, Sicariidae, Drymusidae, Filistatidae a Hypochilidae nacházíme již jednodušší karyotyp ve formě X_1X_2Y (Král a kol., 2006).

Z hlediska genomiky se jedná o poměrně málo zkoumanou skupinu a mnoho vlastností pavoučích genomů zůstává neznámých. Jedním z důvodů je i jejich velikost způsobená pravděpodobně vysokým podílem repetic, které komplikují bioinformatické analýzy. Gregory a Shorthouse (2003) zkoumali velikost genomu u 115 druhů pavouků a zjistili, že hodnota pro haploidní genom se pohybuje od 0.74 pg (0.72 Gbp) do 5.73 pg (5.60 Gbp) s průměrnou velikostí 2.40 pg (2.30 Gbp). Velkým otazníkem zůstává i pavoučí telomerická sekvence (Kuntner, 2022). Ačkoliv drtivá většina členovců sdílí mikrosatelitní sekvenci $(TTAGG)_n$, pavouci na přítomnost tohoto motivu byli několikrát

testováni s negativním výsledkem. V práci Vítková a kol., (2005) bylo u pavouků rodu Brachypelma za využití Southernova přenosu a FISH dokázáno, že jejich telomery nejsou tvořené motivem $(TTAGG)_n$, ani $(TTAGGG)_n$.

1.6 Sekvenování dat typu „Linked-Reads“

Na základě zatím nepublikovaného výzkumu (Voleníková A. a Nguyen P., osobní komunikace) se zdá, že data typu „Linked-Reads“ od společnosti 10X Genomics by mohla být vhodná pro zkoumání telomer. Sekvenování tohoto typu dat funguje na principu vložení fragmentu vysokomolekulární DNA do gelové kuličky, která se následně přidá do oleje za vzniku emulze. Každá gelová kulička obsahuje tzv. barcode – poznávací technickou sekvenci (Obrázek 2). Každý „barcode“ je unikátní pro každou gelovou uličku a existuje velmi nízká pravděpodobnost, při které se vyskytne stejný „barcode“ na dvou různých gelových kuličkách. Směs těchto gelových kuliček v oleji s navázanou DNA projde isotermní inkubací, což vede k nafragmentování vysokomolekulární DNA molekuly. Po inkubaci vznikne mnoho fragmentů, avšak každý fragment obsahuje právě takový „barcode“, na jaké gelové kuličce byl navázaný. S touto vlastností můžeme zpětně poskládat původní vysokomolekulární DNA molekuly z malých fragmentů. Přítomnost „barcode“ umožňující snadnou rekonstruci krátkých čtení do tzv. assembly představuje hlavní výhodu tohoto sekvenování (Marks a kol., 2019).

Připravená knihovna fragmentů se následně sekvenuje na platformě společnosti Illumina. Každý z fragmentů DNA obdrží tzv. adapter (technická sekvence). Sekvence komplementární k adapterům se přilepí na skleněný čip. Fragmenty DNA s adaptery se následně komplementárně navážou na sekvence přilepené na sklíčko. V dalším kroku se provede PCR reakce, při které se dvojité vlákno DNA za vysoké teploty 94 °C rozvolní za vzniku jednovláknových molekul DNA, na které nasedne DNA polymeráza a dosyntetizuje komplementární DNA sekvenci. Na sklíčku tak vznikne velké množství DNA fragmentů o stejné velikosti, každý obsahující adapter, jeden přilepený na sklíčku, druhý na opačným konci fragmentu DNA. Takto naamplifikovaná DNA vytvoří mosty, což znamená, že se konec DNA fragmentu s adapterem, který není přilepený na sklíčko, ohne a volný konec s adapterem se spojí se sekvencí komplementární k danému adaptéru na sklíčku. Celý proces se opakuje v mnoha cyklech.

Illumina sekvenování produkuje čtení typu „single end“ nebo „paired end”. V prvním případě dochází ke čtení pouze jedné strany fragmentu, zatímco v druhém se čte

fragment z obou stran a vzniká „forward" a „reverse" čtení. Typická délka čtení je 150 párů bazí. S touto znalostí poté můžeme osekvenovat fragmenty o různých délkách na základě toho, jaké množství informace chceme skutečně získat. Chceme-li znát kompletní sekvenci fragmentu, zvolíme fragment o velikosti 300 páru bazí (150 bp „forward"a 150 bp „reverse"). Ve většině případů je ale délka fragmentu delší a neosekvenovaná sekvence mezi párovými čtení se nazývá „insert" (Solodskikh a kol., 2019). Illumina sekvenování patří mezi metody tzv. Next generation sequencing, neboli sekvenování nové generace, které se snaží osekvenovat velké množství informace za krátký čas a za nízkou cenu.

Obrázek 2: Struktura čtení sekvenovaného metodou „Linked-Reads" od 10x Genomics. Sekvence označená *P5* a *P7* představují adaptery pro Illumina sekvenování. Sekvence označené *Read 1* a *Read 2* označují místa primerů pro párové sekvenování. *10x Barcode* je „barcode", sloužící k rekonstrukci původních fragmentů DNA a jehož velikost činí 16 párů bazí. *Sample index* slouží k identifikaci sekvenační knihovny. *gDNA Insert* je sekvenovaný vzorek DNA (Marks a kol, 2019).

2. Cíle práce

Telomery jsou nukleoproteinové struktury, které chrání konce chromozomů. Jejich sekvence bývá v rámci vývojových linií poměrně konzervovaná a zpravidla se jedná o tandemově opakované krátké motivy (Vítková a kol., 2005; Gomes a kol. 2010). U většiny členovců (Arthropoda) můžeme nalézt mikrosatelitní telomerický motiv $(TTAGG)_n$ (Vítková a kol., 2005). V rámci Arthropod však existují výjimky – skupiny, které telomerickou sekvenci $(TTAGG)_n$ nemají. I přesto však jejich telomerická sekvence často následuje podobný vzorec, jako například u čmeláků z řádu Hymenoptera, kteří mají telomerickou sekvenci s motivem $(TTAGGTTGGGG)_n$ (Fajkus a kol., 2022), či u brouků (Coleoptera), u kterých lze najít $(TCAGG)_n$ (Prušáková a kol., 2021). Extrémním případem jsou někteří zástupci dvoukřídlých (Diptera), konkrétně octomilka *Drosophila* a komáři rodu *Anopheles* a *Chironomus*, jejichž telomery jsou tvořené mobilními elementy (Biessman a kol., 1998).

Jedním z velkých taxonů, kde telomerická sekvence zůstává neznámá, jsou pavouci (Araneae). Ačkoliv pavoukům blízce příbuzní Amblypygi z třídy Arachnida mají telomerickou sekvenci $(TTAGG)_n$ (Lerma a kol., 2021), předchozí výzkum vyvrátil přítomnost tohoto motivu v pavoučích telomerách. Stejného výsledku bylo dosaženo i při testování motivu $(TTAGGG)_n$, konzervovaného u mnoha linií živočichů (Metazoa) (Vítková a kol., 2005). Cílem této bakalářské práce je objevit telomerickou sekvenci u pavouků. Ke způsobu identifikace telomerické sekvence bude využit bioinformatický přístup a následná fyzická lokalizace kandidátních motivů na pavoučích chromozomech metodou fluorescenční *in situ* hybridizace.

3. Metody

3.1 Sekvenační data

K získání dat byly využity veřejně dostupné databáze – Darwin Tree of Life (DToL; www.darwintreeoflife.org) a National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov). Vyhledána byla pouze data osekvenována metodou „Linked-Reads“ od společnosti 10X Genomics (10X Genomics, Pleasanton, USA). Data tohoto typu byla zvolena na základě předchozích zkušeností garanta a školitele projektu, kteří u datasetů osekvenovaných tímto přístupem pozorovali telomerickou sekvenci ve vyšší míře než u standardních Illumina dat (A. Voleníková a P. Nguyen, osobní komunikace).

3.2 Konverze dat

Data stažená z databáze DToL byla v komprimovaném formátu cram. Ke konvertování z cram formátu na formát fastq potřebný pro další analýzy bylo využito balíčku SAMtools v. 1.11 (Li a kol., 2009), konkrétně modulů sort a fastq. Data stažená z databáze NCBI byla již ve formátu fastq a tudíž konverzi nevyžadovala.

3.3 Kontrola kvality

K vizualizaci kvality dat byl využit program FastQC v. 0.11.5 (Andrews, 2010). Tento program je vhodný pro všechna sekvenační data druhé generace a poskytuje informace o počtu a kvalitě čtení, přítomnosti technických nebo nadreprezentovaných sekvencí a dalších základních statistikách datasetu.

3.4 Filtrování dat

Technické sekvence a čtení o špatné kvalitě představují problém při analýze repetitivních sekvencí, jelikož programy uzpůsobené k analýze repetic nedokážou rozlišit mezi uměle přidanou technickou sekvencí a pravým čtením. Z tohoto důvodu byly technické sekvence a čtení o špatné kvalitě odstraněny pomocí nástroje Trimmomatic v. 0.39 (Bolger a kol.,

2014) s následujícími parametry: *SLIDINGWINDOW:5:20, CROP:135, HEADCROP:30, MINLEN:105*. Pomocí příkazu SLIDINGWINDOW byla stanovena minimální hranice kvality čtení a čtení o nižší kvalitě byla odstraněna. Pomocí příkazu CROP byl oříznut určitý počet páru bazí od konce čtení, aby byla zajištěna stejná délka výsledných čtení, potřebná pro další aplikace. Příkaz HEADCROP odstraní daný počet páru bazí od začátku čtení – „barcode". Parametr MINLEN udává minimální délku čtení. Soubor s technickými sekvencemi byl poskytnut RNDr. Martinou Dalíkovou, Ph. D. (https://github.com/avolenikova/forStudents/blob/master/Truseq_PE_frankenstein.fa). Úspešné odstranění technických sekvencí a čtení o špatně kvalitě bylo ověřeno opětovnou vizualizací v programu FastQC, popsaného v sekci 3.3 Kontrola kvality.

3.5 Komparativní analýza repetic

Porovnání repetitomu napříč získanými datasety pavouků bylo provedeno pomocí programu RepeatExplorer2 (Novák a kol. 2020). Před samotnou analýzou bylo nutné provést úpravu dat. Sekvenační čtení byla konvertována do formátu fasta pomocí příkazu sed a na začátek „headeru" (hlavičky) čtení bylo přidáno označení druhu a na jeho konec znak /1 označující „forward" čtení a /2 označující „reverse" čtení pomocí příkazu awk (pokud se již na konci čtení nevyskytovaly). Vzhledem k tomu, že program RepeatExplorer2 vyžaduje, aby vstupní soubory měly „coverage" (pokrytí) menší než 1x velikost genomu, z každého datasetu byly vytvořeny tři pseudo náhodné repliky o velikosti 500 000 čtení. Komparativní analýza repetic byla provedena v grafickém rozhraní programu RepeatExplorer2 běžícím na serveru Galaxy (The Galaxy Community, 2022), s nastavením pro párové čtení, databází repetic Metazoa verze 3.0 a automatickým filtrováním abundantních satelitů. Ostatní parametry byly ponechány v základním nastavení. Výsledky byly následně zpracovány v programu RStudio (RStudio Team, 2020).

