MASARYKOVA UNIVERZITA
Přírodovědecká fakulta
Centrum pro výzkum toxických látek v prostředí

Bakalářská práce

Brno 2020 **Michal Bubeník**

MASARYKOVA
UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
CENTRUM PRO VÝZKUM TOXICKÝCH LÁTEK V PROSTŘEDÍ

Vysokokapacitní screening telomerových sekvencí ve veřejně dostupných databázích sekvenčních dat nové generace

Bakalářská práce

Michal Bubeník

Vedoucí práce: Mgr. Vratislav Peška, Ph.D. **Brno 2020**

Bibliografický záznam

Autor:	Michal Bubeník Přírodovědecká fakulta, Masarykova univerzita Centrum pro výzkum toxických látek v prostředí

Název práce:	Vysokokapacitní screening telomerových sekvencí ve veřejně dostupných databázích sekvenčních dat nové generace
Studijní program:	Experimentální biologie
Studijní obor:	Matematická biologie
Vedoucí práce:	Mgr. Vratislav Peška, Ph.D.
Akademický rok:	2019/2020
Počet stran:	viii + 81
Klíčová slova:	automatizace; databáze; fylogeneze; high-throughput screening; NGS; SRA; telomery

Bibliographic Entry

Author:	Michal Bubeník Faculty of Science, Masaryk University Research Centre for Toxic Compounds in the Environment
Title of Thesis:	High-throughput screening of telomeric sequences in publicly available next generation sequence data databases
Degree Programme:	Experimental Biology
Field of Study:	Mathematical Biology
Supervisor:	Mgr. Vratislav Peška, Ph.D.
Academic Year:	2019/2020
Number of Pages:	viii + 81
Keywords:	automation; databases; phylogeny; high-throughput screening; NGS; SRA; telomeres

Abstrakt

Hlavním cílem této bakalářské práce je vyvinout postup automatického zpracování veřejně dostupných sekvenčních dat nové generace, jež jsou uložena v databázi SRA, pomocí nástrojů pro nalezení tandemových repetic. Screening zahrnuje alespoň jeden datový soubor pro každý (v té době) dostupný sekvenovaný eukaryontní druh. Jeho hlavním účelem je identifikace kandidátů na neobvyklé (změněné nebo nové) telomerové sekvence. Celý screening je uzpůsoben k provedení v Národní gridové infrastruktuře. Druhým účelem práce je využít kvantitativní informace o zastoupení tandemových repetic v genomech a provést nad vybranou částí dat fylogenetickou analýzu.

Abstract

The main aim of this bachelor thesis is to develop a procedure of automatic processing of publicly available next-generation sequencing data stored in the SRA database by tools for finding tandem repeats. The screening covers at least one data set for each (at that time) available sequenced eukaryotic species. Its main objective is to identify candidates for unusual (modified or new) telomeric sequences. The whole screening is designed to be performed in the National Grid Infrastructure. The second goal of this thesis is to utilize the quantitative information about the tandem repeats' abundance in the genomes and to carry out a phylonenetic analysis on a selected part of the data.

MUNI
SCI

MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
KOTLÁŘSKÁ 2, 611 37 BRNO
IČ: 00216224
DIČ: CZ00216224

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Akademický rok: 2019/2020

Ústav:	RECETOX
Student:	Michal Bubeník
Program:	Experimentální biologie
Obor:	Matematická biologie

Ředitel *RECETOX* PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje bakalářskou práci s názvem:

Název práce:	Vysokokapacitní screening telomerových sekvencí ve veřejně dostupných databázích sekvenčních dat nové generace
Název práce anglicky:	High-throughput screening of telomeric sequences in publicly available next generation sequence data databases
Jazyk závěrečné práce:	čeština

Oficiální zadání:
Telomery jsou nukleoproteinové struktury na koncích eukaryotických chromozomů. Jejich cílem je vytvářet ochranný komplex telomerové DNA a proteinů. Cílem bakalářské práce je vyvinout automatický postup pro HTS (high-throughput screening) ve veřejně dostupných NGS (next generation sequencing) datových souborech. Výstupy plánovaného HTS poskytnou cenné informace o celé řadě neznámých telomerových sekvencí, například u prvokových parazitů, houbových mikroorganismů a škůdců ze skupiny bezobratlých živočichů. Student vypracuje bakalářskou práci ve spolupráci s oddělením Radiobiologie a buněčné biologie Biofyzikálního ústavu AVČR, v.v.i. Úkolem studenta bude: 1. Navrhnout automatizovaný postup high-throughput screeningu telomerových sekvencí v databázi Sequence Read Archive (NCBI), 2. Vyhledat změněné a nové telomerové sekvence napříč eukaryotickou evolucí, 3. Analyzovat vlastní sekvenační data z vybrané čeledi rostlin, 4. Sestrojit fylogenetický strom z NGS dat.

Literatura:
MOUNT, David W. *Bioinformatics : sequence and genome analysis*. 2nd ed. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press, 2004. xii, 692. ISBN 0879697121.

Bioinformatics for high throughput sequencing. Edited by Naiara Rodríguez-Ezpeleta - Michael Hackenberg - Ana M. Aransay. New York, NY: Springer, 2012. xi, 255. ISBN 9781461407812.

Vedoucí práce:	Mgr. Vratislav Peška, Ph.D.
Datum zadání práce:	12. 8. 2019
V Brně dne:	23. 1. 2020

Souhlasím se zadáním (podpis, datum):

.....	.....	.....
Michal Bubeník	Mgr. Vratislav Peška, Ph.D.	prof. RNDr. Ladislav Dušek, Ph.D.
student	**vedoucí práce**	**zástupce ředitele RECETOX pro pedagogické záležitosti**

Poděkování

Tímto děkuji vedoucímu své bakalářské práce, Mgr. Vratislavu Peškovi, Ph.D., za to, že měl na mě vždycky čas, že na něm bylo znát, jak ho telomery opravdu baví, a vůbec za ochotné, přívětivé a inspirativní vedení.

Tato práce vznikla za podpory projektu SYMBIT reg. č.: CZ.02.1.01/0.0/0.0/15_003/0000477 financovaného z EFRR

Výpočetní zdroje byly poskytnuty projektem „e-Infrastruktura CZ“ (e-INFRA LM2018140).

Prohlášení

Prohlašuji, že jsem svoji bakalářskou práci vypracoval samostatně pod vedením vedoucího práce s využitím informačních zdrojů, které jsou v práci citovány.

Brno 14. června 2020 .....

Michal Bubeník

Obsah

Úvod

Telomery jsou nukleoproteinové komplexy na koncích lineárních chromozomů eukaryontních organismů. Podílejí se na udržování stability genomu (brání nežádoucím chromozomovým přestavbám) a současně umožňují zdárný průběh replikace deoxyribonukleové kyseliny (DNA) při buněčném dělení.

Ze sekvenčního hlediska se většinou jedná o tandemově repetitivní sekvence (jeden relativně krátký sekvenční motiv se opakuje mnohokrát za sebou „hlava k patě“). Telomerové sekvence bývají značně evolučně konzervované, přesto existují různé varianty konkrétní podoby dané sekvence. Smyslem této práce je vyvinout a aplikovat automatický bioinformatický postup hledání takových výjimek.

Zkoumání (neobvyklých) telomer je relevantní jednak z hlediska fylogenetického, protože změny v telomerových sekvencích (a tandemových repeticích vůbec) souvisejí s evolucí druhů, a jednak potenciálně i z hlediska medicínského. Telomery se totiž významně podílejí na procesu stárnutí buněk, tedy i celého organismu, a s telomerovými abnormalitami jsou spjaty i vznik a šíření nádorů. Tato práce je však součástí základního výzkumu a necílí na přímé medicínské aplikace získaných poznatků.

V práci se při hledání neobvyklých telomerových sekvencí *in silico* vychází z několika předpokladů:

1. telomerová sekvence má podobu tandemové repetice,
2. je v genomu hojně zastoupena (vysoká relativní abundance vůči jiným tandemovým repeticím),
3. kanonická telomerová sekvence sekvence není přítomna,
4. neobvyklá telomerová sekvence je „podobná“ té kanonické.

V práci byly použity již existující nástroje pro identifikaci tandemových repetic v sekvenčních datech. Screeningem prošel alespoň jeden datový soubor od každého dostupného[1] eukaryontního druhu. V získaném seznamu přítomných repetic byly při nepřítomnosti kanonické telomerové sekvence hledány kandidátní sekvence na neobvyklé telomery.

Bakalářská práce je členěna na čtyři kapitoly. V první, teoretické kapitole jsou obsaženy základy telomerové biologie – stručná historie základních objevů, struktura telomer, stavba a princip fungování enzymu telomerázy, která telomery syntetizuje, i alternativní způsoby prodlužován telomer. Dále jsou v ní představeny principy sekvenování nové generace (NGS).

[1] Myšleno v době běhu screeningu, sekvenčních dat velmi rychle přibývá.

Ve druhé kapitole je popsána použitá metodika, zvláště postup samotného screeningu, použité programy a skripty, získávání datových souborů z databází, práce se soubory, ukládání velkých objemů dat (řádově jednotky TB), způsob identifikace telomerových kandidátů, uspořádání a řazení vygenerovaných dat, resp. různé související problémy. Popsán je také postup odesílání dávkových úloh do plánovacího systému Národní gridové infrastruktury (NGI).

Třetí kapitola obsahuje výsledky automatického screeningu, jsou v ní uvedeny nejzajímavější druhy bez obvyklých kanonických sekvencí a nalezené tandemové repetice, které představují nejslibnější kandidáty na sekvenčně neobvyklé telomery daných druhů (pokud takové sekvence byly v datech nalezeny).

Čtvrtá kapitola je určitým tematickým doplňkem vycházejícím ze snahy využít i zbylé informace o zastoupení tandemových repetic v genomech, a to pro účely orientační fylogenetické analýzy. Je v ní diskutován problém předběžné úpravy a kontroly vstupních dat, jejich další potřebné předzpracování a nakonec jsou popsány výsledky analýzy na konkrétní skupině sekvenovaných genomů.

Použité zkratky

Zkratky jsou uvedeny v abecedním pořádku.

A	adenin (purinová báze v DNA nebo RNA)
$\mathtt{A}^*$, $\mathtt{C}^*$, $\mathtt{G}^*$, $\mathtt{T}^*$	daná báze se v sekvenci na daném místě opakuje v počtu 0 či více
$\mathtt{A}^+$, $\mathtt{C}^+$, $\mathtt{G}^+$, $\mathtt{T}^+$	daná báze se v sekvenci na daném místě opakuje v počtu 1 či více
ALT	„alternative lengthening of telomeres“ – alternativní prodlužování telomer
C	cytosin (pyrimidinová báze v DNA nebo RNA)
ddNTP	dideoxyribonukleotid
DNA	deoxyribonukleová kyselina
dNTP	deoxyribonukleotid
emPCR	„emulsion polymerase chain reaction“ – emulzní polymerázová řetězové reakce
G	guanin (purinová báze v DNA nebo RNA)
HTS	„high-throughtput screening“ – vysokokapacitní screening
ITS	„intersticial telomeric sequences“ – vmezeřené telomerové sekvence
NCBI	National Center for Biotechnology Information
NGI	Národní gridová infrastruktura
NGS	„next-generation sequencing“ – sekvenování nové generace
NJ	„neighbour joining“ – klastrovací metoda
PCR	„polymerase chain reaction“ – polymerázová řetězové reakce
RNA	ribonukleová kyselina
T	thymin (pyrimidinová báze v DNA)
TERC	„telomerase RNA component“ – telomerová RNA-podjednotka
TERRA	„telomeric repeat-containing RNA“ – RNA obsahující telomerové repetice
TERT	„telomerase reverse transcriptase“ – telomerázová reverzní transkriptasa (katalytická podjednotka enzymu telomerázy)
TGS	„third-generation sequencing“ – sekvenování třetí generace
TRFi	Tandem Repeats Finder – název softwaru
TRMe	Tandem Repeats Merger – název softwaru
SRA	„Sequence Read Archive“ – Archiv sekvenčních čtení
U	uracil (pyrimidinová báze v RNA)
WGS	„whole genome sequencing“ nebo „whole genome shotgun“ – celogenomové sekvenování

Kapitola 1

Biologické základy tématu

V této kapitole bude podán výklad biologického a bioinformatického pozadí práce. Nejprve bude objasněna funkce a struktura telomer, včetně příslušných evolučních souvislostí, potom bude obecně popsán proces vzniku dat, s nimiž se bude zacházet v praktické části.[2]

1.1 Telomerová biologie

Jaderný genom eukaryontní buňky je tvořen lineárními molekulami DNA zvanými chromozomy. Na obou koncích chromozomu se nachází komplex DNA a proteinů označovaný pojmem telomera[3].

1.1.1 Počátky výzkumu telomer

Konce mitotického chromozomu[4] nevykazují při mikroskopickém pozorování žádné zjevné morfologické odlišnosti od zbytků chromozomových ramen. Na zvláštní funkci konců chromozomů usoudil teprve ve 30. letech 20. st., několik desetiletí po objevu chromozomu, americký genetik Hermann Muller. Prostřednictvím ionizujícího záření dosahoval cílené mutageneze u chromozomů octomilek (*Drosophila melanogaster*) a pozoroval, že rozsáhlé chromozomové aberace[5] nenastávají u těch chromozomů, jejichž konce zůstaly nepoškozeny [63]. Vyvodil závěr, že *„terminální gen musí mít speciální funkci zapečetit konce chromozomů a chromozom z nějakého důvodu nemůže přetrvávat bez takto uzavřených konců“* ([63], citováno v [60], volně přeloženo). Právě on také přišel s pojmem „telomera“.

Rovněž Barbara McClintock koncem 30. let pozorovala u kukuřice (*Zea mays*) fúze chromozomů s poškozenými konci a dospěla tak k závěru o existenci ochranné struktury na jejich koncích [58], [59]. Budiž podotknuto, že octomilka a kukuřice jsou

[2] Konkrétní popis postupu získání dat a také problematika NGI budou popsány až v kapitole 2.

[3] Z řeckých slov $\tau\varepsilon\lambda\text{o}\varsigma$ = konec a $\mu\varepsilon\rho\text{o}\varsigma$ = část, doslova tedy „koncová část (chromozomu)“.

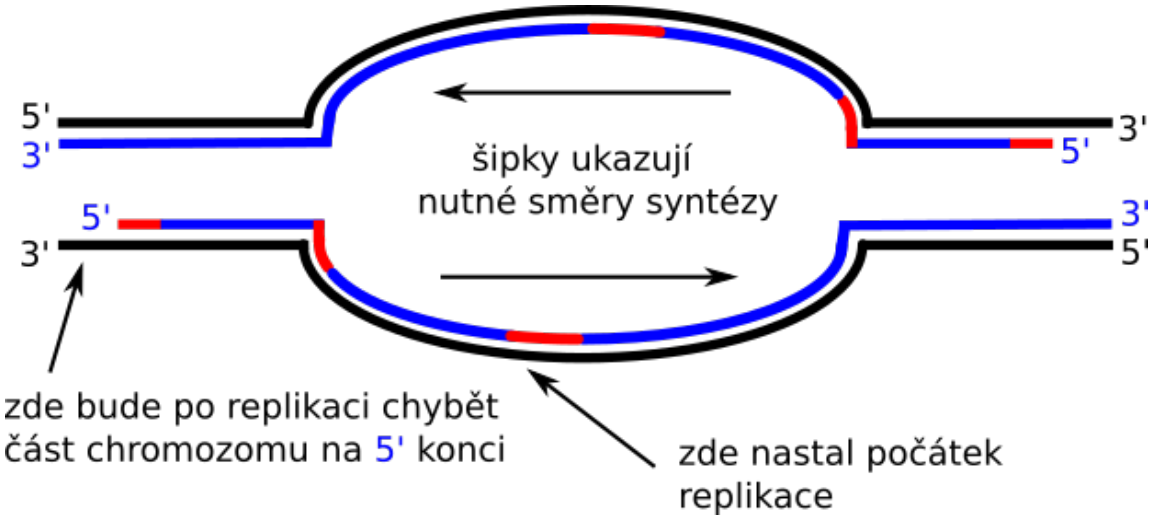
[4] Během buněčného dělení (mitózy, meiózy) chromozomy nabývají charakteristického tvaru „X“ a jsou snáze pozorovatelné i proto, že jaderná membrána se (u většiny eukaryontů) během dělení rozpadá. Během interfáze buněčného cyklu je chromatin organizován jinak, blíže např. v [6].