3.6 Zjišťování velikosti genomu

Pro složení genomové sekvence a druhově specifické analýzy bylo třeba znát informace o velikosti zkoumaných genomů. Toho bylo dosaženo pomocí programu GenomeScope v. 2.0 (Ranallo-Benavidez a kol., 2020), který na základě frekvence k-merů sekvenačních dat

dokáže in silico odhadnout velikost genomu, charakterizovat míru heterozygosity, nebo repetitivní kontent. 21-merové spektrum bylo vytvořeno v programu jellyfish v. 2.3.0 (Marçais a Kingsford, 2011). Vstupním souborem byly veškeré párové čtení získané v sekci 3.4 Filtrování dat, které byly spojené do jednoho souboru. Výsledný histogram byl analyzován ve webovém rozhraní http://qb.cshl.edu/genomescope/genomescope2.0/, kde je možné si informace o genomu prohlédnout.

3.7 Složení genomové sekvence („assembly“)

Tvorbu tzv. assembly z dat osekvenovanými metodou „Linked-Reads“ nabízí exkluzivně program Supernova v. 2.1.1 taktéž vyvinutá společností 10x Genomics (Weisenfeld a kol., 2017). Tzv. barcode přiřazený jednotlivým fragmentům DNA se zde uplatňuje k sestavení původní vysokomolekulární DNA. Vstupním souborem jsou „forward“ a „reverse“ čtení ve formátu fastq z odstavce 2.2 Konvertování dat (data nezbavená technických sekvencí a nefiltrovaná na kvalitu). Styl vytvořeného „assembly“ byl zvolen *pseudohap* („assembly“ tvořené jedním pseudohaplotypem). Parametr *maxreads* byl vypočítán následovně: (velikost genomu x 56) / velikost čtení. Velikost genomu byla zjištěna pomocí programu GenomeScope (viz výše). Číslo 56 odkazuje na 56x coverage genomu, která je doporučená autory programu.

3.8 Vyhodnocení kvality „assembly“

Kvalita vytvořeného „assembly“ byla prověřena pomocí programů Quast v. 4.6.3 (Gurevich a kol. 2013) a BUSCO v. 5.2.2 (Manni a kol. 2021). Tyto programy ukazují, jak je dané „assembly“ kompletní, jak moc je fragmentované, z kolika kontigů je se skládá apod. Quast výstup obsahuje tabulku, ve které jsou vypsané počty kontigů a jejich velikosti, celková velikost a délka, procentuální zastoupení C a G bazí a hodnota N50. N50 je číslo udávající délku, pro kterou kontigy o stejné délce či delší zaujímají alespoň polovinu všech bazí v „assembly“. Čím větší je toto číslo, tím menší je fragmentovanost „assembly“. Výsledky programu BUSCO poskytují informace o kompletnosti „assembly“ na základě úspěšné identifikace genů konzervovaných v příslušné vývojové linii. U programu BUSCO je tedy třeba specifikovat dataset, se kterým se dané „assembly“ porovná. V našem případě byl vybrán dataset *arachnida*, který obsahuje geny typické pro

pavoukovce. Čím více BUSCO genů společných pro tento dataset a „assembly“ pavouka nalezneme, tím je zrekonstruovaná genomová sekvence kompletnější. V případě kontrolního datasetu z motýla *Yponomeuta evonymella* byl použit dataset *lepidoptera*.

3.9 Ověření molekulární identity „assembly“

Pro kontrolní identifikaci druhu ze kterého bylo vytvořeno „assembly“ byl použit program BLAST (Altschul, 1990) implementovaný v GeneiousPrime v. 2020.1.2 (https://www.geneious.com). BLAST vyhledává podobnosti mezi vybranou sekvencí a datasety nahranými v konkrétní databázi. V našem případě byla zvolena sekvence genu pro *cytochrom oxidázu I* (COI) příslušného druhu získaná z databáze NCBI, která byla hledána ve zhotoveném „assembly“ a nalezená sekvence byla zpětně porovnána s NCBI databází nt/nr. Gen COI má silně konzervované části, které umožňují jeho identifikaci napříč fylogenetickými liniemi, ale zároveň obsahuje variabilní části, které rychle mutují a tudíž jsou druhově specifické, a proto je pro tuto aplikaci vhodný (Hebert a kol., 2003)

3.10 Příprava druhově specifických datasetů

Pro jednodušší identifikaci telomerické sekvence bylo přistoupeno k vytvoření datasetu obohaceného o koncové sekvence kontigů. Na základě informací z literatury a kontinuity vytvořeného „assembly“ byla zvolena délka konců kontigů 5 000 páru bazí. Pro izolaci těchto konců byl použit script v programovacím jazyce Python (Van Rossum a kol., 2009), jehož autorem je BSc. Marco Badici (Příloha 1). Tento skript vyselektuje kontigy větší než 5000 páru bazí, ze kterých extrahuje pouze koncové části o délce 5 000 páru bazí. U kontigů menších než 5000 páru bazí k extrakci koncových částí nedochází a v datasetu jsou ponechány.

Na „assembly“ i obohacený dataset byly následně namapovány čtení získaná v kapitole 3.4 Filtrování dat. K mapování byl použit program Bowtie2 (Langmead a Salzberg, 2012) s následujícími parametry: *--very-sensitive-local* (lokální mapování s vysokou přesností), *--al-conc* (vytvoří soubor s čtením, které je mapované konkordantně), *--un-conc* (vytvoří soubor s čtením, které je mapované diskondantně). Konkordantně mapovaná čtení splňují očekávanou vzdálenost mezi sebou, je-li tato vzdálenost porušena, čtení se mapuje diskordantně. Ze vzniklých konkordantních čtení byly pomocí programu

seqtk (Shen a kol., 2016) vytvořeny subsety čtení odpovídající různé velikosti pokrytí genomu dle informací z manuálu k nástrojům použitým k identifikaci repetic (viz následující sekce).

3.11 Analýza repetitivních sekvencí

Druhově specifické datasety ze čtení mapujících na „assembly“ a na obohacený dataset byly analyzovány třemi programy, identifikujícími repetice a potenciální telomerické sekvence. Jako první byl využit již zmíněný RepeatExplorer2 (viz kapitola 2.5) se stejnými parametry, ale bez nastavení pro komparativní analýzu. Velikost repliky odpovídala 0,5x velikosti genomu.

Pro vyhledání satelitů s krátkým monomerem byl použit Tandem Repeats Finder (Benson 1999). Vstupním souborem byla replika čtení o velikosti 0,2x velikost genomu. Čtení byla konkatenována do jedné pseudomolekuly a vzájemně oddělena uměle vloženou sekvencí „NNNNNNNNNNNNNNNNNNNNNN”. Při zadávání parametru *trf* byla zvolena maximální délku monomeru 20 páru bazí. Výstupní soubory z této analýzy ve formátu html byly pomocí programu TRAP (Sobreira a kol., 2006) zpracovány do tabulky, ve které jsou repetitivní sekvence seřazeny od nejabundantnější po nejméně abundantní.

Kandidátní telomerické motivy byly identifikovány programem BioSERF (https://bioserf.org/; Somanathan a Baysdorfer, 2018). Vstupním souborem byla replika čtení o velikosti 0,2x velikost genomu. Minimální počet opakování motivu byl stanoven na tři (*MinBPRepeats=3*). Z kapacitních důvodu serveru byla analýza byla provedena třikrát pro separátní hledání motivů o délce 1-5 bp, 6-10 bp a 10-15 bp.

3.12 Lokalizace kandidátních motivů

Ověření fyzické lokalizace kandidátních sekvencí bylo provedeno u pavouka třesavka velká (*Pholcus phalangioides*; Füssli, 1775), který byl sbírán na území přírodovědecké fakulty Jihočeské univerzity, budova B (48.9775658N, 14.4453183E) v období července a srpna 2022. Vzhledem k tomu, že telomerická sekvence se mezi pohlavími neliší, byly sbíráni samci i samice tohoto druhu.

Před pitvou byli pavouci vloženi na zhruba 10 minut do mrazícího boxu o teplotě -21 °C pro znehybnění. Podchlazeným pavoukům byl odstřihnut abdomen od zbytku tělíčka. Abdomen byl pitván ve fyziologickém roztoku (Glaser 1917 cit. v Lockwood,

1961; 0,9% NaCl, 0,042% KCl, 0,025% CaCl2, 0,02% NaHCO3) a poté z něj byly připraveny chromozomální preparáty podle protokolu Herrera a kol. (2021) „spreading" (roztírací) technikou. Ovaria ze samic a testes ze samců byla vložena do roztoku KCl na 20-25 minut, aby došlo rozvolnění chromozomů. Hypotonizované gonády byly poté fixovány ve fixáži složené z etanolu a kyseliny octové v poměru 3:1 po dobu 10 minut. Po uplynutí této doby se fixáž vyměnila za čerstvou a inkubace pokračovala dalších 20 minut. Následně byla tkáň přesunuta do kapky 60% kyseliny octové na podložním skle očištěném kyselým etanolem (1% HCl v etanolu) a macerována wolframovými jehlami. Vzniklá suspenze byla poté rozetřena na podložním sklíčku na histologické ploténce zahřáté na 45°C. Následovalo odvodnění v etanolové řadě (70%, 80%, 100%) po dobu 30 sekund. Vzniklé preparáty byly uchovávány při -21 °C. Již vytvořené chromozomální preparáty z kontrolního druhu motýla předivky zhoubné (*Yponomeuta evonymella*; Linné, 1758) byly poskytnuty kolegyní Ksenyí Bobryshavou.

Fyzická lokalizace kandidátních motivů byla provedena podle protokolu Cuadrado a kol. (2009). Hybridizační směs pro každý preparát se skládala z 29,4 µl 2x SSC a 0,6 µl 100 µmol oligonukleotidů, značených na 5' konci fluorescenčním barvivem Cy3 (Generi Biotech, Hradec Králové, Česká republika). Z důvodu rozdílné úspěšnosti komerční syntézy jednotlivé sondy obsahovaly kadidátní motivy v různém počtu monomerů, a to konkrétně: $(A)_{35}$, $(AT)_{20}$, $(GT)_{20}$, $(TATGTA)_{11}$, $(TTTGTA)_{11}$ a $(TTAGG)_{12}$. Připravená hybridizační směs byla denaturována v 80°C po dobu 5 minut a poté přesunuta na led, kde se chladila 10 minut. Směs byla následně nanesena na chromozomální preparát, který se překryl krycím sklíčkem o rozměru 24x40 mm. Preparáty se poté inkubovaly ve vlhké komůrce při teplotě 37°C po dobu 2 hodin. Po inkubaci se krycí sklíčko z preparátu odstranilo a nezhybridizovaná sonda byla odmyta roztokem 4x SSC v 0,2% TWEEN-20 (Sigma-Aldrich, St. Loius, USA) po dobu 10 minut a poté roztokem 4x SSC v 0,1% TWEEN-20 (Sigma-Aldrich) po dobu 5 minut. Preparáty byly následně odvodněny v etanolové řadě (70%, 80%, 100%; po 3 minutách). Suché preparáty byly obarveny pomocí 10 µl fluorescenčního barviva DAPI (4,6-Diamidino-2-phenylindol) v DABCO antifade (1,4-Diazobicyclo-[2,2,2-octane]; koncentrace 0,5 ng DAPI/ ml DABCO) a překryty krycím sklíčkem o velikosti 24x32 mm. Krycí sklíčko se zafixovalo pomocí laku na nehty, čímž bylo zároveň zabráněno i vyschnutí preparátu. Preparáty byly uchovávány v 5 °C. Preparáty byly pozorované na konfokálním mikroskopu Olympus FV3000 (Olympus, Tokyo, Japonsko), fotky z mikroskopu byly upraveny pomocí programu Fiji (Schindelin et al., 2012).