[5] Chromozomové přestavby – zejm. delece (ztráty), duplikace (zdvojení), translokace (přesuny), inverze (obrácení směru) – vše jako důsledky zlomů a následných fúzí chromozomů.

organizmy fylogeneticky poměrně vzdálené – to naznačuje, že rozlišování telomer od dvouřetězcových zlomů je obecnou vlastností lineárních chromozomů [77].

V době těchto objevů však ještě ani nebylo prokázáno, že DNA je skutečně nositelkou dědičnosti, ani nebyla známa její molekulární struktura. Další objevy v oblasti telomerové biologie proto přišly až s odstupem tří dekád, v 70. letech.

V uvedené době poukázali nezávisle na sobě Alexej Olovnikov a Jim Watson na tzv. „problém konce replikace“ (the end-replication problem) [71], [107][6]. První funkce telomer, udržení stability genomu, již vyplynula z dřívějších objevů. Problém konce replikace souvisí s jejich druhou funkcí jakéhosi „replikačního pufru“. Daný problém je ilustrován obrázkem 1.1.



Obrázek 1.1: Zjednodušené schéma replikace lineárního chromozomu. Obrázek ilustruje zkracování chromozomu na konci 5’ a vznik přesahu na konci 3’[7].

Enzym DNA-polymeráza zajišťující připojování nukleotidů[8] syntetizovaného řetězce DNA totiž vykazuje aktivitu pouze ve směru od konce 5’ ke konci 3’, a ne naopak. Syntéza tedy nemůže probíhat oběma směry. Od místa počátku replikace proto syntéza nového vlákna probíhá v jednom směru plynule (tzv. vedoucí řetězec), ale ve druhém směru po kratších částech zvaných Okazakiho fragmenty (tzv. váznoucí řetězec). Každý z Okazakiho fragmentů začíná krátkým primerem, sekvencí RNA syntetizovanou enzymem DNA-primázou (tato sekvence je později „vystřižena“ a nahrazena DNA). Primery jsou potřebné proto, že DNA-polymeráza není schopna syntetizovat řetězec DNA *de novo*, vždy musí prodlužovat řetězec již existující, zatímco DNA-primáza[9] syntézu řetězce *de novo* zvládá. Úplný konec chromozomu pak

[6] Olovnikov ve skutečnosti publikoval své výsledky před Watsonem, již v roce 1971, ale v ruštině. Trvalo tedy, než proniky na Západ.

[7] Označení konců molekuly DNA 5’ a 3’ určuje pozici volné hydroxylové funkce na cukru v cukr-fosfátové páteři („uprostřed“ molekuly DNA totiž na cukru deoxyribose žádné volné skupiny OH nejsou), 3’ a 5’ místo 3 a 5 se píše proto, že čísla bez čárek jsou rezervována pro označení atomů dusíkatých bází (A, G, C, T v DNA, resp. U místo T v RNA).

[8] Nukleotid je základní stavební jednotkou DNA, každý nukleotid se skládá z deoxyribosy, fosfátu a jedné z heterocyklických dusíkatých bází, jejichž pořadí je označováno jako sekvence DNA.

[9] Technicky DNA-dependentní RNA-polymeráza.

nemá dostupný primer pro vlastní replikaci, a tak několik desítek nukleotidů není replikováno, tím se chromozom zkracuje.

Už v 60. letech Leonard Hayflick publikoval práce o omezeném čase života lidských buněčných linií – ty po určitém poču generací (tzv. Hayflickův limit) stárnou a odumírají, přestože jsou kultivovány v optimálních podmínkách [35], [36]. Právě Olovnikov v roce 1971 (resp. Watson v roce 1972) jako první vyslovil hypotézu, že Hayflickův limit, tj. buněčné stárnutí, je důsledkem zkracování chromozomů při procesu replikace DNA. Vyvodil také závěr, že musí existovat kompenzační mechanismus tohoto zkracování[10]. Tímto mechanismem je právě množství repetitivní nekódující DNA, ta tvoří jakýsi „nárazník“ a jejím odbouráváním nejsou narušeny ty části chromozomu, které nesou genovou informaci.

1.1.2 Známé telomerové sekvece

V roce 1978 byla objevena první telomerová sekvence – hexanukleotid `5'-TTGGGG-3'`, a to u prvoka rodu *Tetrahymena*[11] [15]. V roce 1980 byla objevena telomerová sekvence jiného nálevníka *Stylonychia pustulata* – `5'-TTTTGGGG-3'` [69] a bylo zjištěno, že i jiní nálevníci mají vesměs podobné telomerové sekvence [47]. Tyto sekvence tvaru `5'-T`$^{+}$`G`$^{+}$`-3'` však nejsou mezi eukaryonty nejběžnější, „kanonická“ telomerová sekvence má spíše tvar `5'-T`$^{+}$`A`*`G`$^{+}$`-3'`. Asi nejběžnější je sekvence `5'-TTAGGG-3'`, ta byla poprvé izolována z lidských buněk v roce 1988, ale je považována za společnou všem obratlovcům (*Vertebrata*) [61], sdílejí ji i některé jiné taxony, a to jak z řad strunatců (*Chordata*) [50] a jim příbuzných ostnokožců (*Echinodermata*) [52], tak i např. některých hub *Fungi* [95] nebo některých rostlin řádu chřestotvarých (*Asparagales*) [110]. Většina rostlin obsahuje telomerovou sekvenci `5'-TTTAGGG-3'` [60], poprvé byla izolována z *Arabidopsis thaliana* v roce 1988 [84]. Pro hmyz je typická sekvence `5'-TTAGG-3'` [70], [88], [30], resp. alternativní sekvence `5'-TCAGG-3'` u některých brouků (*Coleoptera*[12]) [62]. Poznání různosti telomerových sekvencí napříč eukaryontní fylogenezí se i nadále rozšiřuje a stále jsou popisovány nové telomerové sekvenční motivy, část těchto objevů pochází i z PřF MU a přidružených pracovišť [80], [29].[13]

U *Drosophila melanogaster*, jejímž zkoumáním se počíná celá telomerová biologie, paradoxně „standardní“ telomerové repetice nemá. Její telomery se skládají z více typů retrotranspozonů[14] [60], [14], [73].

[10] Je třeba poznamenat, že skutečná rychlost zkracování telomer je vyšší, než jaká by odpovídala pouhým ztrátám při replikaci. Roli totiž hrají i kyslíkové radikály, resp. exonukleázy, jimiž jsou telomery poškozovány, resp. zpracovávány, a tím je zkracování urychlováno, jak píší von Zglinicki et al. v [115].

[11] Tento prvok byl vybrán jako modelový organizmus, protože jeho jádro obsahuje tisíce jednoduchých „minichromozomů“, tedy nadbytek telomerového materiálu [21].

[12] Jde zvláště o brouky nadčeledi *Tenebrionoidea*.

[13] Není cílem práce podat úplný přehled všech známých telomerových sekvencí. Poměrně obsáhlý přehled, byť dnes již neaktuální a neúplný, lze nalézt na `http://telomeráze.asu.edu/sequences_telomere.html` [81].

[14] Transpozon je mobilní element DNA – má schopnost být enzymaticky vyčleněn z chromozomu a začleněn ne jiné místo chromozomu. Retrotranspozon je běžný typ transpozonu podléhající tran-

Pro úplnost je třeba zmínit, že i mezi prokaryonty lze nalézt struktury podobné telomerám. Cirkulární chromozom, díky kterému se prokaryonti bez telomer obvykle obejdou, je totiž pro ně znak obvyklý, ale nikoli nutný. Tato práce se se však týká telomer eukaryontních, proto tato problematika nebude dále rozváděna, přehled viz např. v [19].

1.1.3 Evoluce telomer

Lze se ptát, jaké jsou fylogenetické vztahy mezi jednotlivými telomerovými sekvencemi – které jsou evolučně původní a které odvozené.[15] Za zcela původní telomerovou sekvenci je nejčastěji pokládána „obratlovčí“ sekvence `5'-TTAGGG-3'`, a to na základě jejího bohatého zastoupení napříč eukaryonty a také díky vztahu k předpokládaným kořenům eukaryontního evolučního stromu [26], [31]. Předpokládá se, že „hmyzí“ sekvence `5'-TTAGG-3'` se vyvinula z „obratlovčí“ sekvence `5'-TTAGGG-3'` [106]. Je velmi pravděpodobné, že modifikovaná sekvence `5'-TCAGG-3'` se vyvinula tranzicí[16] ze sekvence `5'-TTAGG-3'`.

Podobnou optikou je třeba nahlížet i na „obratlovčí“ sekvenci v rostlinné říši: Převládající „rostlinná“ sekvence `5'-TTTAGGG-3'` je v rámci rostlin pokládána za původní a sekvence `5'-TTAGGG-3'` je z ní odvozená. Shoda s telomerami obratlovů je tedy evolučně vzato pouhou homoplázií. Rovněž u jiných rostlin, které mají zcela jiné telomerové motivy (*Cestrum*), se předpokládá ztráta původní „rostlinné“ telomery, důvodem tohoto předpokladu je mimo jiné přítomnost proteinů schopných vázat klasickou „rostlinnou“ telomerovou DNA [79].

„Obratlovčí“ telomerová sekvence je velmi hojně zastoupena i mezi houbami (*Fungi*). Dle fylogenetické taxonomie jsou *Fungi* i *Animalia* řazeny do jedné velké skupiny *Opisthokonta*[17] [94]. Na základě toho se lze domnívat, že je mezi oněmi sekvencemi homologický vztah.

1.1.4 Sekundární struktura telomer

Doposud byla řeč o telomerách z hlediska jejich primární struktury (sekvence). Pro úplnost výkladu je však třeba stručně zmínit i jejich vyšší strukturu a jejich interakce s proteiny. Jak již bylo uvedeno, jednou z funkcí telomer je odlišení konce lineárního chromozomu od dvouřetězcových zlomů DNA. Z toho plyne, že telomera musí vykazovat strukturní odlišnost oproti poškozené dvoušroubovici[18]. Jinými slovy, pouhá přítomnost několika kilobází telomerová sekvence by nebyla dostačující pro správné

skripci do RNA a poté reverzní transkripci do DNA (jeho „životní cyklus“ je podobný životnímu cyklu retrovirů).

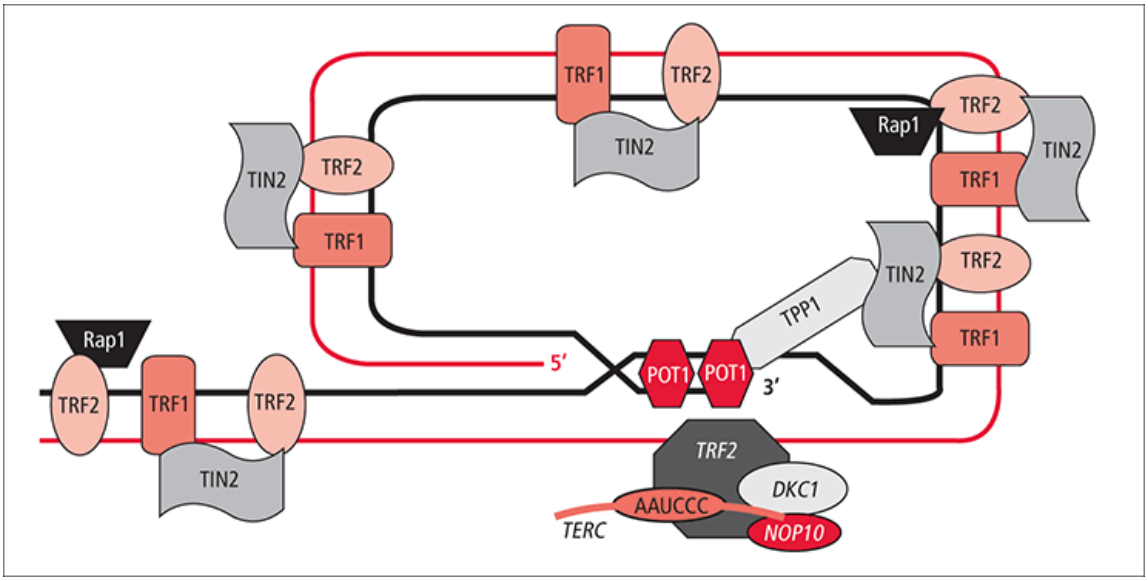
[15] Samozřejmě neplatí, že druhy se stejnou telomerovou sekvencí tvoří vždy monofyletický taxon – vždyť např. *Homo sapiens*, *Aspergillus fumigatus* i *Zostera marina* mají tutéž telomerovou sekvenci.

[16] Tranzice je typ bodové nukleotidové substituce, tedy typ mutace, a to takové, při níž se purinová báze (A, G) mění za purinovou či pyrimidinová (C, T) za pyrimidinovou. (Druhým typem bodové substituce je transverze, při níž se purinová báze mění za pyrimidinovou nebo pyrimidinová za purinovou.)

[17] *Fungi* a *Animalia* nejsou sesterské skupiny.

[18] Genomy obecně vykazují tendenci k cirkularizaci. To se týká většiny bakteriálních genomů, prokaryontních i eukaryontních plazmidů, většiny DNA v mitochondriích a plazmidech i DNA

fungování konce chromozomu [34]. Je známo, že přítomnost dvouřetězcových zlomů v DNA spouští regulační mechanismy zastavující buněčný cyklus a také reparační mechanismy snažící se oddělené konce zpětně spojit, což vede k chromozomovým fúzím, telomery takový účinek nevyvolávají a jeho důsledky naopak potlačují [109].



Obrázek 1.2: Struktura shelterinu savčího typu. Je patrný jednořetězcový přesah na konci 3' a jsou vyobrazeny interagující proteiny. zdroj: [121]

Telomerová struktura je dobře prostudována v případě savčích genomů. Tam dvoušroubovice DNA tvoří na konci 3' jednořetězcový přesah o délce v řádu desítek až stovek nukleotidů [72]. Tento přesah se stáčí nazpět a formuje tzv. telomerovou smyčku, zkráceně t-smyčku (t-loop) [34], pro lepší představu ji lze přirovnat ke konci lasa. Tato smyčka je stabilizována komplexem proteinů zvaným „shelterin" [51] nebo „telozom" [53], tento komplex se skládá z šesti různých typů proteinů, chrání telomeru před fúzemi konců chromozomů, reparačními mechanismy [102], [89] a exonukleázami[19] [77]. Tři proteiny (TRF1, TRF2 a POT1) se přímo vážou k telomerové DNA, zbylé tři (RAP1, TIN2 a TPP1) se vážou ke třem prve zmíněným proteinům a tím spolutvoří strukturu shelterinu [112]. Také ostatní eukaryonti chrání své telomery komplexem proteinů, jeho konkrétní podoba se však může odlišovat od popsaného savčího shelterinu [42]. Schematické znázornění shelterinu je na obrázku 1.2.