4. Výsledky

4.1 Sekvenační data pavouků

Z databáze Darwin Tree of Life byly získány datasety sekvenované metodou „Linked-Reads“ ze 3 druhů pavouků – třesavka velká (*Pholcus phalangioides*; Pholcidae), zápředník tmavý (*Clubiona reclusa*; Clubionidae) a lovčík hajní (*Pisaura mirabilis;* Pisauridae). V databázi National Center for Biotechnology Information byly k dispozici data ze čtyř zástupců rodu Nefila z čeledi Araneidae. Pro účely této práce byl stažen pouze jeden dataset reprezentující tuto skupinu, konkrétně data ze druhu nefila chluponohá (*Nephila pilipes*; identifikační číslo v databázi GenBank DRX230504). Jako kontrolní druh byl vybrán motýl předivka zhoubná (*Yponomeuta evonymella*; Yponomeutidae), jehož telomerická sekvence je známá a jedná se o mikrosatelitní motiv $(TTAGG)_n$. Tento dataset byl stažen z databáze DToL.

Pomocí nástroje FastQC byla analyzována kvalita a množství sekvenačních dat v jednotlivých datasetech (Tabulka 1). Tento nástroj obsahuje i modul k identifikaci nadreprezentovaných sekvencí, v jehož výstupu se u datasetu motýla *Yponomeuta evonemylla* objevila telomerická sekvence $(TTAGG)_n$. Podobný výsledek byl pozorován u datasetu *Clubiona reclusa*, kde byla přítomna nadreprezentovaná sekvence $(TTAGGTTGGGG)_n$. U žádného jiného pavoučího datasetu se však nadreprezentované sekvence nevyskytovaly.

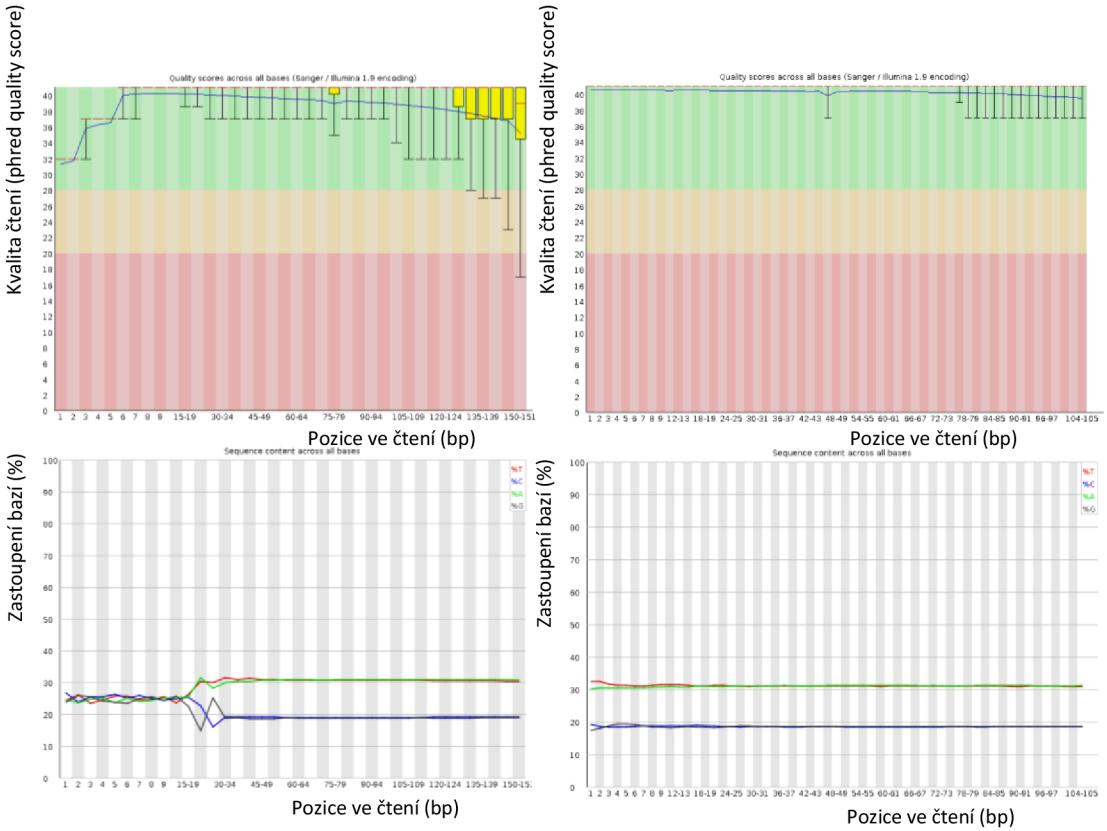
Tabulka 1: Základní vlastnosti použitých datasetů. V tabulce je znázorněn množství dat v jednotkách giga páru bazí (Gbp), průměrná kvalita v jednotkách „phred score“, přítomnost nadreprezentovaných sekvencí a jejich množství v 5 datasetech – *Clubiona reclusa, Pholcus phalangioides, Pisaura mirabilis, Nephila pilipes a Yponomeuta evonymella.* Délka čtení ve všech datasetech činí 151 páru bazí. „Phred quality score“ nad hodnotu 20 je již akceptovatelná kvalita, při které je pouze 1% čtení chybných. n.d. – nadreprezentovaná sekvence nebyla detekována.

Dataset	**Množství dat (Gbp)**	**Průměrná kvalita („phred score“)**	**Nadreprezentované sekvence**	**Průměrný výskyt nadreprezentované sekvence (%)**
Clubiona reclusa	76	39	$(TTAGGTTGGGG)_n$	0,0115
Pholcus phalangioides	140	39	n.d.	n.d.
Pisaura mirabilis	136	39	n.d.	n.d.
Nephila pilipes	136	34	n.d	n.d.
Yponomeuta evonymella	64	39	$(TTAGG)_n$	0,05

Ačkoliv byla získaná data v poměrně vysoké kvalitě, bylo i přesto třeba je pro další analýzy vyfiltrovat. Pomocí programu Trimmomatic byly úspěšně odstraněny technické sekvence a čtení o špatné kvalitě (Tabulka 2). Příklad datasetu před a po odstranění čtení o špatné kvalitě lze pozorovat na Obrázku 3a-b a úspěšné odstranění technických sekvencí na Obrázku 3c-d.

Tabulka 2: Základní vlastnosti datasetů po odstranění čtení o nízké kvalitě a technických sekvencí. Množství dat (Gbp), průměrná kvalita v jednotkách „phred quality score“, přítomnost nadreprezentované sekvence a množství nadreprezentované sekvence v 5 datasetech – *Clubiona reclusa, Pholcus phalangioides, Pisaura mirabilis, Nephila pilipes* a *Yponomeuta evonymella*. Průměrná délka čtení ve všech datasetech činí 105 páru bazí.

Dataset	**Množství dat (Gbp)**	**Průměrná kvalita („phred quality score“)**	**Nadreprezentované sekvence**	**Průměrný výskyt nadreprezentované sekvence (%)**
Clubiona reclusa	37	40	n.d.	n.d.
Pholcus phalangioides	78	40	n.d.	n.d.
Pisaura mirabilis	77	40	n.d.	n.d.
Nephila pilipes	64	35	n.d.	n.d.
Yponomeuta evonymella	29	40	$(TTAGG)_n$	0,027 %



Obrázek 3: Znázornění kvality čtení a přítomnosti technických sekvencí u nefiltrovaných (*a*,*c*) a filtrovaných (*b*,*d*) dat na příkladu datasetu pavouka *Pholcus phalangioides*. Obrázek *a,b* znázorňuje kvalitu čtení - zelená barva představuje čtení o nejlepší kvalitě, oranžová barva představuje čtení o střední kvalitě, červená barva představuje čtení o špatné kvalitě. Krabicové grafy a jejich poloha ukazuje kvalitu čtení. Osa X představuje pozici páru bazí v čtení osa Y představuje kvalitu čtení v jednotkách „phred quality score“. Na obrázku *b* lze pozorovat posun krabicových grafů do zelené oblasti, což je známka zlepšení kvality. Obrázek *c,d* znázorňuje přítomnost technických sekvencí - zelená křivka znázorňuje procentuální zastoupení báze adeninu, červená křivka báze thyminu, černá křivka báze guaninu a modrá křivka báze cytosinu. Osa X představuje pozici páru bazí v čtení, osa Y představuje procentuální zastoupení bazí. Na obrázku *d* lze pozorovat snížení nepravidelnosti a klikatosti křivek v pozici 1-30 bp, což značí úspěšné odstranění technických sekvencí. Délka čtení před filtrováním činí 151 bp, po filtrování 105 bp.

4.2 Komparativní analýza repetic

Vzhledem k tomu, že telomerické motivy bývají ve vývojových liniích konzervované, byla provedena analýza repetic ve které byly porovnány 4 datasety pavouků – *Clubiona reclusa,*

Pholcus phalangioides, Pisaura mirabilis a *Nephila pilipes*. Komparativní analýza identifikovala všechny repetice u studovaných druhů. Celkem bylo anotováno 195 klastrů, z čehož pouze 6 klastrů obsahovalo čtení ze všech datasetů (Tabulka 3). Pouze jeden klastr (klastr 32) byl tandemová repetice, ale obsahoval větší množství čtení z jediného druhu pavouka, a to *Nephila pilipes*. Klastry 10 a 18 tvořily stejný superklastr a byly anotovány jako ribozomální DNA (rDNA). Klastr 8 neměl žádnou anotaci a po bližším prozkoumání obsahoval sekvence s nízkou komplexitou (homopolymery, dimery atd.). Klastr 91 obsahoval sekvenci $(TTAGGTTGGGG)_n$, která však byla přítomna pouze ve čtení z datasetu *Clubiona reclusa* a formoval superklastr s klastrem 13, tvořeným mobilní element typu LINE Ty3/Gypsy (retrotranspozon). Výsledky této analýzy jsou shrnuty v Tabulce 3.

Tabulka 3: Přehled klastrů společných pro všechny 4 datasety pavouků

Klastr	Tandemová repetice	Vlastnosti klastru
8	ne	Neznámá repetice bez anotace o nízké komplexitě (homopolymery, dimery…)
10	ne	Částečná shoda s rDNA, součástí superklastru s klastrem č. 18
13	ne	LINE Ty3/Gypsy, součástí superklastru s klastrem č. 91,
18	ne	Částečná shoda s rDNA, součástí superklastru s klastrem č. 10,
32	ano	Větší množství čtení pouze u pavouka *Nephila pilipes*
91	ne	Součástí superklastru s klastrem č.13, obsahoval nadreprezentovanou sekvenci $(TTAGGTTGGGG)_n$

Tato analýza byla z důvodu uvedeného v následující kapitole zopakována i bez datasetu *Clubiona reclusa*. Ačkoliv bylo identifikováno sedm dalších klastrů, žádný z nich nebyl vyhodnocen jako tandemová repetice. Jednalo se převážně o neznámé repetitivní elementy bez anotace (klastry 17, 43, 54, 87, 92 a 190) a sekvenci obsahující parciální shodu s rDNA (klastr 56).

4.3 Rekonstrukce genomové sekvence

Protože komparativní analýza repetic neodhalila žádné další kandidátní telomerické motivy napříč skupinou, bylo přistoupeno k práci na úrovni jednotlivých druhů. K tomu bylo nezbytné zjistit velikost genomu zkoumaných druhů a následně zrekonstruovat genomovou sekvenci.

U datasetu *Clubiona reclusa,* který obsahoval potenciální telomerický motiv $(TTAGGTTGGGG)_n$, byla velikost genomu v haploidním stavu odhadnuta na 243 702 746 páru bazí (Tabulka 4). Celková délka složené genomové sekvence činila 241 642 818 páru bazí. Zhotovené „assembly" bylo tvořené 2 326 kontigy, z čehož 459 kontigů bylo větších než 5 000 páru bazí a číslo N50 bylo 8 871 374. Výsledky programu BUSCO prozradily, že „assembly" je kompletní z 80,6%. 2,5% BUSCO genů bylo nalezeno fragmentovaných a 16,9% v „assembly" zcela chybělo (Tabulka 5). Při kontrole molekulární identity datasetu však porovnání sekvence genu *cytochrom oxidáza I* odhalilo, že sekvenovaná data nepochází z pavouka *Clubiona reclusa* a dataset byl zřejmě v databázi chybně označen. S nejvyšší pravděpodobností se jedná o data ze čmeláka *Bombus hortorum*, který se v sekvenci *cytochrom oxidázy I* shoduje s vytvořeným „assembly" s 99,7% identitou. Tento dataset byl tedy z dalšího výzkumu vyloučen.

U náhradního druhu, pavouka *Pholcus phalangioides,* byla velikost genomu analyzována s výsledkem 943 638 241 páru bazí (Tabulka 4). Celková délka „assembly" činila 1 252 395 564 páru bazí a složený genom se skládal z 61 816 kontigů. Kontigů větších než 5 000 páru bazí bylo 15 031 a číslo N50 dosáhlo 2 944 798. BUSCO analýza prozradila, že „assembly" bylo kompletní z 68,6%. 5,5 % nalezených BUSCO genů bylo fragmentovaných a 25,9% v genomové sekvenci chybělo (Tabulka 5). Ověření sekvence *cytochrom oxidázy I* potvrdilo, že se jedná o správný druh, tedy pavouka *P. phalangioides*.

Velikost genomu motýla *Yponomeuta evonymella* byla odhadnuta na 456 766 665 párů bazí (Tabulka 4). Celková délka „assembly" vytvořeného kolegyní Ksenyí Bobryshavou činila 851 627 286 páru bazí a byla tvořena 39 263 kontigy. Kontigů větších než 5 000 páru bazí bylo 12 392. Dle výstupu nástroje BUSCO bylo zjištěno, že „assembly" je kompletní z 85,5%, 5,4% nalezených BUSCO genů bylo fragmentovaných a 9,1% v „assembly" chybělo (Tabulka 5).

Tabulka 4: Odhad velkosti genomu zkoumaných druhů pomocí programu GenomeScope2. U datasetů *Clubiona reclusa*, *Pholcus phalangioides* a *Yponomeuta evonymella* byla kromě velikosti odhadnuta míra i heterozygozity a homozygozity, velikosti haploidního genomu a celková úspěšnost předpovídaného modelu.

Dataset	Heterozygozita (%)	Homozygozita (%)	Velikost haploidního genomu (bp)	Úspěšnost modelu (%)
Clubiona reclusa	0,05	99,96	243 702 746	98,42
Pholcus phalangioides	0,98	99,09	943 638 241	98,31
Yponomeuta evonymella	1,24	98,82	456 766 665	98,59

Tabulka 5: Vlastnosti složených genomových sekvencí u datasetů *Clubiona reclusa*, *Pholcus phalangioides* a *Yponomeuta evonymella*. Mezi hlavní charakteristiky „assembly" patří celková délka zrekonstruované sekvence, počet fragmentů (kontigů) ze kterých je tvořená a číslo N50. Tato hodnota udává délku kontigu, který by dosáhl poloviny délky genomové sekvence při seřazení kontigů od nejdelšího po nejkratší. Kompletnost „assembly" byla ověřena pomocí identifikace konzervovaných ortologů nástrojem BUSCO.

Dataset	Celková délka „assembly" (bp)	Počet kontigů	Počet kontigů >5000 bp	N50	Kompletní BUSCO geny (%)	Fragmentované BUSCO geny (%)	Chybějící BUSCO geny (%)
Clubiona reclusa	241 642 818	2 326	459	8 871 374	80,6 %	2,5 %	16,9 %
Pholcus phalangioides	1 252 395 564	61 816	15 031	2 944 798	68,6 %	5,5 %	25,9 %
Yponomeuta evonymella	851 627 286	39 263	12 392	199 298	85,5 %	5,4 %	9,1 %

4.4 Identifikace kandidátních telomerických motivů

Ze zhotoveného „assembly“ byly pomocí skriptu v programovacím jazyce Python izolovány veškeré kontigy delší než 5 000 párů bazí. Z nich byly ponechány pouze jejich konce o délce 5 000 páru bazí, které byly doplněny kontigy kratšími než 5 000 párů bazí. V „assembly“ pavouka *Pholcus phalangioides* bylo 15 031 kontigů větších než 5 000 páru bazí. U motýla *Yponomeuta evonymella* se jednalo o 12 392 kontigů.

Mapováním sekvenačních čtení na „assembly“ a upravený set kontigů byly získány pro každý druh dva datasety – standardní a obohacený. U pavouka *Pholcus phalangioides* mapovalo na „assembly“ 97,01% sekvenačních čtení, zatímco na upravené kontigy pouze 57,47%. U motýla *Yponomeuta evonymella* bylo dosaženo podobných hodnot, tedy 98,11% čtení mapujících na „assembly“ a 58,52% na upravené kontigy. Získané datasety sloužily jako vstupní soubory do tří programů k analýze repetitivních sekvencí – RepeatExplorer2, Tandem Repeats Finder a BioSERF.

Analýza repetic pomocí programu RepeatExplorer2 u pavouka *Pholcus phalangioides* v obohaceném datasetu odhalila celkem 566 klastrů, zatímco ve standardním datasetu pouze 508 klastrů. V obohaceném datasetu bylo s vysokou jistotou identifikováno 29 tandemových repetic, z nichž 2 byly mobilní elementy typu LTR Ty3/gypsy tvořené tandemovým opakováním, avšak s velmi nízkou abundancí 0,086 % a 0,085%. Pouze 7 satelitů mělo odhadovanou abundanci vyšší než 0.1% a délka jejich monomeru se pohybovala od 28 do 2 321 párů bazí (Tabulka 6). U motýla *Yponomeuta evonymella* analýza obohaceného datasetu odhalila 317 klastrů a analýza standardního datasetu 496 klastrů, přičemž žádný z těchto klastrů neobsahoval telomerickou sekvenci $(TTAGG)_n$. Přehled identifikovaných tandemových repetic s abundancí alespoň 0.1 % v obohaceném datasetu je shrnut v tabulce 6.

Tabulka 6: Přehled tandemových repetic identifikovaných programem RepeatExplorer2 s abundancí alespoň 0,1%. U pavouka *Pholcus phalangioides* bylo identifikováno 7 satelitů, zatímco u motýla *Yponomeuta evonymella* pouze jeden. Telomerická sekvence u *Y. evonymella* nebyla tímto programem detekována.

Druh	Dataset	Klastr	Délka monomeru (párů bazí)	Zastoupení sekvence (%)	Zastoupení sekvence ve standardním datasetu (%)
Pholcus phalangioides	obohacený	1	28	1,6	0,32
		2	186	0,47	0,24
		3	28	0,27	0,13
		4	177	0,18	0,089
		5	361	0,12	0,058
		6	2321	0,11	0,058
		7	178	0,1	0,051

Na rozdíl od RepeatExplorer2 je program Tandem Repeats Finder vhodný pro hledání satelitů s krátkým monomerem. U pavouka *P. phalangioides* patřily mezi nejabundantnější repetice identifikované jak ve standardním, tak v obohaceném datasetu mikrosatelity s velmi krátkými motivy, jako například $(CT)_n$, $(TA)_n$, a $(TG)_n$. Následovaly satelity s delším monomerem, nicméně žádný neodpovídal vzorci typickému pro telomery, tedy $T_nA_nG_n$ (Tabulka 7). U motýla *Y. evonymella* se telomerický motiv nacházel na prvním místě jak ve standardním, tak v obohaceném datasetu, přičemž jeho množství bylo v obohaceném datasetu větší (Tabulka 7).

Tabulka 7: Tandemové repetice identifikované programem Tandem Repeats Finder. Uvedeno je 10 nejabundantnějších motivů pro každý druh. Zatímco u pavouka *P. phalangioides* nebyla pozorována žádná sekvence odpovídající typickému telomerickému vzorci $T_nA_nG_n$, u motýla *Yponomeuta evonymella* je telomerická sekvence $(TTAGG)_n$ nejabundantnější repeticí jak ve standardním tak v obohaceném datasetu. V obohaceném datasetu je její abundance vyšší než ve standardním.

Druh	Dataset	Sekvence	Celkový počet bazí	Celkový počet lokusů	Počet opakování monomeru za sebou	Průměrný počet opakování monomeru za sebou v lokusu
Pholcus phalangioides	standardní	TA	123	4	61,5	15,37
		CT	156	4	78,5	19,62
		AATA	104	3	26	8,66
		TG	237	3	118,5	39,5
		TCACTGACTCG GTCACTCAA	237	3	11,7	3,9
		TGAATAAATA AATAAATAAA	80	2	4	2
		AAAACCATTA	68	2	6,8	3,4
		TTATCAGTAAC TCAGT	32	1	2	2
		ACTGAGATAC TGATAA	32	1	2	2
		CATATTTTG	83	1	8,9	8,9
Pholcus phalangioides	obohacený	TG	639	9	319,5	35,5
		TA	174	6	87	14,5
		T	479	5	479	95,8
		TTTGTA	125	4	20,8	5,2
		ACTCGGTCACT CGCTCAATC	208	2	10,6	5,3
		GCGGTGAATT GTTG	140	2	9,6	4,8
		AATTCACCGCT AGGAA	184	2	11,8	5,9
		CACTCACTAG ATCACT	72	2	4,5	2,25
		TATGTA	198	2	35	17,5
		TTTCATTAATT CTGAAATTA	53	1	2,6	2,6
Yponomeuta evonymella	standardní	TTAGG	6040	61	1208	19,80
		GT	551	9	275,8	30,64
		GACA	305	5	76,3	15,26
		CAT	94	3	31,4	10,46

		GTGTGTGTGTGC	134	2	11	5,5
		GTGC	123	2	30,7	15,35
		TGTGTC	208	2	34,7	17,35
		AGGT	56	1	14,2	14,2
		GTTAAACGTGGTGCCAT	54	1	3,2	3,2
		CTAAT	35	1	7,2	7,2
Yponomeuta evonymella	obohacený	TTAGG	7794	76	1559	20,51
		TG	348	4	174	43,5
		CATAA	420	4	84	21
		CTGT	275	3	68,6	22,86
		ATCT	176	3	43	14,33
		A	160	2	160	80
		TTGAAT	202	2	33,7	16,85
		GTTA	210	2	52,4	26,2
		TC	68	2	34	17
		CTAC	86	1	21,8	21,8

Specifická predikce telomerických motivů byla provedena pomocí programu BioSERF. U pavouka *P. phalangioides* opět k nejčastnějším motivům patřily velmi krátké mikrosatelity, zejména $(A)_n$ a $(TA)_n$ (Tabulka 8). Dále se jednalo o jeden delší motiv, $(TTTGTA)_n$, který byl pozorován také v předchozí analýze, konkrétně v obohaceném datasetu analyzovaném nástrojem Tandem Repeats Finder (Tabulka 7). U motýla *Y. evonymella* byla telomerická sekvence identifikována pouze v obohaceném datasetu, kde patřila k nejabundantnějším sekvencím (Tabulka 8).