Podrobnější popis struktury shelterinu a jeho rolí v regulaci buněčných signálních drah není námětem této práce a lze jej najít např. v [97], [5], [98] nebo [72]. Právě zmíněná účast telomer v regulaci buněčného metabolismu, zajišťování stability genomu a souvislost s buněčným stárnutím činí jejich výzkum významným i z hlediska medicíny, včetně onkologie. Pro přehled viz např. [105], [7], [9], [32].

mnohých virů (i viry s lineární DNA se mohou replikovat přes kružnicové meziprodukty). Lineární chromozomy eukaryontů tedy představují jistou anomálii; jejím účelem je možnost meiózy [34].

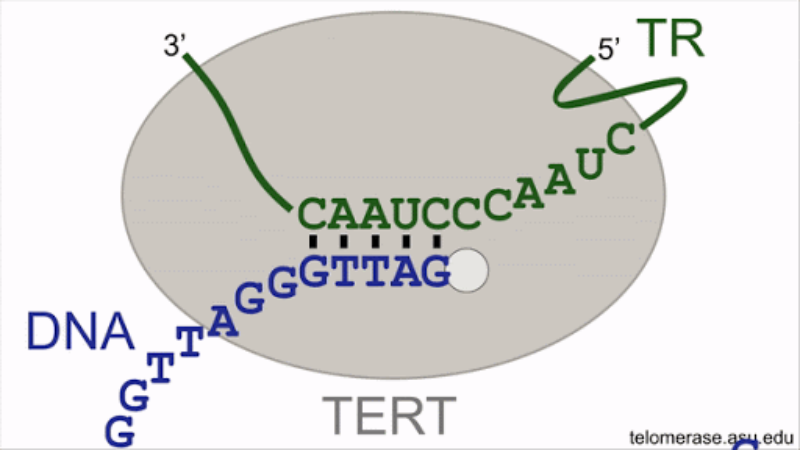
[19]Nukleáza je enzym štěpící nukleovou kyselinu. Exonukleáza odděluje jednotlivé nukleotidy od konce řetězce, zatímco endonukleáza štěpí řetězec „uprostřed".

1.1.5 Telomeráza

Telomerové repetice se zkracují při každém buněčném dělení a důsledkem tohoto zkracování chromozomů je buněčná senescence a buněčná smrt. Už jen proto, aby se život mohl předávat do dalších generací, musí existovat mechanismus, pomocí něhož jsou telomery prodlouženy a tím obnoveny.

Existuje více mechanismů prodlužování telomer a jejich různost bude ilustrovat sekce 1.1.6. V této podkapitole bude řeč o „kanonickém“ a patrně nejběžnějším způsobu prodlužování telomer. Zajišťuje jej enzym zvaný terminální transferáza nebo také telomeráza. Aktivita telomerázy byla poprvé zaznamenána v roce 1985, a to u téhož modelového organizmu, u kterého byl o několik let dříve popsán první telomerový sekvenční motiv (prvok *Tetrahymena*) [33].

Telomeráza se skládá ze dvou esenciálních podjednotek. Jednou z nich je telomerázová reverzní traskriptáza (TERT), katalytická podjednotka, která zajišťuje syntézu telomerové DNA. Druhou částí enzymu je telomerázová RNA-podjednotka (TERC, nebo také TR – „telomerase RNA“). Ta obsahuje templát pro syntézu DNA. Jinak řečeno: Konkrétní podoba telomerové repetice je určena sekvencí ribonukleotidů tvořících „templátovou část“ telomerázové RNA-podjednotky. TERC se dále účastní i procesu tvorby daného holoenzymu, jeho lokalizace v buňce a regulace jeho aktivity. S tím souvisí i existence konzervovaných sekundárních struktur TERC [116], [118], [20]. Právě existencí RNA-podjednotky se tak telomeráza liší od jiných reverzních transkriptáz, jaké jsou známy např. u retrovirů. Struktura telomerázy je znázorněna na obrázku 1.3 Telomerázu lze, co do její funkce, označit jako RNA-dependentní



Obrázek 1.3: Struktura telomerázy, modře vyznačen konec 3’ chromozomové ssDNA, zeleně TERC (zde popsáno TR „telomeráze RNA“), s vypsanými nukleotidy templátové oblasti, velký šedý ovál značí TERT, šedé kolečko je „kotevní místo“ („anchor site“) zajišťující interakce chromozomové ssDNA a TERT [111]; zdroj: [81]

DNA-polymerázu. Proces syntézy telomerové DNA probíhá tak, že volný konec 3’ chromozomu hybridizuje k templátové části TERC a katalytická podjednotka zajistí připojení několika nukleotidů tvořících jednu repetici. (TERC tedy musí být o něco delší než jedna telomerová repetice.) Enzym se pak posouvá, syntetizuje se další úsek telomery a celý proces se cyklicky opakuje [118].

Telomeráza nepatří k proteinům s konstitutivní[20] aktivitou. Ve většině somatických buněk tedy není aktivní. Telomerázovou aktivitu však vykazují silně proliferující buňky jako např. buňky zárodečných linií, kmenové buňky, případně buňky rakovinné. Z tohoto důvodu se telomeráza stává cílem kancerostatických léčiv [108].

1.1.6 Nestandardní způsoby prodlužování telomer

V předchozích odstavcích byla řeč o „kanonickém“ způsobu prodlužování telomer působením enzymu telomerázy. Pro úplnost zbývá zmínit se o ostatních známých procesech, jež mohou k obnově telomer sloužit.

Jak již bylo naznačeno v 1.1.2, jedním z organizmů, jejichž buňky nevykazují telomerázovou aktivitu, je *Drosophila melanogaster*. Její telomery jsou rovněž tvořeny repetitivními sekvencemi, ale délka jedné repetice je řádově tisíckrát větší než u „běžných“ eukaryontů – obsahují šest až dvanáct tisíc bází. Tyto telomery jsou udržovány pomocí retrotranspozonů[21]. Ty jsou tří typů: „Healing Transposon“ (HeT-A), „Telomere Associated Retrotransposon“ (TART) a „Telomere Associated and HeT-A Related“ (TAHRE) [74].

Vnitřní struktura daných repetic zahrnuje promotory[22], geny i regulační sekvence. Díky nim se mohou v genomu svépomocí replikovat a tím udržovat – jinak řečeno – enzymatická aktivita nutná k replikaci těchto retrotranspozonů je zajištěna geny v nich obsaženými [18]. Princip jejich replikace přitom v mnohém připomíná telomerázu – replikují se totiž přes ribonukleázový meziprodukt a dále reverzní transkripcí do DNA, která je poté začleněna do genomu. Tato podobnost s telomerázovým mechanismem naznačuje společný evoluční původ obou retroelementů [78].

Dalším příkladem organizmu s netypickým prodlužováním telomer je bourec morušový (*Bombyx mori*). Je zvláštní tím, že na koncích jeho chromozomů se nacházejí tisíce kilobází kanonické hmyzí telomery (`5'-TTAGG-3'`), ale TERT je v jeho buňkách exprimována v malé míře a její aktivitu není možné detekovat. Přitom se při koncích chromozomů nachází i velké množství subtelomerických retrotranspozonů. Taková spolupráce telomerázy a retrotranspozonů do telomerové oblasti lze z funkčního hlediska považovat za evoluční mezičlánek obou typů udržování telomer [57].

Třetím typem prodlužování telomer (vedle telomerázy a retrotransposonů) je tzv. alternativní prodlužování telomer („alternative lengthening of telomeres“, ALT). Tento mechanismus se vyskytuje u rakovinných buněčných linií – z nich asi 10 až 15 % vykazuje ALT, ostatní prodlužují telomery kanonickou cestou. Četnost ALT se liší v závislosti na původu daného nádoru – např. u mezenchymálních tumorů se ALT vyskytuje častěji než u epiteliálních. Výskyt ALT je také možné korelovat s přežitím pacientů, ale tato korelace se liší dle typu nádoru či skupiny pacientů, nejdená se tedy o jednoduchý, obecný vztah [117].

Mechanismus ALT je založen na homologní rekombinaci, proto může být indukován např. dvouřetězcovými zlomy DNA [87]. Pro buněčné linie s ALT jsou typické

[20] Stále stejnou, nepodléhající regulaci.

[21] Jiným typem mobilních elementů jsou tzv. DNA-transpozony, ty se ale při přesouvání v genomu nekopírují, pouze mění polohu, zatímco u retrotranspozonu se vždy vytvářejí nové kopie, zatímco templát zůstává na místě.

[22] Promotor je místem, kde začíná transkripce.

dlouhé a heterogenní telomerové motivy v rámci repetitivního úseku telomerové DNA s různými odchylkami od kanonické sekvence (např. `5'-TCAGGG-3'`, `5'-TGAGGG-3'` či `5'-TTGGGG-3'`) [104].

1.1.7 Telomerové transkripty a vmezeřené telomerové sekvence

Telomerové repetice nekódují žádné proteiny, přesto však dochází k jejich transkripci do RNA. Tyto transkripty se nazývají „telomeric repeat-containing RNA" neboli RNA obsahující telomerové repetice (TERRA) a jsou značně variabilní co do své délky. Existence TERRA byla objevena až v roce 2007 [4]. Sekvence TERRA má v případě kanonické telomery podobu `5'-UUAGGG-3'` [23].

V této souvislosti budiž zmíněno, že chromatin – komplex DNA a proteinů (zvláště tzv. histonů), který tvoří strukturu chromozomu – se může vyskytovat ve dvou stavech: jako „rozvolněnější" euchromatin, vyznačující se vyšší mírou transkripce, nebo „kompaktnější" heterochromatin s nižší mírou transkripce.[24] Telomerové transkripty vznikají v subtelomerických oblastech chromozomů, které vykazují právě povahu heterochromatinu[25] [23] (u samotných telomer je epigenetický status chromatinu dosud předmětem diskusí, ačkoli má znaky heterochromatinu, vyznačuje se dynamikou pro heterochromatin neobvyklou [22], [43]).

Promotory TERRA mívají povahu oblastí bohatých na CpG (tj. místa obsahující sekvenci `5'-CG-3'`; ta se např. v lidských buňkách nacházejí i v subtelomerových oblastech), nicméně i mezi těmito promotory existuje značná heterogenita, což může souviset se zmíněnou růzností délek TERRA [83].

Bylo zjištěno, že TERRA mohou asociovat s různými proteiny, jmenovitě i s proteiny shelterinu TRF1 a TRF2 (viz 1.1.4), stabilizují je a zprostředkovávají jejich další interakce. Úbytek TERRA může mít za následek špatnou funkci telomer [25], [23]. TERRA však mohou dle uvedených studií interagovat i s jinými proteiny heterochromatinu, úbytek TERRA byl asociován se snížením methylace histonů (TERRA interaguje s methyltransferázou zajišťující methylaci histonů). To naznačuje kauzální vazby mezi transkripcí telomerových repetic a organizací chromatinu, tedy mírou genové exprese. Metodami kvantitativní hmotnostní spektrometrie[26] bylo identifikováno množství proteinů, které se specificky vážou k TERRA [96], což naznačuje komplexitu regulací, jichž se tyto molekuly RNA účastní.

[23] Že se `5'-TTAGGG-3'` přepíše jako `5'-UUAGGG-3'`, a ne jako `5'-CCCUAA-3'` je dáno tím, že daná sekvence se vždy přepisuje z komplementárního C-bohatého řetězce DNA.

[24] Na přechod mezi oběma formami chromatinu mají vliv zvláště methylace histonů – podporuje heterochromatin – nebo acetylace histonů – podporuje euchromatin. Tyto chemické děje patří k důležitým epigenetickým regulačním mechanismům.

[25] V tomto ohledu byla zjištěna souvislost mezi mírou methylace subtelomerových oblastí chromozomu a mírou transkripce TERRA. [114]

[26] Hmotnostní spektrometrie (mass spectrometry, MS) je skupina analytických metod, jejichž principem je fragmentace vzorku zkoumané látky za vzniku iontů (pozitivně, či negativně nabitých – dle typu metody), ty jsou urychleny v elektromagnetickém poli, detekovány a je stanovena jejich hmotnost (přesněji – poměr hmotnosti a náboje). Podstatou je specifita frangmentace různých látek, čehož lze využít k jejich přesné identifikaci, a to i za podmínek extrémně malých množství vzorku.

Jako příklad lze uvést souvislost s aktivací opravných mechanismů v oblasti telomer (jakožto domnělých dvouřetězcových zlomů). U laboratorních myší bylo pozorováno, že úbytek exprese TERRA z jedné konkrétní telomery vedl k široké dysfunkci telomer napříč celým genomem dané buňky, a to právě kvůli aktivaci (nežádoucích) opravných mechanismů [99].

Z hlediska sekvenční bioinformatiky spočívá význam telomeromých transkriptů v možnosti identifikovat telomerové repetice také v transkriptomických datech. Této varianty však při screeningu nebylo plošně využito.

Závěrem této „epigenetické vsuvky" zbývá zmínit se o tzv. vmezeřených telomerových sekvencích (intersticial telomeric sequences, ITS). Jejich existence byla poprvé popsána již v roce 1992 [76].

Telomery se *ex definitione* vyskytují na koncích chromozomů. Bylo však zjištěno, že u více druhů obratlovců se příslušné telomerové motivy (`5'-TTAGGG-3'`) vyskytují nejen v terminálních částech chromozomů, nýbrž i na jiných místech. [16] Mohou být nalezeny blízko centromery[27] či přímo v její oblasti (tzv. (peri)centromerické ITS). Mohou se ale také nacházet v oblasti mezi centromerou a telomerou (takto jsou někdy chápány ITS v užším slova smyslu).

ITS mohou tvořit až desítky procent všech telomerových repetic daného genomu. Jedním z rozdílů oproti pravým telomerám je to, že ITS zřídka tvoří „dokonalé" tandemové repetice, jsou méně sekvenčně konzervovány než pravé telomery a běžné telomerové repetice jsou v nich prokládány nepravidelnými motivy [22].

ITS jsou asociovány s genomovou nestabilitou. Bylo zjištěno, že jde o frekventovaná místa (tzv. hotspots) chromozomových zlomů, rekombinací, přestaveb a jiných strukturních změn. Jejich sekvence se mohou v genomu amplifikovat, místa jejich výskytu jsou fragilnější a mají i podíl na regulaci genové exprese [16].

Přestože biologické funkce ITS dosud nejsou zcela objasněny, je patrné, že se skrze svou účast na působení genomové nestability podílejí i na evoluci chromozomů [16].

1.2 Sekvenování DNA

Účelem podkapitoly o sekvenování nukleových kyselin (stanovení pořadí jejich nukleotidů) je objasnit původ dat, s nimiž se zachází v praktické části práce.

První metody stanovení primární sekvence nukleových kyselin byly popsány už v 60. letech 20. století [38] Zprvu se jednalo o sekvenace malých molekul RNA a sekvenační metody byly založeny na úplné, nebo částečné degradaci těchto molekul [37].