Tabulka 8: Kandidátní telomerické motivy identifikované programem BioSERF. Datasety byly analyzovány na přítomnost kandidátních motivů, které měly délku 1-5, 6-10 a 11-15 párů bazí. U pavouka *P. phalangioides* k nejčastějším motivům patřily mikrosatelity $(A)_n$, $(TA)_n$ a $(TTTGTA)_n$. U datasetu motýla *Yponomeuta evonymella* lze u obohaceného datasetu pozorovat telomerickou sekvenci $(TTAGG)_n$.

Druh	Dataset	Délka monomeru (párů bazí)	Počet opakování sekvence	Sekvence
Pholcus phalangioides	standardní	1-5	>50 000	$(A)_n$ $(TA)_n$
		6-10	>12 000	$(TTTGTA)_n$
		11-15	-	-
	obohacený	1-5	>600 000	$(A)_n$ $(TA)_n$
		6-10	>2 000	$(TACAAA)_n$ $(TTTGTA)_n$
		1-15	>100	$(A)_n$ $(C)_n$
Yponomeuta evonymella	standardní	1-5	>2 000 000	$(TA)_n$ $(A)_n$
		6-10	>1500	$(A)_n$ $(ACGCAC)_n$ $(ATGTTT)_n$ $(TGTGCG)_n$
		11-15	>200	$(A)_n$
	obohacený	1-5	>200 000	$(A)_n$ $(TTAGG)_n$ $(TA)_n$
		6-10	>800	$(ACATAA)_n$ $(CATACA)_n$ $(TATGTT)_n$
		11-15	>80	$(A)_n$ $(TTTCATTTTCA)_n$ $(AAATGAAAATG)_n$

Souhrn získaných mikrosatelitních motivů je uveden v Tabulce 9. Analýza repetic pomocí nástroje RepeatExplorer2 neodhalila žádný kandidátní motiv tohoto typu a jeho výstup tedy není v tabulce uveden.

Tabulka 9: Přehled nejabundantnějších kandidátních motivů z analýzy Tandem Repeats Finder a BioSERF. Tučně jsou zvýrazněny repetice vybrané pro fyzické mapování.

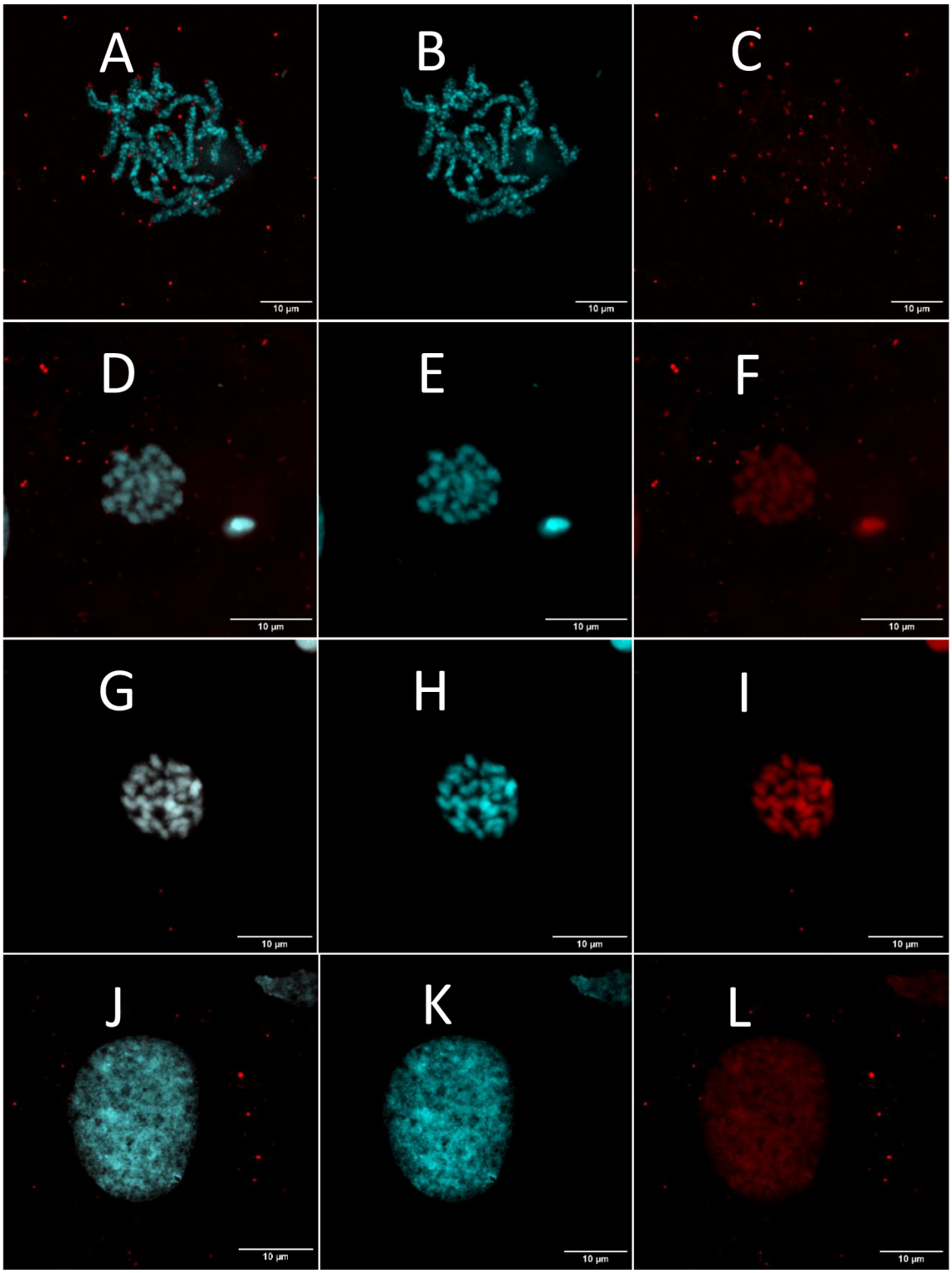
Druh	Dataset	Kandidátní sekvence Tandem Repeats Finder	Kandidátní sekvence BioSERF
Pholcus phalangioides	mapovaný na „assembly“	**TA**, CT, AATA, **TG**, TCACTGACTCGGTCACTCAA, TGAATAAATAAATAAATAAA, AAAACCATTA, TTATCAGTAACTCAGT, ACTGAGATACTGATAA, CATATTTTG	**A**, **TA**, **TTTGTA**
	mapovaný na konce	**TG**, **TA**, **T**, **TTTGTA**, ACTCGGTCACTCGCTCATCGCG GTGAATTGTTG, AATTCACCGCTAGGAA, CACTCACTAGATCACT, **TATGTA**, TTTCATTAATTCTGAAATTA	**A**, **TA**, **TTTGTA**, C
Yponomeuta evonymella	mapovaný na „assembly“	**TTAGG**, GT, GACA, CAT, GTGTGTGTGTGC, GTGC, TGTGTC, AGG, T, GTTAAACGTGGTGCCAT, CTAAT	TA, A, ACGCAC, ATGTTT, TGTGCG
	mapovaný na konce	**TTAGG**, TG, CATAA, CTGT, ATCT, A, TTGAAT, GTTA, TC, CTAC	A, **TTAGG**, AT, ACATAA, CATACA, TATGTT, TTTCATTTTCA, AAATGAAAATG

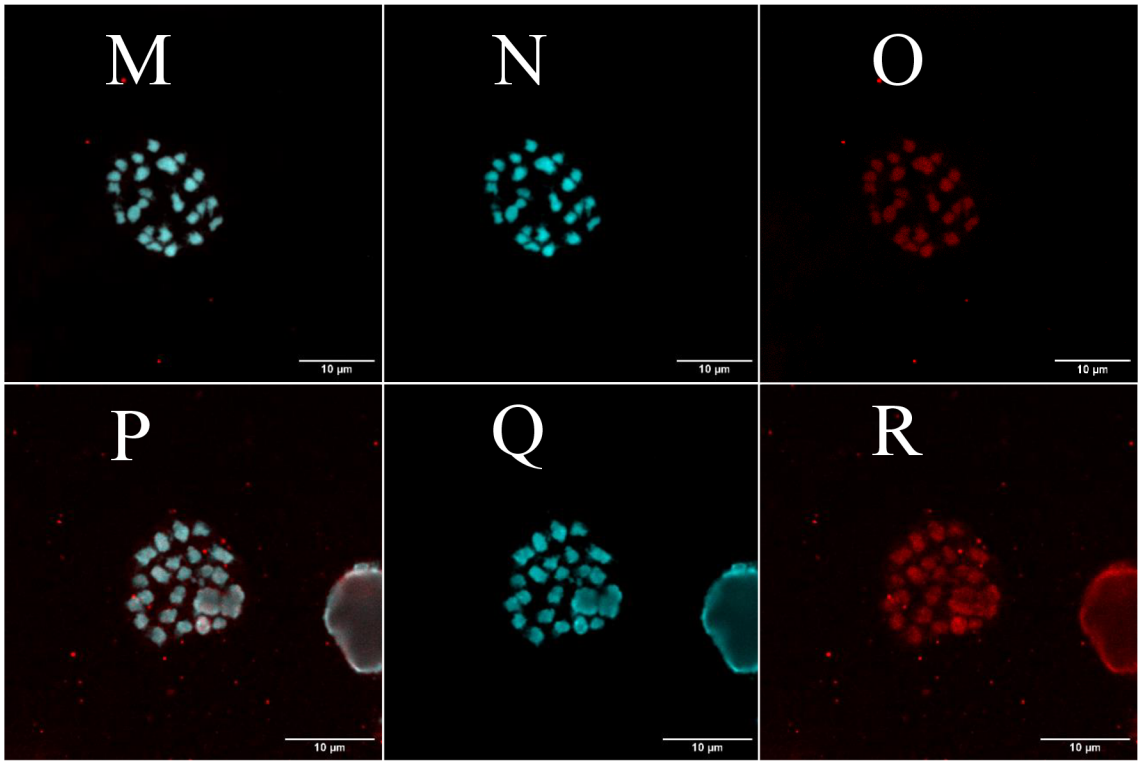
4.5 Lokalizace vybraných motivů na chromozomech

Kandidátní telomerické motivy u *P. phalangioides* $(TTTGTA)_n$, $(TATGTA)_n$, $(A)_n$, $(TA)_n$, $(TG)_n$ a kontrolní telomerická sekvence u *Y. evonymella* $(TTAGG)_n$ byly nasyntetizovány jako fluorescenčně značené oligonukleotidy a nahybridizovány na chromozomální preparáty příslušného druhu metodou nedenaturující fluorescenční *in situ* hybridizace (FISH).

U motýla *Yponomeuta evonymella* sonda úspěšně označila konce pachytenních chromozomů a navzdory pozadí byl signál jasně lokalizován v telomerické oblasti (Obrázek 4A-C). U pavouka *Pholcus phalangioides* se terminální lokalizace žádného

z kandidátních motivů na chromozomech nepotvrdila a sondy značily celé chromozomy. Detekce navíc byla ztížená přítomností artefaktů způsobených barvením DAPI (Obrázek 4N, Q) a nízkým počtem chromozomálních figur na preparátech. Výsledky analýzy FISH jsou znázorněny na Obrázku 4.


A
B
C
D
E
F
G
H
I
J
K
L
10 µm



Obrázek 4: Lokalizace vybraných kandidátních telomerických motivů metodou nedenaturující fluorescenční *in situ* hybridizace (FISH). Modrá barva představuje chromatin barvený DAPI, červená sondy kandidátních sekvencí značené fluorescenční barvou Cy3. Obrázky v levém sloupci značené písmeny A, D, G, J, M, N obsahují složený obraz DAPI a sondy. Obrázky v prostředním sloupci B, E, H, K, N, Q obsahují pouze DAPI barvený chromatin bez přítomnosti sondy. Obrázky v pravém sloupci obsahují chromosomální preparáty pouze s navázanou sondou. Obrázky označené písmeny A, B, a C náleží motýlu *Yponomeuta evonymella*. Obrázek A značí přítomnost úspěšně nahybridizované telomerické sondy $(TTAGG)_n$ na pachytenním jádře, sloužící jako pozitivní kontrola. Signál je jasně lokalizován v terminální oblasti chromozomů. Obrázky D-R zobrazují chromozomální preparáty pavouka *Pholcus phalangioides* s nahybridizovanými sondami $(TTTGTA)_n$ (obr. 4D-F) a $(TATGTA)_n$ (obr. 6G-I) na mitotických chromozomech, $(A)_n$ na interfázním jádře (obr. 4J-I), a sondy $(TA)_n$ (obr. 4M-O) a $(TG)_n$, (obr. 4P-R) na mitotických chromozomech. Lokalizace kandidátních motivů v telomerách se nepotvrdila a sondy značily celé chromozomy.

5. Diskuse

Pavouci jsou jednou z mála vývojových linií, u nichž telomerická sekvence zůstává neznámá. U pavoukům blízce příbuzné skupiny Amblypygi (bičovci), patřící mezi Arachnida (pavoukovci), můžeme nalézt telomerický motiv s motivem $(TTAGG)_n$ (Lerma a kol., 2021). Předchozí výzkum ukázal, že pavoučí telomery neobsahují mikrosatelitní telomerickou sekvenci typickou pro Arthropoda s motivem $(TTAGG)_n$, ani sekvenci typickou pro Metazoa s motivem $(TTAGGG)_n$ (Vítková a kol., 2005). Cílem této bakalářské práce bylo telomerickou sekvenci pavouků odhalit, a to s použitím již osekvenovaných dat veřejně dostupných v databázích jakou jsou Darwin Tree of Life (DToL; www.darwintreeoflife.org) a National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov).

Prvním krokem bylo získat vhodná data. Vybrána byla data osekvenovaná metodou „Linked-Reads" od 10x Genomics. Tento druh dat byl zvolen k analýze telomerických sekvencí na základě předchozích zkušeností školitelky a vedoucího laboratoře, protože u jiných druhů nemodelových organismů v datech osekvenovaných touto metodou pozorovali telomerickou sekvenci s větší četností, než u standardních Illumina dat (A. Voleníková a P. Nguyen, osobní komunikace). Z DToL jsme získali sekvenační data ze 3 druhů pavouků – *Pholcus phalangioides* (třesavka velká), *Clubiona reclusa* (zápředník tmavý) a *Pisaura mirabilis* (lovčík hajní). Z databáze NCBI byla získána sekvenační data pavouka *Nephila pilipes* (nefila chluponohá). Tyto 4 druhy pavouků reprezentují 4 fylogenetické větve v rámci infrařádu Araneomorphae (Wheeler a kol., 2016) a zároveň jsou od sebe značně fylogeneticky vzdálení, čímž je zajištěné dostatečné fylogenetické pokrytí skupiny. Z infrařádu Mygalomorphae bohužel žádný zástupce touto metodou osekvenovaný nebyl. Pozitivní kontrolou byl zvolen dataset motýla *Yponomeuta evonymella*, jehož telomerická sekvence s motivem $(TTAGG)_n$ je dobře známá (Provazníková, 2023).

Při kontrole kvality získaných dat v programu FastQC byla u *Y. evonymella* pozorována nadreprezentovaná sekvence $(TTAGG)_n$, odpovídající telomerickému motivu u tohoto druhu. Podobná situace byla u datasetu pavouka *Clubiona reclusa*, kde se objevil motiv $(TTAGGTTGGGG)_n$, částečně odpovídající obecnému telomerickému vzorci $T_nA_nG_n$, a tedy představující kadidátní telomerickou sekvenci. U žádného jiného pavouka nadreprezentovaná sekvence pozorovaná nebyla. Důvodů, proč se kandidátní sekvence objevila pouze u některých datasetů, může být několik. Použitý modul analyzuje 100 000 sekvenačních čtení a vyskytuje-li se v nich určitá sekvence alespoň z 0,1%, pak ji

program vyhodnotí jako nadreprezentovanou. Přítomnost sekvence ve výstupu tohoto programu je tedy ovlivněna náhodným výběrem čtení a abundancí dané sekvence v genomu. V případě telomerické sekvence hraje roli délka telomer, počet chromozomů a velikost samotného genomu, kdy s větším počtem chromozomů a s delšími telomerami roste i obsah telomerické sekvence v datasetu a naopak.

Na základě znalostí o konzervovanosti telomerických motivů v rámci vývojových linií (Vítková a kol. 2005, Gomes a kol. 2010) byla provedena komparativní analýza repetic u všech získaných druhů pavouků s předpokladem, že tandemová repetice společná pro všechny datasety by představovala vhodný kandidátní telomerický motiv. Satelit přítomný u všech druhů se však podařilo identifikovat pouze jeden a po bližším prozkoumání bylo zjištěno, že je abundantní pouze u druhu *Nephila pilipes* (Tabulka 3). U ostatních datasetů se vyskytoval pouze v minimální míře (jednotky sekvenačních čtení). Další sekvence společné pro všechny datasety byly anotovány jako rDNA a neznámé repetice, které však netvořily tandemové uspořádání, očekávané pro telomery. Kandidátní sekvence $(TTAGGTTGGGG)_n$ získaná v předchozím kroku byla přítomna pouze v sekvenačních čtení z datasetu *Clubiona reclusa,* a to ve formě superklastru, kde se pojila s mobilním elementem specifickým pro tento druh.

Vzhledem k tomu, že komparativní analýza repetic neodhalila žádný kandidátní motiv, bylo přistoupeno k analýze repetic na úrovni druhu. Pro lepší detekci telomerické sekvence byl vytvořen *in silico* obohacený dataset, kdy se ze sekvenačních dat zrekonstruovala genomická sekvence a izolovaly se pouze konce složených fragmentů. U huseníčku (*Arabidopsis*) bylo zjištěno, že minimální délka funkčních telomer je 400 párů bazí (Watson a kol., 2021). Při působení BAL31 exonukleázy pro detekci telomer (podrobně popsáno v kapitole 1. Úvod) dochází k odtrávení přibližně 10 kilo párů bazí (Peška a kol. 2017). Délka izolovaných konců kontigů by se tedy měla pohybovat v rozsahu 400-10 000 párů bazí, v závislosti na kontinuitě získané genomické sekvence.

Z důvodu zjištěné přítomnosti nadreprezentované sekvence $(TTAGGTTGGGG)_n$, silně se podobající telomerické sekvenci $(TTAGG)_n$ byla předmětem dalšího studia zvolena zápřednice *Clubiona reclusa*, pro níž bylo vytvořeno genomové „assembly“. Při následné kontrole molekulární identity datasetu pomocí porovnání sekvence genu *cytochrom oxidáza I* však bylo zjištěno, že jeho sekvence se shoduje z 99,7% se sekvencí čmeláka *Bombus hortorum* (Hymenoptera). Sekvence zápřednice se v „assembly” vůbec nevyskytovala a nemohlo tedy jít o kontaminaci například z potravy. S nejvyšší pravděpodobností došlo k záměně sekvenovaného vzorku nebo dat, které byly chybně

nahrané do databáze. Původ dat ze čmeláka podporovala i nalezená nadreprezentovaná sekvence $(TTAGGTTGGGG)_n$, která odpovídá telomerickému motivu pro tuto skupinu (Fajkus a kol., 2022).

Po vyřazení datasetu *C. reclusa* ze studovaných druhů byla nejprve zopakována komparativní analýza repetic u zbývajících datasetů. Ani v tomto případě se však nepodařilo získat kandidátní telomerickou sekvenci, která by byla společná všem pavoukům a vykazovala tandemovou organizaci. Jako náhradní druh ke studiu byla vybrána třesavka *Pholcus phalangioides*, která je v České republice hojná a bylo tedy možné z ní získat chromozomální preparáty pro pozdější ověření lokalizace kandidátních motivů. Genomové „assembly" *P. phalangioides* se skládalo z 61 816 fragmentů s hodnotou N50 přibližně 3 mega párů bazí (Tabulka 5). Na základě těchto hodnot a informací z literatury uvedených výše byl připraven obohacený dataset z izolovaných konců kontigů o délce 5 000 párů bazí. Stejným způsobem bylo postupováno i v případě kontrolního druhu, motýla *Yponomeuta evonymella*.

Následná identifikace kandidátních repetic byla provedena na sekvenačních čtení, mapujících na „assembly" a obohacený dataset, a to pomocí tří různých bioinformatických nástrojů. Tyto programy byly vybrány tak, aby byly analyzovány tandemové repetice s dlouhým (RepeatExplorer2) i krátkým (Tandem Repeats Finder) monomerem a nezávisle predikovány telomerické motivy (BioSERF).