1.2.1 Klasické sekvenační metody

První generaci sekvenačních metod představovaly metody vyvinuté ve druhé polovině 70. let. Jednou z nich bylo Maxamovo–Gilbertovo sekvenování, jehož principem

[27] Centromera je místem chromozomu, v němž jsou při buněčném dělení spojeny dvě sesterské chromatidy. Je také místem připojení mikrotubulů dělicího vřeténka při mitotické fázi buněčného cyklu.

je štěpení malé molekuly DNA na fragmenty, jejichž konce jsou opatřeny radioaktivními značkami. Délkami značených fragmentů[28] jsou pak určeny pozice jednotlivých nukleotidů na polyakrylamidovém gelu s rozlišovací schopností jeden nukleotid [55].

Až do nástupu sekvenování nové generace byla velmi hojně využívána metoda Sangerova sekvenování [90]. Principem je využití tzv. dideoxyribonukleotidů (ddNTP) – od běžných deoxyribonukleotidů (dNTP) v DNA se liší absencí hydroxylové skupiny na pozici 3'. Je-li taková báze při syntéze komplementárního řetězce začleněna do sekvence, komplementární řetězec už se v daném místě nemůže prodlužovat. Reakční směs obsahuje jak dNTP, tak i ddNTP – tyto se náhodně začleňují na různá místa v syntetizovaném řetězci, čímž ukončí jeho prodlužování. Tím jsou (pouze na základě pravděpodobnosti) získány řetězce „všech možných" délek. Fluorescenční nebo radioaktivní značení ddNTP umožní rozlišit koncový nukleotid každého fragmentu (zda jde o A, G, C, či T) a tím rekonstruovat původní sekvenci.

V 80. letech pak již vznikaly automatické sekvenátory schopné číst fragmenty dlouhé řádově stovky bází [2]. Pokud bylo třeba sekvenovat delší úsek DNA, byla použita metoda zvaná „shotgun sequencing". Ta spočívá v sekvenování dílčích fragmentů původní molekuly DNA. Tyto fragmenty se ovšem překrývají. Výsledek sekvenování je tak nutné sjednotit do jediné souvislé sekvence (tzv. kontig) na základě zjištěných překryvů [1]. K tomuto sestavování byly již „od počátku" užívány počítačové programy [101].

Během 80. let byla vyvinuta i jiná, velmi citlivá metoda, která nevyužívá fluorescenční ani radioaktivní značení, ale pracuje na principu chemiluminiscence – pyrosekvenování[29], jehož průkopníkem byl Pål Nyrén [68].

Podstatou je, že každé připojení nukleotidu je následováno emisí světla, intenzita emitovaného světla je úměrná počtu začleněných dNTP (vždy v počtu 0 nebo více). Metoda pracuje v cyklech, v každém cyklu je přidán vždy jen jeden typ dNTP a jejich typy se postupně střídají[30]. Metoda tedy *de facto* sleduje aktivitu DNA-polymerázy a z této informace rekonstruuje podobu sekvence [67], [40].

Podobně jako Sangerovo sekvenování, a na rozdíl od Maxamovy–Gilbertovy metody, pracuje pyrosekvenování na principu syntézy komplementárního řetězce. Přesnost této metody trpí tím, že při začlenění většího počtu (čtyř a více) nukleotidů roste intenzita emitovaného záření nelineárně, a jejich počet tak nemusí být správně detekován [85].

1.2.2 Sekvenování nové generace

Když se po přelomu tisíciletí ujala metody pyrosekvenování biotechnologická společnost *454 Life Sciences*, stala se tato metoda prvním zástupcem sekvenování nové

[28] Délku fragmentů lze měřit experimenty na bázi tzv. elektroforézy – migrace záporně nabitých molekul DNA gelem, jímž prochází stejnosměrný proud, rychlost migrace přitom závisí na velikosti daného fragmentu.

[29] Název pochází od sloučeniny pyrofosfátu – to je anion kyseliny difosforečné ($P_2O_7^{4-}$), tato molekula vzniká při každém začlenění dNTP do syntetizovaného řetězce DNA.

[30] V původním provedení bylo vždy třeba původní směs s nukleotidy vymýt. Tento poněkud nepraktický krok byl později optimalizován zavedením enzymatické degradace nežádoucích nukleotidů [85].

generace (NGS)[31]. První komerční platforma pro NGS vznikla v roce 2005 [75]. Obecný princip NGS spočívá v masivní paralelizaci sekvenačních reakcí [8] (myšleno po milionech). K provedení takového sekvenování musí být nejprve připraven soubor fragmentů DNA, každý z nich v mnoha kopiích, tento soubor fragmentů se nazývá knihovna DNA (DNA library).

Při sekvenování metodou 454 je DNA nejprve fragmentována a ke koncům jednotlivých úseků jsou připojeny krátké sekvence zvané adaptory. Každý z těchto fragmentů je posléze podroben klonální amplifikaci metodou emulzní polymerázové řetězové reakce (emPCR) [54]. Principem emPCR je příprava emulze vody v oleji – každá kapička vody obsahuje (má obsahovat) právě jeden unikátní fragment DNA a kuličku, na které je onen fragment jedním svým řetězcem imobilizován pomocí zmíněné adaptorové sekvence, celý tento objekt se někdy označuje jako „mikroreaktor“. Proces emPCR probíhá právě v těchto kapičkách (mikroreaktorech) a jeho výsledkem je, že každá kulička obsahuje řádově 10^7 identických kopií původního fragmentu[32] [64], [49], [54].

Připravená knihovna DNA pak může být umístěna na tzv. pikotitrační destičku (picotiter plate) obsahující jamky (picotiter wells)[33]. Do každé jamky by měla zapadnout právě jedna kulička s imobilizovanými (identickými) fragmenty DNA. Do jamek jsou poté přidávány další potřebné reaktanty, proběhne pyrosekvenování dle výše popsaného principu a emitované fotony jsou snímány čipem CCD, čímž je stanovena sekvence DNA.

Ve srovnání s jinými metodami NGS je protokol 454 velmi rychlý. Další výhodou je, že poskytuje velmi dlouhá čtení (přes 700 bp [56]). Nevýhodou je vysoká cena potřebných reaktantů a také již zmíněná nízká spolehlivost sekvenování homopolymerních úseků.

Metodě 454 je podobná platforma „Ion Torrent“. Ta je specifická způsobem, jakým je detekováno začlenění příslušného nukleotidu do vznikající sekvence. Většina ostatních metod k tomu využívá fluorescenci nebo jinou luminiscenci, zatímco metoda Ion Torrent využívá měření pH, inkorporace nukleotidu je totiž doprovázena uvolněním protonu. Ve srovnání s 454 umožňují absence optických komponent v sekvenátoru a elektronická detekce vznikající sekvence snížit náklady na sekvenaci a celý proces zrychlit. Metoda je schopna produkovat poměrně dlouhá čtení. Problém se sekvenací homopolymerních úseků však přetrvává [86].

Nejpoužívanější současnou metodou NGS je platforma Illumina[34]. Jejím principem je využití fluorescenčně značených terminačních nukleotidových analogů. Podobně funguje i tradiční Sangerovo sekvenování. V případě platformy Illumina ale nejsou použity nevratně terminační ddNTP, nýbrž nukleotidy opatřené odstranitelnou terminační skupinou.

Fluorescenční substráty enzymu DNA-polymerázy opatřené reverzibilně vázanou chránicí skupinou na konci 3’ byly popsány již v roce 1994 [17]. Trvalo však více jež

[31] Někdy se tyto metody označují i jako „high-throughput sequencing“.

[32] Jedná se o jednořetězcové kopie.

[33] Počet jamek na jedné destičce je v řádu jednotek milionů – tím je dána úroveň paralelizace příslušné sekvenační reakce – každá jamka je snímána zvlášť.

[34] Původně uvedena na trh firmou *Solexa*, ta záhy přešla pod firmu *Illumina*.

jedno desetiletí, než byl uvedený princip využit v komerčním sekvenátoru a než byl poprvé významně zužitkován pro NGS [93], [13].

Základní postup je podobný jako u pyrosekvenační platformy 454 – je třeba fragmentovat genetický materiál, provést jeho amplifikaci pomocí PCR a poté jej paralelně sekvenovat – opět se jedná o sekvenování syntézou komplementárního řetězce. Metoda opět začíná rozštěpením DNA a připojením krátkých sekvencí – adaptorů – ke koncům vzniklých fragmentů.

Pro klonální amplifikaci na rozdíl od sekvenování 454 není uplatněna emPCR, ale amplifikace na pevném povrchu. Amplifikace probíhá v reakční komůrce (flow cell), na jejímž povrchu jsou imobilizovány oligonukleotidy, jež posléze slouží jako primery pro PCR. Tyto oligonukleotidy jsou současně komplementární k adaptorům přichyceným k fragmentům DNA, které se díky tomu přichytí k povrchu reakční komůrky. Případné nepřichycené fragmenty jsou ze směsi odmyty. Ke směsi jsou poté přidány reaktanty potřebné pro PCR a enzym DNA-polymeráza syntetizuje komplementární řetězec k tomu přichycenému. Poté dojde k denaturaci, původní řetězec se odmyje pryč (protože nebyl imobilizován, pouze nekovalentně vázán ke komplementárnímu adaptoru) a reverzní řetězec zůstává.

Amplifikace dále probíhá metodou „přemostění“, ta spočívá v hybridizaci volného adaptoru syntetizovaného fragmentu ke komplementárním adaptorům, jimiž je povrch reakční komůrky hustě pokryt – aby k tomu došlo, musí se tento fragment ohnout a vytvořit strukturu podobnou mostu[35]. DNA-polymeráza potom syntetizuje komplementární řetězec, zvýšením teploty řetězce denaturují a cyklus se může opakovat. Výsledkem PCR je stav, v němž je povrch reakční komůrky pokryt shluky (clusters) fragmentů DNA. Všechny fragmenty v rámci jednoho shluku jsou identické (reverzní komplementy se odštěpí a odmyjí, zůstávají pouze přímá čtení) [49].

DNA-polymeráza vykazuje při začleňování jednotlivých nukleotidů jistou chybovost, inkorporací špatného nukleotidu mohou nastávat bodové substituce, a to zvláště na koncích 3’ [44].

Vlastní sekvenace probíhá přidáním směsi dNTP[36] s fluorescentní skupinou na pozici 3’. DNA-polymeráza začlení příslušný nukleotid, ale potom už žádný další, protože skupina 3’ je blokována fluoroforem. Teprve až po chemickém odštěpení chránicí skupiny dojde k připojení dalšího nukleotidu. Sekvenování tak může probíhat synchronizovaně [103]. Než je ale fluorofor odštěpen, kamera CCD identifikuje, o který nukleotid se jedná (každý nukleotid nese jiný fluorofor). Tím je stanovena sekvence DNA.

V porovnání s jinými metodami je platforma Illumina cenově příznivá. Jednotlivé sekvenační běhy ovšem poskytují jen kratší čtení – délka zde závisí na typu použitého přístroje, obvykle 200, 250, 300, někdy až 600 bp [49]. Dalším nedostatkem je vychýlené pokrytí sekvenovaných fragmentů bohatých na AT, nebo GC. Metoda také vykazuje poměrně vysokou chybovost [49]. Rovněž chemické odštěpení floroforu může na molekule DNA zanechat „jizvu“, která ztíží další průběh chemických reakcí [41]. Za svou rozšířenost Illumina vděčí hlavně zmíněné nízké finanční náročnosti

[35] V angličtině se proto tato metoda označuje „bridge amplification“.

[36] Při pyrosekvenování byly v každém cyklu přidány dNTP jen jednoho typu, u tohoto postupu se přidávají všechny čtyři typy současně.

a schopnosti produkovat velké objemy dat, které se při správně nastavených standardech mohou využívat i pro citlivá stanovení, např. klonální variabilitu tumorů, detekce SNP v širokém měřítku apod.

S výjimkou Maxamova–Gilbertova sekvenování byly všechny doposud zmíněné postupy založeny na sekvenování syntézou, tedy na použití DNA-polymerázy. Existují však i metody NGS spočívající v principu ligace – spojení nukleotidů jiným enzymem, DNA-ligázou[37].

Sekvenování pomocí ligace bylo poprvé uplatněno v roce 2005 u bakteriálního genomu a už tehdy byla pozorována jeho vysoká přesnost (podíl chyb okolo 10^{-6}) [91].

Protokol opět obnáší fragmentaci DNA, připojení adaptorů, klonální amplifikaci pomocí PCR a vlastní sekvenování. Amplifikace je prováděna podobně jako u metody 454 – je uplatněna emPCR. Fragmenty jsou upevněny na malých kuličkách, ty jsou imobilizovány na sklíčku a sekvenovány [82], [3].

Komerční platforma využívající sekvenování ligací se nazývá SOLiD[38]. Zatímco přípravou knihovny DNA může SOLiD připomínat 454, princip sekvenování je odlišný. Sekvenační reakce spočívá v použití oktamerů DNA[39]. Každý oktamer je tvořen dvěma bázemi DNA na konci 3' (je tedy třeba 16 různých typů oktamerů – pro všechny variace), dále třemi „univerzálními" bázemi (které hybridizují s libovolným templátem) a třemi nukleotidy nesoucími fluorofor jedné ze čtyř barev[40]. V prvním sledu dochází k hybridizaci oktameru k sekvenované molekule DNA. DNA-ligáza kovalentně spojí konec 3' daného oktameru s koncem 5' již připravené části komplementárního řetězce. Následuje detekce barvy příslušného fluoroforu a odštěpení tří nukleotidů nesoucích tento fluofor. Z oktameru tak zůstává pentamer dvou známých bází a tří univerzálních bází. Proces se poté opakuje.

Je patrné, že tímto algoritmem nelze získat úplnou informaci o dané sekvenci, protože nelze zjistit, k jakým bázím hybridizovaly univerzální báze v oktameru. Celou reakci je tak nutné ještě několikrát opakovat, při každém opakování se ale čtecí rámec posune o jeden nukleotid. Spojením dílčích informací ze všech iterací lze rekonstruovat původní sekvenci [92].

Sekvenování pomocí ligace se oproti jiným metodám vyznačuje vysokou přesností – vždyť z povahy metody musí být každý nukleotid přečten dvakrát, nejběžnější chybou je substitce nukleotidu [49]. Jeho slabinou je ovšem špatná schopnost sekvenovat delší palindromatické sekvence[41]. Tyto sekvence totiž tvoří nežádoucí sekundární struktury – tzv. vlásenky (hairpin) – jedna polovina vlásenky hybridizuje s její druhou polovinou, a kvůli tomu k nim už nemůže hybridizovat oktamer [39]. Nepřesnost

[37] Rozdíl mezi DNA-polymerázou a DNA-ligázou záleží v tom, že DNA-polymeráza kovalentně připojí nukleotid k části už hotového komplementárního vlákna, zatímco DNA-ligáza potřebuje, aby se k části téhož vlákna hybridizací (vodíkovými můstky) připojil nějaký oligonukleotid a ona poté vytvoří fosfodiesterovou vazbu mezi tímto oligonukleotidem a předchůdně syntetizovanou částí komplementárního řetězce.

[38] Sequencing by Oligonucleotide Ligation and Detection, uvedena na trh někdejší společností *Applied Biosystems* v roce 2007 [3].

[39] Oktamerem se zde rozumí oligonukleotid délky 8.