Analýza RepeatExplorer2 nalezla u pavouka *Pholcus phalangioides* množství satelitů, z nichž se však žádný nepodobal známým telomerickým motivům. Délka monomeru nejkratšího abundantního satelitu byla 28 párů bazí (Tabulka 6), přičemž u známých minisatelitních telomer se délka jednotky pohybuje pod touto hodnotou (Jenner a kol., 2022). Identifikovány byly i dva satelity, obsahující domény mobilních elementů. Telomery tvořené tandemovým opakováním retrotranspozonů byly popsány u vybraných druhů dvoukřídlého hmyzu (shrnuto v Kuznetsova a kol. 2020), nicméně abundance těchto repetic u *P. phalangioides* byla velmi nízká (<0.1%). Vzhledem k častému výskytu rDNA v subtelomerických oblastech (Roa a Guerra, 2012; Provazníková a kol.) byly podrobně prozkoumány i všechny klastry superklastru obsahujícího rDNA, žádny z nich však nebyl označený jako satelit. U motýla *Y. evonymella* RepeatExplorer2 telomerickou sekvenci nebyl schopný identifikovat, a to z důvodu příliš krátkého monomeru (Novák a kol., 2020).

V případě Tandem Repeats Finder a BioSERF se u pavouka mezi nejabundantnějšími motivy shodně objevovaly velmi krátké repetice, a to konkrétně $(TA)_n$, $(TG)_n$, $(A)_n$, jak u standardního, tak obohaceného datasetu (Tabulka 7 a Tabulka 8).

V obohaceném datasetu byly dále identifikovány dva delší velmi podobné kandidátní motivy, $(TTTGTA)_n$ a $(TATGTA)_n$ (Tabulka 9), ani jeden však neodpovídal vzoru $T_nA_nG_n$. U motýla byla telomerická sekvence $(TTAGG)_n$ nalezena oběma výše zmíněnými nástroji. Ve výstupu Tandem Repeats Finder byla na prvním místě u obou datasetů, ale lišila se abundancí, která byla v obohaceném datasetu vyšší (Tabulka 7). Podobný jev byl pozorován u výsledků BioSERF analýzy, kde byl telomerický motiv úspěšně odhalen v obohaceném datasetu, ale nikoliv ve standardním (Tabulka 8). Tato data ukazují, že *in silico* obohacení bylo úspěšné a u kontrolního druhu napomohlo k lepší identifikaci telomerické sekvence.

Fyzické mapování mikrosatelitních motivů identifikovaných u *P. phalangioides* bylo ověřeno na chromozomálních preparátech pomocí fluorescenční *in situ* hybridizace (FISH). Ačkoliv uvedená metoda fungovala a u kontrolního druhu *Y. evonymella* bylo možné telomery detekovat (Obrázek 4A-C), u *P. phalangioides* se telomerický patern kandidátních motivů nepotvrdil. Z důvodu limitovaného množství preparátů s pachytenními jádry a DAPI artefaktů (Obrázek 4N, Q) bylo možné až na motiv $(A)_n$ lokalizaci ověřit na mitotických chromozomech, sondy ale hybridizovaly po celé délce chromozomů a na koncích v telomerických oblastech nepřevažovaly (Obrázek 4C-R). Testované mikrosatelity se tedy pravděpodobně vyskytují napříč celým genomem. Ke stejným výsledkům nezávisle došli i Voráčková a Forman (osobní komunikace, nepublikovaná data), kteří na *P. phalangioides* hybridizovali stejné motivy $(TA)_n$, $(TG)_n$ a $(A)_n$ s totožným výsledkem.

I přesto, že telomerickou lokalizaci zkoumaných mikrosatelitů u třesavky *Pholcus phalangioides* nebylo možné potvrdit, se zdá že sekvence s nízkou komplexitou se na chromozomálních koncích u pavouků opravdu vyskytují, ať už v telomerických či subtelomerických oblastech. Dle výsledků z databáze telomerických sekvencí automaticky predikovaných z konců genomových „assembly“ na úrovní chromozomů (https://github.com/tolkit/a-telomeric-repeat-database), by se u pavouka *P. phalangioides* mohla vyskytovat telomerická sekvence s motivem $(A)_n$. Tato sekvence skutečně je i mezi motivy identifikovanými v rámci této práce. Predikované telomerické motivy u ostatních druhů pavouků uvedených v této databázi se však téměř zcela liší mezi jednotlivými druhy a zpravidla se jedná o velmi jednoduché mikrosatelity, neodpovídající typickým telomerickým motivům.

Ze získaných dat vyplývá, že telomerická sekvence u pavouků pravděpodobně není variací na již známé motivy a nenásleduje klasický vzorec $T_nA_nG_n$. K vyřešení této

problematiky bude třeba provést podrobnější analýzy repetitivních sekvencí na dalších druzích pavouků, a to zejména druhů s malými genomy a velkým počtem chromozomů, včetně přesného ověření lokalizace na chromozomálních koncích.

6. Závěr:

Hlavním cílem této bakalářské práce bylo objevit telomerickou sekvenci u pavouků. Ke způsobu identifikace telomerické sekvence bylo přistoupeno bioinformaticky, a to s použitím sekvenačních dat dostupných ve veřejných databázích. Kandidátní sekvence byly zkoumány nejprve komparativní analýzou repetic napříč zástupci infrařádu Araneomorphae, která však potenciální telomerické motivy neodhalila. Podrobná analýza byla tedy provedena na vybraném druhu *Pholcus phalangioides* tak, že byl dataset obsahující sekvenační čtení pavouka obohacen o telomerickou sekvenci pro snadnější odhalení analýzou repetitivních motivů. Paralelně se postupovalo stejným stylem na datasetu motýla *Yponomeuta evonymella*, jehož telomerická sekvence $(TTAGG)_n$ je dobře známá a byl tedy použit jako pozitivní kontrola. Z analýzy repetitivních sekvencí na koncích obohacených o telomerickou sekvenci bylo odhaleno 5 kandidátních motivů u pavouka *P. phalangioides*: $(TTTGTA)_n$, $(TATGTA)_n$, $(A)_n$, $(TA)_n$, a $(TG)_n$. Výsledky fyzické kolalizace metodou FISH kandidátní sekvence v telomerách pavouka *Pholcus phalangioides* nepotvrdily a sondy značily celé chromozomy, ačkoliv u kontrolního datasetu *Yponomeuta evonymella* tento způsob fungoval a telomerická sekvence $(TTAGG)_n$ byla úspěšně detekována. Získaná data ukazují, že u pavouků se pravděpodobně nevyskytuje telomerická sekvence následující klasický telomerický vzorec $T_nA_nG_n$, ale jedná se zatím neznámý motiv, který může být tvořený sekvencemi s nízkou komplexitou.

7. Bibliografie

Altschul, S., F., et al. (1990) Basic local alignment search tool. *Journal of molecular biology,* **215**: 403-410.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Arkhipova, I., R., and Yushenova, I., A. (2019) Giant transposons in eukaryotes: is bigger better?" *Genome biology and evolution,* **11**: 906-918.

Arnone, M., I., and Davidson, E., H. (1997): The hardwiring of development: organization and function of genomic regulatory systems. *Development*, **124**: 1851-1864.

Belicard, T., Jareosettasin, P., and Sarkies, P. (2018) The piRNA pathway responds to environmental signals to establish intergenerational adaptation to stress. *BMC biology,* **16**: 1-14.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research,* **27**: 573-580.

Biessmann, H., et al. (1998) DNA organization and length polymorphism at the 2L telomeric region of Anopheles gambiae. *Insect molecular biology,* **7**: 83-93.

Biscotti, M., A., Ettore O., and Heslop-Harrison, J., S. (2015) Repetitive DNA in eukaryotic genomes. *Chromosome Research,* **23**: 415-420.

Blackburn, E., H. (1991) Structure and function of telomeres. *Nature,* **350**: 569-573.

Bolger, A., M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* **30**: 2114-2120.

Bull, J., J. (1983) *Evolution of sex determining mechanisms*. USA, The Benjamin/Cummings Publishing Company, Inc.

Carthew, R., W., Agbu, P., and Giri, R. (2017) MicroRNA function in Drosophila melanogaster."*Seminars in cell & developmental biology*, **65**. Academic Press.

Chan, S., R., W., L, and Blackburn, E., H. (2004) Telomeres and telomerase. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences,* **359**: 109-122.

Coddington, J., A. (2005) Phylogeny and classification of spiders. *Spiders of North,* edited by Ubick, D., Cushing, P. E., and Paquin, P., *American: an identification manual*.

Cong, Y., Wright, W., E., and Shay, J., W. (2002) Human telomerase and its regulation. *Microbiology and molecular biology reviews,* **66**: 407-425.

Cooke, H., J., and Smith, B., A. (1986) Variability at the telomeres of the human X/Y pseudoautosomal region. *Cold Spring Harbor symposia on quantitative biology*, **51**. Cold Spring Harbor Laboratory Press.

Dimitri, C., Pickett, H., A., and Reddel, R., R. (2013) Alternative lengthening of telomeres: remodeling the telomere architecture. *Frontiers in oncology,* **3**: 27.

Doksani, Y. (2019) The response to DNA damage at telomeric repeats and its consequences for telomere function. *Genes,* **10**: 318.

Fajkus, P., et al. (2023) Telomerase RNA in Hymenoptera (Insecta) switched to plant/ciliate-like biogenesis. *Nucleic Acids Research,* **51**: 420-433.

Feschotte, C., and Pritham, E., J. (2007) DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics,* **41**: 331-368.

Fujiwara, H., et al. (2005) Telomere-specific non-LTR retrotransposons and telomere maintenance in the silkworm, Bombyx mori. *Chromosome Research,* **13**: 455-467.

Fulnečková, J., Ševčíková, T., Fajkus, J., et al. (2013) A Broad Phylogenetic Survey Unveils the Diversity and Evolution of Telomeres in Eukaryotes, *Genome Biology and Evolution*, **5**: 468–483.

Girard, M., B., Kasumovic, M., M., and Elias, D., O. (2011) Multi-modal courtship in the peacock spider, Maratus volans (OP-Cambridge, 1874). *PLoS One,* **6**: e25390.

Gomes, N., M., V., Shay, J., V., W. and Wright, W., E. (2010), Telomere biology in Metazoa, *FEBS Letters*, **584**.

Gregory, T. R., and Shorthouse, D., P. (2003) Genome sizes of spiders. *Journal of Heredity,* **94**: 285-290.

Greider, C., W. (1999) Telomeres do D-loop–T-loop. *Cell,* **97**: 419-422.

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies, *Bioinformatics*, **29**: 1072–1075.

Harley, C., B. (1991) Telomere loss: mitotic clock or genetic time bomb? *Mutation Research/DNAging,* **256**: 271-282.

Herrera, I., M., Á., Král, J., Pastuchová, M. et al. (2021). Evolutionary pattern of karyotypes and meiosis in pholcid spiders (Araneae: Pholcidae): implications for reconstructing chromosome evolution of araneomorph spiders. *BMC Ecology and Evolution*, **21**, 75.

Hebert, P., D., N., Cywinska A., Ball, S., L., deWaard, JR. (2003) Biological identifications through DNA barcodes. *Proceedings: Biological Sciences*, **270**: 313–321.

Lockwood, A., P., M. (1961) "Ringer", solutions and some notes on the physiological basis of their ionic composition, *Comparative Biochemistry and Physiology*, **2**: 241-289.

Jenner, L., P., et al. (2022) Telomeres and Their Neighbors. *Genes,* **13**: 1663.

Jones, B., C., et al. (2016) A somatic piRNA pathway in the Drosophila fat body ensures metabolic homeostasis and normal lifespan. *Nature communications,* **7**: 13856.