[40] Jedna barva tedy může znamenat čtyři různé variace příslušného dinukleotidu.

[41] Palindromatická sekvence se v přímém směru čte stejně jako v reverzním směru, např. `5'-TAGGCATGCCTA-3'`.

může být způsobena i „směsnými“ fragmenty DNA na jedné kuličce, přílišnou blízkostí kuliček při sekvenaci či ubýváním fluorescenčního signálu při opakování cyklu [3], [46]. Jiným nedostatkem platformy SOLiD je malá délka sekvenčních čtení [39].

Platformě SOLiD se podobá i sekvenační metoda „DNA Nanoballs“ vyvinutá někdejší firmou *Complete Genomics*. Tato metoda je unikátní provedením replikace DNA. Při přípravě knihovny totiž fragmenty DNA tvoří kruhové molekuly, jež jsou replikovány mechanismem otáčející se kružnice s využitím fágové polymerázy. Tím vzniknou dlouhé tandemové repetice daného fragmentu. Tyto úseky se uspořádají do podoby „nanokuliček“ a v této podobě dochází k jejich sekvenaci podobným principem jako u platformy SOLiD. Nevýhodou této metody je krátká délka sekvenčních čtení. [28]

1.2.3 Sekvenování třetí generace

Tato práce je postavena především na zpracování dat pocházejících z „klasických“ metod NGS. V posledních letech se ale rozvíjejí i postupy tzv. sekvenování třetí generace (third-generation sequencing, TGS), někdy se označuje i jako jednomolekulové sekvenování (single-molecule sequencing). Tyto metody se od NGS liší absencí amplifikační fáze. Tím se snižují jak finanční náklady, tak i chybovost způsobená amplifikací [49]. V databázi SRA se již nacházejí data pocházejících z TGS – zvláště pak z platforem PacBio SMRT a Nanopore. Jejich společnou vlastností je produkce velmi dlouhých čtení [37].

Kapitola 2

Metodika

Tato kapitola popisuje, jak bylo postupováno při provedení screeningu. Bude v ní uvedeno, které databáze tato sekvenční data obsahují a spravují, jak byla data extrahována z databáze a jak s nimi bylo dále zacházeno.

Jelikož se jedná o popis metodiky, není cílem specifikovat zde výsledky práce. Je však třeba říci, že mnohé metodické přístupy vznikaly průběžně jako reakce na dílčí mezivýsledky a potřebu jejich vylepšení, proto musejí být takové výsledky naznačeny už v této kapitole, ačkoli jejich systematický a úplný popis bude podán v kapitole 3.

2.1 Databáze sekvenčních dat

Veškerá dostupná sekvenační data jsou uložena v databázích zastřešených projektem „The International Nucleotide Sequence Database Collaboration“ (INSDC) [45]. Tato platforma zahrnuje tři velké databáze, mezi nimiž denně probíhá sdílení všech sekvenčních dat i dalších přidružených informací. Jsou to

- ENA (European Nucleotide Archive) pod EMBL-EBI (European Molecular Biology Laboratory's European Bioinformatics Institute),
- GenBank pod americkým NCBI (National Center for Biotechnology Information) [10],
- DDBJ (DNA Data Bank of Japan).

Vzhledem k tomu, že „fyzický“ obsah všech tří databází je identický, rozdíl mezi nimi je spíše otázkou uživatelského rozhraní a nabídky nadstavbových nástrojů.

V rámci platformy INSDC existuje rozsáhlá databáze zvaná Archiv sekvenčních čtení (Sequence Read Archive, SRA), která se rovněž podílí na zmíněné politice sdílení dat [48].

Databáze SRA se soustřeďuje na shromažďování „surových“ sekvenčních dat pocházejících z nejrůznějších laboratoří. Vysoká kapacita sekvenátorů a snižující se náklady na sekvenace mají pochopitelně za následek rychlý nárůst objemu dostupných dat. Valná většina těchto dat je přitom veřejně přístupná komukoli na Internetu. Kolem 80 % sekvenčních dat v SRA v současnosti pochází ze sekvenační platformy Illumina.

Kromě vlastního datového souboru je obvykle k dispozici i množství metadat zahrnujících informace o organismu, způsobu sekvenace a (důležitých) detailech jejího provedení, přispívající laboratoři apod.

Ačkoliv se nejedná o absolutně otevřenou databázi, protože její obsah je institucionálně spravován, relevance dat, správná anotace a spolehlivost údajů v metadatech je ponechána přispívajícím autorům dané studie s případným odkazem na publikace. Tento způsob je skvělá příležitost pro sdílení vstupních dat, ověřování jejich využití, a případně recyklaci pro studie jiného zaměření než bylo původní. Taková otevřenost systému nicméně přináší i riziko tzv. zmatečných souborů (např. kontaminace houbovými sekvencemi v sekvenačních datech muzejních a herbářových položek, záměna pojmů genome/transcriptome, příprava sekvenační knihovny random/selection, selection/size selection atd.)

Databáze obsahuje i obecnější metadata, např. ohledně taxonomie všech dostupných organismů. V době přípravy dat pro screening (podzim 2018) v ní bylo zhruba čtyři sta tisíc eukaryontních genomů, objem dat však stále rychle roste.

2.2 Tandem Repeats Finder

Tandem Repeats Finder (TRFi) je program určený k vyhledávání tandemových repetic v sekvenčních datech [11][42]. Existuje i řada jiných podobných programů, např. Repeat Explorer, ale v rámci návaznosti na předchozí projekty a již zčásti implementovanou nástavbu nad TRFi byl pro analýzu použit právě tento program [66]. Program TRFi poskytuje internetové uživatelské rozhraní pro zpracování sekvencí ve formátu FASTA (`https://tandem.bu.edu/trf/trf.submit.options.html`). Sem mohou uživatelé své sekvence přímo kopírovat, případně nahrávat z připravených textových souborů. Takový postup je ale vhodný pro malé objemy dat a jeho automatizace by byla neschůdná. Proto byla pro screening použita volně dostupná instalace ve verzi pro Linux.

MetaCentrum, v jehož rámci byly analýzy prováděny, poskytuje TRFi v jednom ze svých modulů, v rámci této práce ale byla použita vlastní kopie programu, protože spouštění předchystané verze provázely technické obtíže. Při screeningu byla použita verze programu 4.07b[43].

2.2.1 Princip a algoritmus programu TRFi

TRFi funguje na principu sekvenčního přiložení (sequence alignment) [113] (str. 31 až 35), [24] (str. 28 až 30). Alignment je metoda k určení míry podobnosti dvou sekvencí.[44] Dvě přiložené sekvence mohou vypadat např. takto:

```
GTGGCGTGGAATATTATTAGGC---
GTC-CGTGGTA-ATTATTCGGCAAA
```

[42] Z citované Bensonovy práce [11] (popřípadě ze stránek TRFi pro vyšší aktuálnost – [12]) pocházejí téměř všechny informace o TRFi uvedené v těchto sekcích.

[43] Existuje již novější verze 4.09, která se liší schopností zacházet s centromerovými sekvencemi, ty však pro tuto práci nejsou zásadní.

[44] V tomto případě nukleotidových, ale přikládat lze i sekvence jiné, např. proteinové.

Často se rozlišuje párové přiložení (pairwise alignment, pouze dvě sekvence) a vícenásobné přiložení (multiple alignment, obecně více než dvě sekvence). Kvalita alignmentu se hodnotí pomocí různých skóre. To je kvantita zohledňující počet shod (matches), neshod (mismatches) a mezer (gaps) v přiložených sekvencích. Každý z těchto jevů aditivně přispívá k celkovému skóre. Konkrétní „kladné body“ za shody nukleotidů, případně penalizace za různost sekvencí, jsou do značné míry arbitrární záležitost. Skóre proto může sloužit spíše k porovnání více různých alignmentů a nalezení nejlepšího nežli k „separátnímu“ určení objektivní kvality konkrétního alignmentu.

Protože je sekvenční přiložení (zvláště vícenásobné) výpočetně náročné, softwary nehledají zcela optimální alignment, ale využívají heuristických metod dynamického programování.

Algoritmy sekvenčního přiložení se dělí na lokální a globální. Máme-li dvě sekvence podobné délky, které si mají vzájemně odpovídat (např. dvě sekvence téže funkce z příbuzných organismů), můžeme použít globální přiložení (tak je tomu např. v příkladu uvedeném výše). Může však nastat situace, že chceme krátkou sekvenci přiložit k nějaké mnohem delší sekvenci. Přitom nás zajímá jen to, nakolik přesně je kratší sekvence obsažena v té delší. Použití globálního přiložení by nevyhnutelně vygenerovalo velmi dlouhé mezery, což by bezdůvodně snižovalo skóre alignmentu. V takovém případě použijeme lokální přiložení, viz příklad:

```
GCATTAGCAGTCTTTGGTTA--TTACGTGGAGGGATTCGATCCGTT
           TTAGGTTAGGTTAGG
```

Sekvenční přiložení (ve své lokální verzi) hraje zásadní roli i v TRFi: Program nejprve pomocí statistických kritérií najde kandidátní repetice (dle původní Bensonovy terminologie tzv. detekční část programu) a takto vygenerované úseky se pokusí přiložit k původní sekvenci (analytická část programu). Jestliže má výsledný alignment dostatečné vysoké skóre, program repetici vyhodnotí jako přítomnou v datech a zapíše do výstupního souboru.

Detekce tandemových repetic probíhá na základě skenování vstupní sekvence kratším úsekem (sondou – „probe“), přičemž program hledá úplné shody použité sondy se vstupní sekvencí. Přitom jsou zaznamenávány indexy nukleotidů, u nichž se daná sonda vyskytla. Při jejím opakovaném výskytu se pak hodnotí vzdálenosti indexů nukleotidů a tato vzdálenost se případně stává délkou periody příslušné repetice. Použití statistických kritérií k vyhodnocení přítomnosti repetice souvisí s tím, že TRFi uplatňuje jistou toleranci vůči „nepravidelnostem“ v repetitivních sekvencích.

Jestliže nalezený kandidát na repetici projde statistickými testy, analytická část TRFi se pokusí daný vzor přiložit k sekvenci. Pokud jsou k sekvenci úspěšně přiloženy alespoň dvě opakování daného vzoru a přiložení má dostatečně vysoké skóre, program tuto sekvenci stanoví jako repetitivní a zapíše do výstupního souboru.

2.2.2 Nastavení parametrů TRFi

Program je spouštěn z příkazové řádky příkazem `trf`, při použití vlastního spustitelného binárního souboru je volán přímo název souboru (včetně případné cesty

k němu – `trf409.linux64`, `trf407b.linux64` apod.), za tímto příkazem následují parametry. Příklad volání programu je:

```
$ ./trf409.linux64 ./SRR7003228.fasta 2 7 7 80 10 50 35 -h -ngs
```

- Prvním nutným parametrem je název vstupního souboru (`SRR7003228.fasta`). V uvedeném příkladě se předpokládá, že spustitelná kopie programu i vstupní datový soubor se nacházejí v témže adresáři, který je současně aktuálním pracovním adresářem.
- Další tři parametry definují nastavení alignmentu (`2 7 7`). Dokumentace je nazývá „match“, „mismatch“ a „delta“. Matoucí může být skutečnost, že shoda zde má váhu 2, zatímco neshoda či indel 7 – jde o to, že hodnoty 7 algoritmus interpretuje jako záporná čísla, tedy jako -7. Uvedené hodnoty jsou doporučené v dokumentaci programu, mohou však být změněny.
- Hodnoty `80` a `10` se týkají předem stanovených pravděpodobností shody ($p_{\mathrm{M}} = 0{,}8$), resp. indelu ($p_{\mathrm{M}} = 0{,}1$). Pravděpodobností model detekční části TRFi s nimi pracuje v rámci statistických kritérií. Změna těchto hodnot může dle dokumentace několikanásobně zvýšit čas potřebný pro výpočet.
- Další parametr (`50`) se nazývá „minscore“. Je to nejmenší skóre alignmentu, při jakém je daný repetitivní motiv reportován do výstupu. Pro příklad uvažme tento alignment s 28 shodami, jednou neshodou a jednou mezerou:

  ```
  TTTAATTTATTTTACCCTAA-CCTAACCCTAACCCTAAGCCTAAC
                      CCCTAACCCTAACCCTAACCCTAACCCTAA
  ```

  Jeho skóre je při nastavení parametrů např. `2 3 7`

  $$28 \cdot 2 + 1 \cdot (-3) + 1 \cdot (-7) = 56 - 3 - 7 = 46.$$

  Při tomto nastavení (`minscore=50`) by repetice `CCCTAA` nebyla programem identifikována. (Je asi zřejmé, že míra „benevolence“ při hodnocení aligmentů v TRFi je převážně věcí uživatelových zkušeností a preferencí.)
- Poslední povinný parametr je nazýván „maxperiod“ (zde `35`). Je to maximální délka periody repetice, kterou program uvažuje. Minimální délka je vždy 1. Maximální hodnota tohoto parametru je 2000[45]. Dle zkušeností při výpočtech se jeho snížením celé zpracování i několikanásobně urychlí.

Program dále nabízí použití přepínačů, ty však již nejsou povinné. Mezi nimi např.:

- Volba `-m` vede k vygenerování tzv. maskovaného souboru („masked sequence file“), v němž jsou nukleotidy, které byly v rámci analýzy zařazeny do tandemové repetice, nahrazeny znakem `N`.

[45] Je třeba uvážit, jaká je délka sekvenčních čtení a podle toho parametr snížit.

- Volba `-h` potlačuje výstup ve formátu HTML.
- Přepínač `-u` vypíše nápovědu.
- Parametr `-v` vypíše verzi TRFi.
- Volba `-ngs` je určena zvláště pro vstupní datové soubory s velkým množstvím dílčích sekvencí. Výstup je pak kompaktnější, sekvenční čtení bez repetic nejsou v souborech `.out` za účelem šetření paměti vůbec uváděny.

2.2.3 Výstupy programu TRFi

Hlavním výstupem programu je soubor s příponou `.out`, který obsahuje informace o zpracovaných sekvenčních čteních. Jestliže byla v daném čtení identifikována tandemová repetice, program ji zapíše do výstupního souboru, a to včetně různých souvisejících kvantitativních údajů. Tento výstup je ale jednak velmi objemný (zvláště, byl-li původní soubor `.fasta` velký jednotky až desítky GB) a jednak poměrně špatně čitelný. Níže je uvedena ukázka takového výstupu (jedná se o genom motýla *Papilio ambrax*, který se vyskytuje v Oceánii):[46]

```
@SRR5879299.481 481 length=200
158 187 1 30.0 1 100 0 60 100 0 0 0 0.00 a~AAAAAAAAAAAAAAAAAAAAAAAAAA
AAAA AGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT GATGAACAATACA
@SRR5879299.483 483 length=200
69 103 7 5.0 7 85 0 52 17 25 28 28 1.97 TATGCGC TATGCGCTATGCGCTATGCAC
TATGCGGTATGCGC TATGCGATTTTCGAGTTTCGACTTTCGAGTTTTTGATAAGCGATACGCGT GTA
TCGCTTATCAAAAACTCGAAAGTCGAAACTCGAAAATCGCGTATCGC
163 200 7 5.4 7 87 0 58 31 26 23 18 1.97 ATAGCGC ATAGCGCATACCGCATAGTG
CATAGCGCATAGCGCATA ATCAAAAACTCGAAAGTCGAAACTCGAAAATCGCGTATCGCGTTATGCAA .
@SRR5879299.485 485 length=200
1 138 2 72.0 2 82 8 90 36 2 1 59 1.19 TA TATATATATATATATATATATATATA
TATATATATTATAATATTTGTATATTTTTTTTTTTTTTCTATTATCTATATTGTATTATTTTTCTTAT
TTATATATATATATATATATATATATATATATATATATATA . ATTGAGAGTTGTGCGTGTAGAGAGT
GTGTGTCTATCTTTTTGATCGATAT
@SRR5879299.487 487 length=200
46 90 7 6.4 7 84 0 63 15 24 28 31 1.96 TATGCGC TATGCGCTATGCGCT
ATGCGCTATGCACTGTGCGGTATGCGCTAT GATTTTCGAGTTTCGACTTTCGAGTTTTTGATAAGCGA
TACGCGT TTGCATAACGTATCGCTTATCAAAAACTCGAAAGTCGAAACTCGAAAATC
```

Jsou zde uvedeny výsledky pro čtyři čtení: 481, 483, 485 a 487. Čtení s čísly spadajícími do uvedeného rozmezí, která zde nejsou uvedena, program nevypsal proto, že v nich nenalezl žádnou repetici. Každý výstup začíná symbolem `@`, názvem zpracovávaného souboru, identifikátorem sekvenčního čtení a údajem o jeho délce.