Kajino, T., et al. (2019) Divergent lnc RNA MYMLR regulates MYC by eliciting DNA looping and promoter-enhancer interaction. *The EMBO Journal,* **38**: e98441.

Král, J., et al. (2006) Evolution of the karyotype and sex chromosome systems in basal clades of araneomorph spiders (Araneae: Araneomorphae). *Chromosome Research,* **14**: 859-880.

Král, J., Kořínková, T., Krkavcová, L., Musilová, J., et al. (2013) Evolution of karyotype, sex chromosomes, and meiosis in mygalomorph spiders (Araneae:Mygalomorphae). *The Biological Journal of the Linnean Society*, **109**: 377- 408.

Kuntner, M. (2022) The seven grand challenges in arachnid science. *Frontiers in Arachnid Science*, **2**.

Langmead, B., and Salzberg, S., L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods,* **9**: 357-359.

Levsky, J., M., and Singer, R., H. (2003) Fluorescence *in situ* hybridization: past, present and future. *Journal of cell science,* **116**: 2833-2838.

Li, H., et al. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics,* **25**: 2078-2079.

Logsdon, G.A., Eichler, E., E. (2023) The Dynamic Structure and Rapid Evolution of Human Centromeric Satellite DNA. *Genes*, *14*, 92.

Lucchetta, E., M., et al. (2009) The endo-siRNA pathway is essential for robust development of the Drosophila embryo. *PloS one* **4**: e7576.

Ludwig, M., Z. (2002) Functional evolution of noncoding DNA. *Current opinion in genetics & development,* **12**: 634-639.

Manni, M., et al. (2021) BUSCO: assessing genomic data quality and beyond. *Current Protocols,* **1**: e323.

Marçais, G., Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers, *Bioinformatics*, **27**: 764–770.

Marks, P., et al. (2019) Resolving the full spectrum of human genome variation using Linked-Reads. *Genome research,* **29**: 635-645.

Mason, J., M., and Biessmann, H. (1995) The unusual telomeres of Drosophila." *Trends in Genetics,* **11**: 58-62.

McClintock, B. (1941) The stability of broken ends of chromosomes in Zea mays. *Genetics,* **26**: 234.

McHugh, C., A., et al. (2015) The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature,* **521**: 232-236.

McEachern, M., J., et al. (2000) Telomere fusions caused by mutating the terminal region of telomeric DNA. *Proceedings of the National Academy of Sciences,* **97**: 11409-11414.

Mora, P., et al. (2023) Making the Genome Huge: The Case of Triatoma delpontei, a Triatominae Species with More than 50% of Its Genome Full of Satellite DNA. *Genes,* **14**: 371.

Novák, P., et al. (2020) Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nature Protocols*, **15**: 3745-3776.

Ohno, S. (1972) So much "junk" DNA in our genome. In „Evolution of Genetic Systems". *Brookhaven Symposium in Biology*, **23**.

Olovnikov, A., M. (1973) A theory of marginotomy: the incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon. *Journal of theoretical biology,* **41**: 181-190.

Pappalardo, X., G., Barra, V. (2021) Losing DNA methylation at repetitive elements and breaking bad. *Epigenetics & Chromatin,* **14**, 25.

Pardue, M. L., and DeBaryshe, P., G. (2008) Drosophila telomeres: A variation on the telomerase theme. *Fly,* **2**: 101-110.

Pennacchio, L., A., and Rubin, E., M. (2001) Genomic strategies to identify mammalian regulatory sequences. *Nature reviews genetics,* **2**: 100-109.

Peška, V., et al. (2017) BAL31-NGS approach for identification of telomeres de novo in large genomes. *Methods,* **114**: 16-27

Peška, V., et al. (2015) Characterisation of an unusual telomere motif (TTTTTTAGGG)n in the plant Cestrum elegans (Solanaceae), a species with a large genome. *The Plant Journal*, **82**: 644-654.

Provazníková, I., et al. (2021) Large-scale comparative analysis of cytogenetic markers across Lepidoptera. *Scientific Reports,* **11**, 12214.

Provazníková, I. (2023) Drivers of karyotype evolution in Lepidoptera. Disertační práce. Jihočeská univerzita v Českých Budějovicích, Přírodovědecká fakulta, České Budějovice, Česká republika, 131 stran

Prušáková, D., et al. (2021) Telomeric DNA sequences in beetle taxa vary with species richness. *Scientific Reports,* **11**, 13319.

Ranallo-Benavidez, T., R., et al. (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature communications,* **11**: 1432

Reyes Lerma, A., C., et al. (2021) Insights into the Karyotype Evolution of Charinidae, the Early-Diverging Clade of Whip Spiders (Arachnida: Amblypygi). *Animalsm,* **11**: 3233.

Riethman, H., et al. (2005) Human subtelomere structure and variation. *Chromosome Research,* **13**: 505-515.

Roa, F., Guerra, M. (2012) Distribution of 45S rDNA sites in chromosomes of plants: Structural and evolutionary implications. *BMC Evolutionary Biology,* **12**, 225.

Rowell, D., M. (1991): Chromosomal fusion and meiotic behaviour in Delena cancerides (Araneae:Sparassidae). I. Chromosome pairing and X-chromosome segregation. *Genome*, **34**: 561-566

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Saez, N., J., and Herzig, V. (2019) Versatile spider venom peptides and their medical and agricultural applications. *Toxicon* **158**: 109-126.

Salehi, S., et al. (2020) Spider silk for tissue engineering applications. *Molecules,* **25**: 737.

Schindelin, J., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods*, **9**: 676–682.

Schroeder-Reiter, E., and Wanner, G. (2009) Chromosome centromeres: structural and analytical investigations with high resolution scanning electron microscopy in combination with focused ion beam milling. *Cytogenetic and Genome Research*, **124**: 239-250.

Sember, A., et al. (2020) Patterns of sex chromosome differentiation in spiders: Insights from comparative genomic hybridisation. *Genes,* **11**: 849.

Servant, G., and Deininger, P., L. (2016) Insertion of retrotransposons at chromosome ends: adaptive response to chromosome maintenance. *Frontiers in Genetics,* **6**: 358.

Shapiro, J., A. (2022) Repetitive DNA is Functional and Encodes Parts of the Non-Coding RNA Repertoire. *Advanced Genetics,* **3**: 2200026.

Shen, W., et al. (2016) SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS one,* **11**: e0163962.

Solodskikh, S., A., et al. (2019) Postgenomic technologies for genomic and proteomic analysis in biological and medical research, *Ukrainian Journal of Ecology*, **9**: 765-776.

Soibam, B., and Zhamangaraeva, A. (2021) LncRNA: DNA triplex-forming sites are positioned at specific areas of genome organization and are predictors for Topologically Associated Domains. *BMC genomics,* **22**: 1-10.

Somanathan, I., and Baysdorfer, Ch. (2018) A bioinformatics approach to identify telomere sequences. *Biotechniques,* **65**: 20-25.

Sobreira, T., J., P., et al. (2006) TRAP: automated classification, quantification and annotation of tandemly repeated sequences. *Bioinformatics*, **22**: 361–362.

Southern, E. (2006) Southern blotting. *Nature protocols,* **1**: 518-525.

Suzuki, S. (1954): Cytological studies in spiders. III. Studies on the chromosomes of fiftysevenspecies of spiders belonging to seventeen families, with general considerations on chromosomalevolution. J. Sci. Hiroshima Univ. B, **15**: 23-136.

The Galaxy Community. (2022) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, **50**: W345–W351.

Van Rossum, G., and Drake, F., L. (2009) *Introduction to python 3: python documentation manual part 1*. CreateSpace.

Van Straalen, N., I., and Roelofs, D. (2006) *Introduction to Ecological Genetics* (New York, Oxford University Press)

Venter, J., C., et al. (2001) The sequence of the human genome. *science* **291**: 1304-1351.

Vítková, M., et al. (2005) The evolutionary origin of insect telomeric repeats, (TTAGG) n. *Chromosome research,* **13**: 145-156.

Watson, J., M., et al. (2021) A hypomorphic allele of telomerase uncovers the minimal functional length of telomeres in Arabidopsis. *Genetics,* **219**: iyab126.

Weisenfeld, N., L., et al. (2017). Direct determination of diploid genome sequences. *Genome Research*, **27**: 757-767.

Wells, J., N., and Feschotte, C. (2020) A field guide to eukaryotic transposable elements. *Annual review of genetics,* **54**: 539-561.

Wheeler, W., C., et al. (2017) The spider tree of life: phylogeny of Araneae based on target-gene analyses from an extensive taxon sampling. *Cladistics,* **33**: 574-616.

Willard, H., F. (1990) Centromeres of mammalian chromosomes. *Trends in Genetics,* **6**: 410-416.

World Spider Catalog (2023). World Spider Catalog. Version 24. Natural History Museum Bern, online at http://wsc.nmbe.ch, accessed on 21.2.

Yang, F., et al. (2015) The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome biology,* **16**: 1-17.

8. Přílohy

Příloha 1: Script pro přípravu obohaceného datasetu

```
# Script for telomeric enrichment by Marco Badici
# Prepare environment
import os
from Bio import SeqIO
import numpy as np

os.chdir('C:\Users\lukas\Desktop\Data\')

# Prepare input and output files
input_file = 'assembly.fasta'

fasta_sequences = SeqIO.parse(open(input_file),'fasta')

#Divide sequences based on length
with open('under_5000.fasta', 'w') as under:
    with open('over_5000.fasta', 'w') as over:
        for fasta in fasta_sequences:
            if len(fasta.seq) <= 5000:
                SeqIO.write(fasta, under, 'fasta')
            else:
                SeqIO.write(fasta, over, 'fasta')


list_ids = []
list_descriptions = []
list_sequences = []

over = SeqIO.parse(open('over_5000.fasta'),'fasta')
for entries in over:
    list_ids.append(entries.id)
    list_descriptions.append(entries.description)
    list_sequences.append(str(entries.seq))

list_ids = np.repeat(list_ids, 2)

redo_list = []

for i in list_ids:
    for j in range(1,3):
        redo_list.append(i+'_'+str(j))

redo_list = list(dict.fromkeys(redo_list))

list_descriptions = np.repeat(list_descriptions, 2)

```

```
#Extract 5 000 bp from ends of long contigs
first_5000 = []
last_5000 = []
for i in list_sequences:
    first_5000.append(i[:5000])
    last_5000.append(i[-5000:])

list_sequences = []
while True:
    try:
        list_sequences.append(first_5000.pop(0))
        list_sequences.append(last_5000.pop(0))
    except IndexError:
        break

descriptions = []

for i in list_descriptions:
    element = str(i).split(' ', 1)[-1]
    descriptions.append(str(element))

print(len(redo_list))
print(len(descriptions))
print(len(list_sequences))

print(descriptions)


ofile = open("over_5000_modif.fasta", "w")

for i in range(len(list_ids)):
    ofile.write(">" + redo_list[i] + ' ' + descriptions[i] + "\n" +
list_sequences[i] + "\n")

ofile.close()

# Merge all files and create enriched dataset
with open('under_5000.fasta') as fp:
    data = fp.read()

with open('over_5000_modif.fasta') as fp:
    data2 = fp.read()

data += "\n"
data += data2

with open('combined.fasta', 'w') as fp:
    fp.write(data)

```