Následuje tabulka s údaji o nalezené repetici. Tabulka bude interpretována např. pro čtení č. 483 (to je zde pro srovnání uvedeno ve své původní, vstupní podobě):

[46] Jde o kompaktnější výstup za použití přepínače `-ngs`.

```
>SRR5879299.483 483 length=200
GCTATTTGCATAACGCGATATGCGATTTTCGAGTTTCGACTTTCGAGTTTTTGATAAGCGATACGCGTTA
TGCGCTATGCGCTATGCACTATGCGGTATGCGCGTATCGCTTATCAAAAACTCGAAAGTCGAAACTCGAA
AATCGCGTATCGCGTTATGCAAATAGCGCATACCGCATAGTGCATAGCGCATAGCGCATA
```

Číselné údaje `69 103 7 5.0 7 85 0 52 17 25 28 28 1.97` mají následující význam:

- První dvě čísla (`69`, `103`) jsou indexy nalezené oblasti s repeticí, jsou vztaženy k začátku daného sekvenčního čtení.
- Třetí číslo (`7`) představuje délku periody, s níž se daná repetitivní sekvence opakuje.
- Čtvrtá položka (`5.0`) může být chápána jako délka celého úseku tvořeného repeticemi v jednotkách počtu period. Přesněji – je to počet kopií nalezené repetice, které byly přiloženy k originální sekvenci.
- Pátý údaj (`7`) značí délku konsenzu v původní sekvenci, jenž byl použit pro alignment s nejlepším skóre. Obvykle má stejnou délku jako je délka periody přikládané repetitivní sekvence, ale může se i lišit.
- Šesté a sedmé číslo (`85` a `0`) znamenají procenta shody v daném alignmentu, resp. procenta malých inzercí a delecí (indelů).
- Osmý údaj (`52`) představuje souhrnné skóre alignmentu.
- Další čtyři položky (`17`, `25`, `28`, `28`) udávají procentuální zastoupení jednotlivých nukleotidů v dané repetici.
- Poslední údaj (`1.97`) je entropie vypočtená z nukleotidového složení příslušného úseku. Nejvyšší entropie bude dosaženo při zcela rovnoměrném zastoupení nukleotidů, naopak homopolymerní úsek (např. „polyA“) má nulovou entropii.

V přehledu uvedeném výše byla vypisována čísla příslušející motivu `TATGCGC`. Ve čtení 483 však byla identifikována ještě jiná repetice – `ATAGCGC`. Při prohlédnutí počátečního úseku původního čtení si lze všimnout náznaku třetí repetice – `TTTCGAG`, její přítomnost však program nevyhodnotil jako významnou, proto ve výstupu není uvedena. Srovnáním popsaného výstupu s výstupem pro čtení 481, v němž se nachází dlouhý homopolymerní úsek adeninů, je patrné, že tento úsek byl programem vyhodnocen jako repetice délky 1.

Zbývá dodat, že dalším výstupem programu jsou soubory s alignmenty, na základě nichž byly příslušné repetice identifikovány.

2.2.4 Tandem Repeats Merger

Tandem Repeats Finder není nastavitelný pro paralelní zpracování většího množství datových souborů. Navíc v případě velkého množství čtení v jednom vstupním souboru (u NGS běžně jednotky, ba desítky milionů) postrádá stručné a přehledné shrnutí nalezených tandemových repetic.

Z těchto důvodů byl již dříve na školitelově pracovišti vyvinut nástroj Tandem Repeats Merger (TRMe) určený pro dodatečné zpracování výstupů z TRFi [100]. Je schopen paralelně zpracovat i desítky datových souborů. Kromě standardních výstupů TRFi poskytuje seznamy nalezených tandemových repetic včetně počtů čtení, v nichž byly identifikovány, a souhrnné tabulky těchto četností pro všechny zpracované soubory. Tento nástroj má i funkce, jež nebyly při screeningu využity – jde o porovnávání obsahu tandemových repetic ve vzorcích ošetřených nukleázou BAL-31 štěpící telomery a ve vzorcích kontrolních. To nebylo využito, protože v databázích zpravidla nejsou vhodná data k takovému porovnávání.

Tato nástavba posloužila jako základ pro provedení screeningu. TRMe byl však programován s ohledem na zpracování spíše malých skupin souborů, pročež jeho použití (zvláště pro velká množství souborů) obnáší některá úskalí:

- Názvy datových souborů je v něm nutné (manuálně) zadávat do jednoho ze skriptů.
- Program generuje výstupní soubory s velmi dlouhými názvy. OS Linux je však opatřen limitem pro délku názvu souboru. Pokus o paralelní zpracování velkého množství souborů tedy skončí chybou.
- Pokud z jakéhokoli důvodu program skončí chybou, uživatel neobdrží žádné výstupní soubory. Výpočty totiž probíhají v úložišti scratch.
- Pro potřeby širokého screeningu mnoha souborů nejsou výstupy optimální – dílčí seznamy repetic s četnostmi obsahují některé sekvence redunantně, tabulky jsou místy špatně formátované (u palindromatických sekvencí).
- Paralelní zpracování mnoha datových souborů je výpočetně neefektivní, vykazuje velmi špatné škálování procesorů. Délka výpočtu se totiž odvíjí od doby trvání nejkomplikovanějšího výpočtu, zatímco mnohé jiné už jsou hotové a procesory, na kterých byly spuštěny, pak běží „naprázdno“.

Z uvedených důvodů byl TRMe v rámci screeningu využit, ale bylo nutné alespoň částečně obejít zmíněné problémy a upravit jeho výstupy.

2.3 Implementace screeningu

Již před tímto projektem byly uplatňovány přístupy ke hledání tandemových repetic založené na TRFi a TRMe, převážně s využitím služeb MetaCentra. V rámci této práce bylo třeba navrhnout analýzu tak, aby byly zpracovány řádově tisíce genomů (pokud možno) všech různých eukaryotních druhů dostupných v SRA.

2.3.1 Výběr vstupních dat

Již v době spouštění screeningu obsahovala databáze stovky tisíc datových souborů se sekvenovanými genomy. Výpočetní kapacita v kombinaci s dostupnými softwarovými nástroji a časovým horizontem bakalářské práce neumožňuje zpracovat veškerá dostupná data. Lze však provést racionalizaci a vhodným filtrováním a částečným vzorkováním dat screening uskutečnit v masivním rozsahu ve smyslu zastoupení jednotlivých druhů.

Uživatelské rozhraní databáze SRA (v rámci serveru https://www.ncbi.nlm.nih.gov) nabízí možnosti filtrace dat podle různých kritérií. Jestliže je filtrovaných souborů relativně malý počet, lze jejich seznam zobrazit v interaktivním prostředí „SRA Run Selector“.V něm se zobrazí tabulka datových souborů, údaje o jejich počtu a velikosti, je možné pokračovat ve filtraci dat, řadit je dle jednotlivých kritérií (druhu, sekvenační platformy, přispívajícího pracoviště, velikosti, ...), případně manuálně vybírat konkrétní datové soubory pro zařazení do konečného seznamu. Výstupem tohoto rozhraní je tzv. „SRA Accession List“ – textový soubor o jednom sloupci obsahující identifikátory zvolených datových souborů, např:

```
SRR2154279
SRR4031068
SRR3571026
...
SRR7699527
SRR7784022
```

Alternativně lze získat i celou tabulku s metadaty (tzv. „SRA Run Table“ ve formátu `.csv`), která zahrnuje všechny zvolené soubory.

Tento způsob výběru datových souborů je vhodný pro menší projekty zaměřené na relativně úzkou skupinu organismů, a to ze dvou důvodů:

1. Do rozhraní „Run Selector“ lze odesílat jen malé skupiny souborů (řádově stovky), jsou-li jich tisíce a víc, server je zahlcován a nepracuje optimálně.
2. Žádný z nabízených filtrů neobsahuje volbu, která by v rámci jednoho kritéria (např. druhu) automaticky vybrala zvolený počet souborů (např. právě jeden) od každé varianty. Právě to však bylo pro screening zapotřebí.

Proto byl výběr dat proveden níže popsaným způsobem.

1. V rámci webového rozhraní NCBI byly vybrány soubory obecně vyhovujících kritérií:
   - Byly filtrovány jen eukaryontní organismy.
     dotaz: `("Eukaryota"[Organism] OR eukaryotes[All Fields])`
   - Byla zahrnuta jen veřejně dostupná data.
     dotaz: `cluster_public[prop]`
   - Přestože telomerové sekvence lze teoreticky zaznamenat i v transkriptomických datech, screening byl omezen na molekuly DNA.
     dotaz: `"biomol dna"[Properties]`

- Byly uvažovány jen celogenomové sekvenace, tj. ne např. jen sekvenace konkrétní skupiny genů, zvolené části chromozomu apod.
  dotaz: `"strategy wgs"[Properties]`
- Telomerová sekvence je v některých sekvenačních experimentech z principu ignorována (např. sekvenace, při nichž se vytváří knihovna z „vybraných“ restrikčních fragmentů klonovaných do knihovny BAC, sekvenace „vybraných“ restrikčních fragmentů o konkrétní jednotné délce, sekvenace genomových úseků „vybraných“ využitím PCR nebo jiné amplifikační metody atd.). Proto byly uvažovány jen ty experimenty, v nichž byla příprava knihovny fragmentů DNA provedena na způsob náhodného výběru (tj. sekvence ve výsledné knihovně byly zastoupeny optimálně jen vzhledem k frekvenci výskytu v genomu, vychýlení způsobené např. vyskokým obsahem párů GC a podobného typu je v postupu zanedbáváno). Je třeba uvést, že poskytovatelé hrubých dat zde často chybují. Parametr PCR tu správně znamená, že vstupní populace fragmentů DNA se omezuje na specifickou frakci genomových sekvencí. Přitom chtějí jen vyjádřit, že příprava knihovny obsahuje kroky opatřování adapterů pomocí PCR a není tzv. PCR-free. Další chybou bývá, že pro genomová data uvádějí „size selection“, což by připadalo v úvahu u PCR dat, nebo celotranskriptomických dat. „Size selection“ v tom případě snižuje „náhodnost“ vstupních sekvencí vzhledem k celému genomu. Poskytovatelé chtějí vyjádřit, že sekvenační knihovna obsahovala fragmenty DNA v určitém rozsahu, což je nutné pro efektivní využití sekvenační technologie, nicméně výběr cílových sekvencí je zde náhodný. Ke zmatení pojmů dochází i „opačným směrem“, ale pro účely této práce byly použity jen datové soubory apriori označené random.
  dotaz: `random[Selection]`
- Filtrace na použitou sekvenační platformu nebyla provedena.

  Celý dotaz měl tedy tento tvar:

  ```
  ("Eukaryota"[Organism] OR eukaryotes[All Fields]) AND
  (cluster_public[prop] AND "biomol dna"[Properties] AND
   "strategy wgs"[Properties]) AND random[Selection]
  ```

2. Přes tlačítko „Send to“, následnou volbu „Choose destination“ jako „File“, a vybraný formát „RunInfo“ byl vynenerován soubor typu `.csv` o zhruba čtyř stech tisících řádcích. Ten obsahoval všechny dostupné informace o všech filtrovaných datových souborech.

3. Finální příprava seznamu souborů byla provedena pomocí skriptu v jazyce `Python`. Vyvstala otázka, na základě které charakteristiky rozlišit různost druhů – nabízel se latinský název, nebo parametr „TaxID“, jímž je v databázi každému taxonu přiřazeno jednoznačný číselný identifikátor. Výhodou TaxID je jeho přesnost – zadané latinské názvy občas obsahují překlepy či jiné chyby. Problém ale spočívá v tom, že i různé varianty téhož druhu (poddruhy apod.)

se mohou lišit v TaxID a screening by měl při takové filtraci až příliš jemný záběr. Proto bylo nadále zacházeno s latinskými názvy jako identifikátory druhů. Skript procházel tabulku, na prvním řádku zařadil název druhu do iniciovaného seznamu. Na dalším řádku porovnal nový název druhu se seznamem druhů, nebyl-li takový druh dosud zaznamenán, zařadil jej tam a zapsal řádek do výstupního souboru, jinak řádek přeskočil. V rámci textového řetězce označujícího název druhu pak byly vždy uvažovány jen první dvě slova (oddělená bílými znaky), protože mnohdy bylo sekvenováno mnoho poddruhů jednoho druhu, ty však nebylo nutné všechny do screeningu zahrnovat. Tímto způsobem bylo vybráno 5152 řádků. Sloupce s přístupovými kódy k příslušným souborům pak byly extrahovány pomocí nástroje `awk`.

2.3.2 Přístup k datům

Rozhraní databáze SRA popisovaná v 2.3.1 neposkytují (zjevně dostupnou) možnost přímého stažení sekvenčních dat. Je to dáno jejich velikostí – již relativně úzká filtrace souborů může snadno nabýt velikosti jednotek i desítek TB. Data je možné přímo stahovat pomocí adresy URL, kterou lze získat z rozhraní „Run Selector“, resp. z tabulky „SraRunInfo“ (viz 2.3.1). Jejím zadáním do prohlížeče se celý datový soubor začne stahovat na lokální počítač, což je pochopitelně značně nepraktické.

Data by bylo možné stahovat přímo na úložiště NGI příkazem `wget`, ale pro přístup k datům (i jejich případné úpravy, převádění mezi formáty atp.) slouží především již připravený nástroj `SRA Toolkit` [120]. Nejnovější verzi programu lze získat z adresy https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software Starší verze byly dostupné pro OS Linux, Windows i MacOS, nejnovější je již podporována pouze pro Linux.

Návod k instalaci a konfiguraci je dostupný na adrese https://ncbi.github.io/sra-tools/install_config.html. Nápověda je dostupná na https://www.ncbi.nlm.nih.gov/books/NBK158900/, dokumentace na https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc.

V rámci tohoto programu slouží ke stahování dat ze SRA příkaz `prefetch`. Jeho použití však provázely technické obtíže a chyby, proto bylo k datům přistupováno poněkud nekanonickým způsobem – pomocí příkazu `fastq-dump`. Ten primárně slouží k formátování a jiným úpravám dat, ale dokáže rovněž stahovat soubory z databáze. Navíc umožňuje nastavit počet stažených čtení pro příslušný datový soubor.

Při screeningu byla používána verze softwaru `sratoolkit.2.9.2-ubuntu64`. Příkaz `fastq-dump` (i ostatní příkazy) se nacházejí v podadresáři `./bin`. Je-li pracovní adresář nastaven na zmíněný adresář se spustitelnými příkazy, lze je volat přímo zadáním jejich názvu do terminálu, případně včetně cesty. Data byla stahována tímto příkazem:

```
$ ./fastq-dump  -X 10000000 -Z --fasta SRR2154279 > ./SRR2154279.fasta
```

Cesty k příkazu a výstupnímu souboru je v konkrétním případě vhodné modifikovat. Argument `-X` stanovuje maximální počet zapsaných čtení, v tomto případě nastaveno

na 10 milionů, což (velmi orientačně) může odpovídat třem GB stažených dat (jde-li o kratší čtení z platformy Illumina). Přepínač `-Z` určuje, že data mají být zapsána jako standardní výstup (ne např. jako archiv), argument `--fasta` definuje formát výstupních dat. Následuje přístupový kód k datovému souboru. Symbol `>` slouží k přesměrování výstupu do libovolného výstupního souboru, jehož název (včetně případné cesty) je uveden za tímto symbolem.

Bylo zmíněno, že data jsou získávána ve formátu FASTA[47]. Takový soubor může vypadat např. takto (jde o genom pšenice (*Triticum aestivum*)):

```
>SRR1170582.1 1 length=300
CCCATGCCTTCTTCTTCCTCTGCTCATCCATTCTCCCTCACTTGAGATCTAGCGCCCAAGCTTTCACCTT
CTCCGCAAAGTCATTTCAAGTATGTCCGGAGCGGGAGGCAAGTGGATGGTCTCCTCCGTTACGGAGGAGG
ACATCACGAAAAGGCCTCGCACACGATGATAAACACCGAGATGTTGAGGATGAAATTGGGGGCCAGATCA
TGAAACTCCAGCCTGTAGTAGAACATAAGCCCGCGGATGAATGGGTGGACAGGGAATCCCAGTCCACGGA
CAAAGTGAGAGAGAAACACT
>SRR1170582.2 2 length=300
TGCAAACCGACCACCACACCCATGTCGTCTACCGACAAGATAACAACTGTTGATGGTGAGCTTTTGTCTC
CTGCGGATGCCACAGAGTACAGAAGCATTCTTGGTGGACTTCAGTACTTGACGATCACGAGACCAGATAT
CTCTTATGCTAGCAAAGACTGCACCCAAATAATCTCTGCAGTAGCATTAGCCACAGCCTTGTACTCAGCT
TCAGTACTGCTACGTGACACAGTAGCCTGTTTCCGAGCACTCCAGGCGATCAAATTAGAGCCAAAGAATA
CTGCATAACCCCCCGTGGAT
>SRR1170582.3 3 length=300
AGTCCTTATGCTCACCCAGCAGCTGATGCAGACAAATCTTTGCACATTACTAAGCAAGTAAGCATGCCAC
CGTGATTCAAATATCTCCGTTGATAGGCGATTCAAAATATACCTAGTTCGATCTGTTAGAATACGTATAG
GATAGAGATATGCGGGAGCACGACGGCCTTCCATAGAAGGAAATTTCCCCTTGATAGCTTCTCCTGCGGT
GGTGGTCCGAGAGGAACGGAGCTGCTGCTGGATGATGATGTGAAGGGGGAGGATGTGGGGGGAGAAAGCG
CGCTCATGGTGGACGATAGG
...
```

Každé čtení má svůj identifikační řádek. Ten vždy začíná symbolem `>` a poté (bez mezer) následuje identifikátor daného členi. V případě souborů ze SRA má podobu názvu souboru a pořadového čísla čtení oddělených tečkou. Vše, co se nachází za prvním bílým znakem (mezerou), je komentář. Na dalším řádku následuje vlastní sekvence. Tento vzor se pak v celém souboru stále opakuje.

Je třeba poznamenat, že pořadí čtení v souboru `.fasta` nijak nesouvisí s tím, ze kterého chromozomu dané čtení pochází, ani s tím, jaké je jeho pořadí v genomu vůči ostatním fragmentům. Existují sice metody pro rekonstrukci genomu (ve smyslu lokalizace dílčích úseků získaných ze sekvenace), ale pro tento projekt jsou stěžejní nekódující repetice, které pro kompletaci genomu představují spíše problém. V této práci proto o rekonstrukci genomů z dat NGS nebylo nijak usilováno.

2.3.3 Spuštění a běh screeningu

MetaCentrum bylo využito jednak jako zdroj výpočetních kapacit, jednak i jako velký úložní prostor (3 TB na čelním uzlu `alfrid`).

[47] Jiné běžné formáty pro ukládání sekvenčních dat jsou např. FASTQ či SAM.

K přístupu do MetaCentra byly využity programy

- `PuTTY` pro práci v terminálu a
- `WinSCP` pro kopírování souborů na vzdálený server a z něj zpět.[48]

Základem provedení screeningu bylo jeho rozdělení na velké množství dílčích menších úloh, které byly během delší doby postupně odesílány do dávkového systému a jejichž výsledky byly po skončení screeningu shromážděny a souhrnně vyhodnoceny.

Konkrétní postup byl následující:

1. Vzhledem k tomu, že druhy zařazené do screeningu netvořily žádné přirozené, přiměřeně velké skupiny, byly fixně rozděleny na úseky po 10 datových souborech. Celkem bylo 516 takových skupin. Tak vznikly textové soubory (až na poslední) o 10 řádcích, které byly uloženy vlastního adresáře. To bylo provedeno ve skriptovacím jazyce `bash`.[49]
2. Bylo vygenerováno 516 adresářů, každý z nich určený pro jednu dílčí úlohu.
3. V každém z těchto adresářů byly vytvořeny prázdné adresáře `data/` a `res/` určené pro vstupní data (soubory `.fasta`), respektive výsledky TRFi.
4. Skripty, v nichž není třeba nastavovat nic specifického pro jednotlivé úlohy, byly přímo kopírovány do všech 516 adresářů.
5. TRMe obsahuje skript `myScriptTRF.sh` v němž je třeba nastavit čtyři parametry:
   - Volba `data_dir` určuje cestu k adresáři se vstupními daty. Ve všech případech bylo nastaveno `data_dir="data"`.
   - Parametr `shortName` značí název dané úlohy (z pohledu TRMe, nikoli MetaCentra). Soubor byl generován tak, aby byl tento název shodný s číslem úlohy (1 až 516), např. tedy `shortName="423"`.
   - Proměnná `minNumberOfRepeats` stanoví, jaký musí být minimální počet opakování daného motivu, aby jej TRMe zahrnul do finálních výsledků. Nastaveno pevně na `minNumberOfRepeats="3"`.
   - Parametr `minLengthOfPeriod` definuje minimální délku periody tandemové repetice, aby byla ve výstupech TRMe uvažována. Bylo nastaveno `minLengthOfPeriod="3"`, homopolymerní úseky (např. `TTTTTT`) ani opakující se dinukleotidy (např. `CTCTCTCT`) tak nebyly uvažovány. Obvyklá délka periody telomerových repetic je totiž v rozmezí 5 až 10 bp[50].

[48] Jestliže jsou do MetaCentra kopírovány (textové) soubory vytvořené pod OS Windows, je třeba dbát na správná zakončení řádků – oba operační systémy používají různá, vzájemně nekompatibilní, a tato skutečnost se stává zdrojem chyb typu `syntax error near unexpected token`. Zakončení řádků lze ošetřit pod Windows v pokročilejších editorech (např. `PSPad`) nebo v Linuxu různými způsoby – např. příkazem `sed` nebo využitím editoru `vim` je lze hromadně přepsat na „správná“ zakončení.

[49] Skript s názvem `rozdeleni_datasetu.sh`

[50] Existují ale i delší motivy: 12 u *Allium*, 25 u kvasinek rodu *Kluyveromyces* apod.

6. TRMe rovněž používá skript `join.sh`, v němž je třeba zadat názvy zpracovávaných souborů ve tvaru

   ```
   names = ( SRR1800839 SRR1800842 SRR1800844 SRR1800846 SRR1800848
   SRR1800849 SRR1800852 SRR1800860 SRR1800869 SRR1800856 )
   ```

   Tyto soubory bylo třeba automaticky vygenerovat. Byly proto zvlášť vytvořeny příslušné řádky a ty poté spojovány s „invariantními“ zbytky skriptu.[51]

7. Dalším krokem bylo odesílat po skupinách úlohy do dávkového systému MetaCentra. K tomu slouží příkaz `qsub`[52]. Pomocí něj lze v NGI rezervovat výpočetní zdroje (Tyto požadavky lze specifikovat při zadávání příkazu `qsub`, anebo přímo v dávkovém skriptu.), na nichž je potom spuštěna příslušná úloha[53]. Je třeba specifikovat zejména následující[54]

   - Počet „výpočetních jednotek“ a počet procesorů se určuje skrze parametry `select`, resp. `ncpus`, např.:

     ```
     -l select=1:ncpus=10
     ```

     V případě, že úloha v nějakém okamžiku spotřebuje více procesorového času, než je násobek aktuální doby jejího běhu a počtu rezervovaných procesorů, plánovací systém vynutí ukončení úlohy. Proto je třeba požádat o jeden procesor za každý datový soubor zpracovávaný během jedné úlohy.
   - Potřebné množství paměti se nastavuje volbou `mem` a dočasné úložní místo v úložišti scratch volbou `scratch_local`, např.:

     ```
     mem=90gb:scratch_local=70gb
     ```

     Pokud by úloha spotřebovala více paměti, než kolik bylo rezervováno, je ukončena.
   - Je třeba nastavit maximální čas běhu úlohy parametrem `walltime` ve tvaru `HH:MM:SS`[55], např. tedy volba

     ```
     -l walltime=24:00:00
     ```

     představuje jeden den trvání úlohy. Pokud úloha neskončí během této lhůty, je ukončena plánovacím systémem. V závislosti na požadované době výpočtu je úloha zařazena do některé z front a čeká na spuštění. V případě TRFi/TRMe se jeden den ukázal být jako dostačující doba. Některé datové soubory se zpracovávaly i mnohem kratší dobu, a to v závislosti na velikosti souboru a komplexitě genomu (vyšší komplexita genomu znamená menší množství tandemových repetic).

[51] Viz skripty v adresáři `generovani`.

[52] Nápověda např. na https://wiki.metacentrum.cz/wiki/How_to_compute/Batch_jobs [119] (odkaz funkční 21. 2. 2020, dokumentace k příkazu dostupná přes `$ man qsub`.

[53] Spouštět dlouhotrvající úlohy přímo na čelních uzlech je zakázáno.

[54] https://wiki.metacentrum.cz/wiki/How_to_compute/Requesting_resources, [119]

[55] První dvoj/trojčíslí značí počet hodin, druhé počet minut, třetí počet sekund. Počet hodin může být i značně vysoký, např 480:00:00, ale taková úloha bude velmi dlouho čekat na spuštění, zvláště, je-li žádáno o mnoho procesorů.

- Existují i mnohé další parametry příkazu `qsub`. Týkají se např. požadavků na spouštění úlohy na konkrétních klastrech (volba `cluster=`), konkrétních operačních systémech (volba `os=`) či v konkrétní lokalitě[56]. Případně lze nastavit název úlohy (parametr `-N`) nebo zasílání informačních e-mailů ohledně stavu úlohy (volba `-m`).

8. Byl zvolen postup separátního odesílání úloh pro stahování dat a pro jejich následnou analýzu. Pokud jde o stahování dat, byl k tomuto účelu připraven skript, jenž vygeneruje zdrojový kód dávkového skriptu, odešle jej do plánovacího systému a poté jej odstraní. Data pro jednu každou dávkovou úlohu byla získávána ze zmíněných souborů obsahujících seznam deseti datových souborů (viz krok 1). Požadavky na zdroje byly specifikovány takto:

```
#PBS -N d$cislo_ulohy # 1 az 516
#PBS -l select=1:ncpus=1:mem=30gb:scratch_local=50gb
#PBS -l walltime=01:30:00
```

Ne každý datový soubor se skutečně stáhl během požadovaných devadesáti minut. V takovém případě bylo do doby ukončení úlohy staženo méně dat, ale pro následnou analýzu v TRFi jejich konkrétní množství nebylo klíčové. Ve všech případech bylo nastaveno zapsání maximálně 10 milionů čtení, což by měl být dostatečný počet pro identifikaci telomer, byl-li genom fragmentován a sekvenován opravdu náhodně. Skript je schopen odeslat najednou i více úloh (více „desetic“ datových souborů) – čísla datových souborů, jež mají být staženy, se nastavují jako parametry při jeho spouštění, např. tedy příkaz

```
$ bash download_files.sh 341 360
```

generuje dávkové skripty `d341.sh` až `d360.sh`, z nichž každý obsahuje uvedené požadavky na zdroje, přímo je odesílá do plánovacího systému (spouští příkazy `$ qsub d341.sh` až `$ qsub d341.sh`), čímž stahuje skupiny souborů od čísla 341 do čísla 360 včetně (celkem tedy 200 datových souborů).

9. Jakmile byla data připravena, byly spouštěny skripty pro TRFi. Opět byl připraven jeden obecný skript (`main.sh`), jehož parametry představovaly čísla zpracovávaných skupin. Např. tedy:

```
$ main.sh 341 360
```

V rámci tohoto skriptu byl změněn pracovní adresář na adresář dané úlohy, ta byla odeslána příkazem

```
qsub -l select=1:ncpus=10:mem=90gb:scratch_local=70gb
-l walltime=24:00:00 -N TRFi$cislo_ulohy myScriptTRF.sh
```

[56] Geografická blízkost míst, mezi nimiž probíhá např. výměna dat, může ovlivnit rychlost této výměny.

a totéž bylo provedeno pro všechny skupiny ze zadaného rozsahu (zde 341 až 360).

10. Po skončení běhu úlohy byly výsledky dostupné v adresáři `res/` (viz bod 3).

11. Výsledky ze všech dílčích adresářů byly pomocí jiného skriptu[57] shromážděny za účelem dalšího zpracování.

Základním výstupem screeningu byly soubory o názvech ve tvaru `ERR1802055_ppr.txt`, jež obsahovaly identifikované tandemové repetice a počet čtení[58], v nichž byla daná repetice identifikována. Příklad (jde o *Nicotiana tabacum*):

```
8338 TAAACCC
3646 CTC
2891 AAT
2171 TTG
2139 TTTTAGGATATGAATTGAAC
1910 CTTTT
...
```

2.3.4 Identifikace známých telomer

Ve stavu, při němž jsou k dispozici tisíce souborů se záznamy o četnostech tandemových repetic v genomech, bylo třeba naprogramovat skript, jenž by tyto záznamy automaticky prošel a stanovil (ne)přítomnost známých telomerových motivů.[59]

Zatímco práci se soubory v rámci samotného screeningu bylo možné pohodlně implementovat v jazyce `bash`, následné analýzy zahrnující i poněkud složitější (logické) konstrukce, byly takřka výhradně pováděny v jazyce `Python`.

Většina tandemových repetic nemá v genomu žádnou zásadní úlohu, a proto neexistuje (selekční) důvod, proč by se měly v genomu udržovat v extrémně hojné míře. Z tohoto úhlu pohledu vynikne srovnání se sekvencemi telomerovými, které jsou pro fungování buňky nutné, a tedy je jejich abundance v genomu obvykle mnohem vyšší, než jak je tomu u ostatních repetic. Identifikace telomer proto vychází z předpokladu, že ve srovnání s jinými repetitivními sekvencemi jich bude v příslušném datovém souboru velmi mnoho.

Bylo třeba sestavit seznam motivů telomerových sekvencí, s nimiž mají být výstupní soubory porovnány. Tento seznam byl založen na webových stránkách http://telomerase.asu.edu, některé sekvence byly ještě doplněny z jiných zdrojů. Seznam neobsahuje nutně všechny kvasinkové motivy, které jsou velmi dlouhé a variabilní. Byly na něj tedy zařazeny tyto sekvence:[60]

[57] Viz skript s názvem `shromazdit_vysledky.sh`.

[58] Budiž zdůrazněno, že ačkoli je „to číslo před každou sekvencí“ pro stručnost nazýváno abundance, nejde o celkový počet těchto repetic v genomu, ale počet sekvenčních čtení, v nichž TRFi danou tandemovou repetici identifikoval.

[59] Viz skript s názvem `telomery_v_datech.py`.

[60] Velmi dlouhé a variabilní motivy, které lze na seznamu vidět, patří kvasinkám.

```
TTAGGG
TTGGGG
TTGGGT
TTTTGGGG
TTTAGGG
TTAGG
TTAGGC
TGTGGGTGTGGTG
TCTGGGTG
GGGGTCTGGGTGCTG
GGTGTACGGATGTCTAACTTCTT
GGTGTACGGATGTCACGATCATT
GGTGTAAGGATGTCACGATCATT
GGTGTACGGATGCAGACTCGCTT
GGTGTAC
GGTGTACGGATTTGATTAGTTATGT
GGTGTACGGATTTGATTAGGTATGT
TCAGG
TTGCA
TTAGGGTCAACA
TTGTGG
TTTTAGGG
AATGGGGGG
TTTTAGGG
TTTAGG
TAGGG
TTTTTTAGGG
CTCGGTTATGGG
```

Je třeba podotknout, že tento přístup přináší jisté obtíže:

- Sekvence se mohou v datech vyskytovat jako přímé (forward) čtení, nebo jako reverzní čtení (reverse complement).
- Sekvence se může na výstupu objevit ve kterékoli své cyklické permutaci, např. přímá a reverzní čtení kanonické sekvence mohou mít tyto podoby:

  ```
  TTAGGG   CCCTAA
  GTTAGG   CCTAAC
  GGTTAG   CTAACC
  GGGTTA   TAACCC
  AGGGTT   AACCCT
  TAGGGT   ACCCTA
  ```

  Je třeba zohlednit, že se ve všech uvedených případech jedná o tutéž jednotku repetitivní sekvence, pouze zapsanou jiným způsobem. I když ale program nalezne sekvenci v „neobvyklé“ cyklické permutaci, vypsat by ji měl v „kanonické“ formě (pro účely přehlednosti).

- Další problém se týká už samotného pojetí „neobvyklé“ telomery. V rámci tohoto prohledání je jako neobvyklá nutně brána jen sekvence zcela nová. Je ale možné, že se u nějakého taxonu vyskytuje sekvence sice známá, ale typická pro zcela jiné organismy, jako příklad lze uvést „obratlovčí“ telomeru u některých rostlin (viz 1.1.3). Takové abnormality nebudou plánovaným postupem detekovány a k jejich případnému odhalení bude muset sloužit ještě práce s výstupními daty.
- Další otázka souvisí s tím, zda za „obvyklé“ telomery nepovažovat jen tři čtyři nejběžnější motivy a sekvence typu `CTCGGTTATGGG`[61] filtrovat jako neobvyklé. Tato filtrace však byla nastavena spíše „přísněji“, a to i s ohledem na očekávanou vysokou míru falešné negativity. (Ostatně – implementace byla navržena tak, aby i méně obvyklé telomery ze seznamu byly vypsány.)

Posledním aspektem je nastavení prahu pro „přítomnost“ či „nepřítomnost“ dané telomerové sekvence. Je přitom nutné zohlednit zejména kontaminace vzorků. Jestliže byl vzorek správně sekvenován, telomerová sekvence by měla být mezi dominantními motivy. Potom mohou taková data vypadat např. takto:

```
1733 GGGTTTA
1678 TCTT
482 AAG
210 TTCTTACT
152 TAAGAAAGT
...
```

V prvním záznamu můžeme rozpoznat standardní rostlinný motiv `TTTAGGG`, pouze v jiné cyklické permutaci.

Na to, že telomerová sekvence bude na výstupu první v pořadí však nelze spoléhat, protože genom může být špatně navzorkovaný, případně v něm z jiného důvodu skutečně převažuje ještě jiná repetice, např.:

```
184 AGCCA
172 TGCGACATTGGCCT
122 GCAAGGCCAAGGTC
92 GCCGCATGAGC
84 CACATACT
77 AAGTCCACCGC
67 TCATGCGGTGC
63 GTTTAGG
...
```

Opět jde o genom rostliny. Zde však vidíme rostlinnou telomeru až na osmém místě a její abundance oproti jiným několikanásobně zaostává. V rámci tohoto hodnocení přítomnosti telomer proto zvolen práh 5 % nejhojněji zastoupené sekvence jako hranice mezi přítomností telomery a „náhodnou kontaminací“, tedy její nepřítomností.

Výsledky tohoto zpracování budou popsány v kapitole 3.

[61] Jedná se o „cibulovou“ telomeru objevenou nedávno na PřF MU a BFÚ AVČR [29].

2.3.5 Implementace identifikačního postupu

Prvotní výstupy TRFi/TRMe byly zatíženy duplicitními výskyty tandemových repetic, a to kvůli existenci jejich reverzně-komplementárních protějšků. TRMe sice provádí sloučení napříč cyklickými permutacemi, neslučuje však sekvenci s jejím reverzním komplementem (alespoň u těch souborů, s nimiž bylo dále pracováno).

Za tímto účelem (kromě hlavního účelu identifikace známých telomer) byl napsán skript, jenž tento nedostatek odstranil.[62] K tomu bylo využito několik nově definovaných funkcí:

1. Funkce `reverzni_komplement()`, jejímž parametrem je sekvence obsahující pouze znaky `A`, `C`, `T`, `G`. Návratovou hodnotou funkce je opět textový řetězec – reverzně-komplementární sekvence k sekvenci zadané. Příklad:

```
>>> reverzni_komplement('ACGTT')
    'AACGT'
>>>
```

2. Funkce `generuj_cykl_permutace()`. Jejím vstupem je opět sekvence, funkce vrací seznam všech cyklických permutací této sekvence i jejího reverzního komplementu. Příklad:

```
>>> print(generuj_cykl_permutace('TTAGG'))
    ['TTAGG', 'GTTAG', 'GGTTA', 'AGGTT', 'TAGGT', 'CCTAA', 'ACCTA',
     'AACCT', 'TAACC', 'CTAAC']
>>>
```

3. Funkce `test_cykl_permutace(vzor, test)`, má dva argumenty – vzorovou a porovnávanou sekvenci. Funkce stanoví, zda porovnávaná sekvence je, nebo není reverzním komplementem sekvence vzorové. Vrací logické hodnoty `True`, nebo `False`. Patrně nejjednodušší způsob, jak toto testovat, je položení podmínky:

```
...
vzor_rev = reverzni_komplement(vzor)
if (len(vzor) == len(test)) and (test in vzor*2 or test in vzor_rev*2):
...
```

   Příklady použití funkce:

```
>>> test_cykl_permutace('TTAGGG', 'GGTTAG')
    True
>>> test_cykl_permutace('TTAGGG', 'ATCCCA')
    False
>>> test_cykl_permutace('TTAGGG', 'ACCCTA')
```

[62] Viz skript s názvem `redukce_duplicit.py`.

```
    True
>>> test_cykl_permutace('TTAGGG', 'TTAGG')
    False
>>>
```

Tyto funkce byly spolu s dalším kódem využity pro dosažení obou uvedených cílů. V případě redukce duplicitních reverzních komplementů bylo rovněž přistoupeno k jistému výpočetnímu zjednodušení podle podmínky:

```
if (radek > 2000) or ((abundance < 10) and (radek > 100)):
    break
```

Tj., skript neprocházel řádky výstupních souborů s pořadovým číslem vyšším než 2000, případně řádky s abundancí tandemové repetice pod 10 byly ignorovány, bylo-li jejich pořadové číslo větší než 100. Důvodem těchto zjednodušení bylo, že skript musí *de facto* porovnat každou sekvenci s každou (kvadratická výpočetní složitost pro každý datový soubor), což při (deseti)tisících řádků znamená, že jediný datový soubor může být takto zpracováván i celé minuty (a celkem máme tisíce souborů). Při uvedeném zjednodušení výpočet trval „pouze“ několik hodin (bez využití NGI).[63]

2.3.6 Taxonomické řazení a jiné úpravy výstupních dat

Předchozí algoritmus sloužil převážně k rozlišení „pozitivních“ datových souborů, v nichž byla identifikována některá ze známých telomer, resp. „negativních“, v nichž taková nalezena nebyla.

Hlavním výstupem byl seznam negativních výsledků, jenž posloužil jako vstup do další fáze analýzy. Byly rovněž vygenerovány i textové soubory obsahující první řádky abundancí tandemových repetic u všech negativních datových souborů, jež bylo možné „zkusmo“ pocházet a „od oka“ hledat zajímavé kandidáty. Několik jich sice bylo tímto způsobem nalezeno, ale vyvstala nutnost podívat se na data ještě jiným způsobem – uspořádaně dle systému eukaryontní taxonomie. Tímto způsobem je totiž možné vidět výsledky v různých souvislostech, nalézt případné skupiny datových souborů, jež jsou si vzájemně podobné co do (ne)výskytu nějaké sekvence. Zajímavé by bylo i nalezení změněné telomerové sekvence v rámci nějakého malého taxonu (např. na úrovni čeledi, rodu apod.).

Původní seznam datových souborů byl zapsán v pořadí určeném údaji o výběru souborů získanými ze SRA (viz 2.3.1). Tento seznam ale sám o sobě nemá dobré taxonomické pořadí, je převážně „neuspořádaný“.

Kýžené taxonomické řazení by se teoreticky dale provádět i ručně, ale u několika tisíc datových souborů by to trvalo příliš dlouho, a zvláště při opakování screeningu by bylo nutné opětovné seřazení. Proto musel být nalezen automatický postup seřazení. K tomu posloužila metadata databáze „Taxonomy“ [122], jež je přidružena k NCBI. Příslušná data lze získat na stránce ftp://ftp.ncbi.nih.gov/pub/

[63] Řečené se týká odstranění redundantních reverzních komplementů, samotná identifikace telomer podle seznamu je pochopitelně mnohem rychlejší (zhruba lineární výpočetní složitost).

`taxonomy` (dostupné příp. pod volbou „Taxonomy FTP“ z domovské stránky databáze).

Princip řazení spočívá na těchto předpokladech:

1. Každý taxon, jehož osekvenovaný genom je v SRA, má svůj záznam v databázi Taxonomy.
2. Každému takovému taxonu odpovídá jednoznačné identifikační číslo TaxID.
3. V metadatech databáze Taxonomy existuje soubor, v němž jsou veškerá přiřazená TaxID zapsána, a to v „uspořádaném“ pořadí.

Takový soubor se podařilo najít pod názvem `fullnamelineage.dmp` (tehdy (říjen 2019) měl velikost asi 530 MB a více než dva miliony řádků), v němž jsou pro každý taxon zapsány všechny jeho taxony nadřazené. Kromě toho jsou samozřejmě uvedena i příslušná identifikační čísla a řazení taxonů bylo vyhovující – příbuzné taxony blízko sebe. Tento soubor[64] byl proto použit jako templát pro seřazení výsledků HTS.

```
$ head -5 fullnamelineage.dmp # prvnich pet radku souboru
1       |       root    |               |
131567  |       cellular organisms      |               |
2157    |       Archaea |       cellular organisms;     |
1935183 |       Asgard group    |       cellular organisms; Archaea;    |
1936272 |       Candidatus Heimdallarchaeota    |       cellular organisms;
 Archaea; Asgard group;      |
$ awk '{print $1}' fullnamelineage.dmp | head -8 # extrakce TaxID
1
131567
2157
1935183
1936272
2026747
1841596
2012493
```

Tato data byla využita k napsání skriptu[65], jehož vstupem byl původní seznam zpracovaných datových souborů a výstupem byl podobný seznam, ale přístupové kódy nyní již nebyly řazeny dle implicitních nastevení SRA, nýbrž taxonomicky.

Takto získaný seznam pak bylo možné využít ke dvěma dalším výstupům:

1. K přegenerování původních výstupů shrnujcích (ne)přítomnost telomerových repetic. Ukázka (z „oblasti“ hlodavců – ve všech případech si lze všimnout standardní telomerové sekvence):

[64] Velké množství organismů v něm uvedených jsou Prokaryota, příp. jiné skupiny, které do HTS nebyly řazeny, což však ničemu nevadí.

[65] Viz skripty ve složce `taxonomicke_serazeni`.

```
...
SRR5192253;Callosciurus prevostii;TTAGGG;1223;
SRR5192250;Exilisciurus exilis;TTAGGG;3360;TTTAGG;261;
SRR5192246;Lariscus insignis;TTAGGG;14541;
SRR5192213;Rhinosciurus laticaudatus;TTAGGG;10111;
SRR5192206;Sundasciurus;TTAGGG;3758;
SRR5192209;Sundasciurus brookei;TTAGGG;5935;
SRR5192207;Sundasciurus lowii;TTAGGG;1715;
SRR5192215;Ratufa bicolor;TTAGGG;8592;
SRR099459;Ictidomys tridecemlineatus;TTAGGG;3423;
SRR7878800;Marmota flaviventris;TTAGGG;273472;
SRR7613949;Marmota vancouverensis;TTAGGG;7766;TTAGG;473;
...
```

2. Předchozí výstup poskytl jen přehled toho, zda byla v datech nalezena známá telomera, nikoli to, jaké další repetice byly identifikovány. Proto bylo třeba vytvořit tabulku, v níž každý sloupec odpovídal jednomu datovému souboru (tj. jednomu druhu) a v řádcích byly jednotlivé tandemové repetice, ať už telomerové, nebo jiné. V každém řádku byly repetice seřazeny podle své četnosti.[66]

Další funkcí, kterou bylo třeba naprogramovat, bylo vytvoření tabulky podobné předchozímu bodu (2.), ovšem až na to, že jednotlivé tandemové repetice ve sloupcích nebudou řazeny podle abundance v konkrétním datovém souboru, nýbrž podle součtu abundancí ve všech datových souborech[67]. To umožní snadné vizuální srovnání toho, u kterých datových souborů se daný motiv vyskytuje, případně v jakém množství. Tato funkce byla rovněž naprogramována a použita na některé menší skupiny genomů (mimo hlavní HTS).

2.3.7 Ověřování výsledků na nových datech

Na základě průběžných výsledků vyšla najevo nečekaně vysoká míra negativních výsledků, tedy datových souborů, v nichž nebyly identifikovány žádné známé telomerové sekvence.

Z tohoto důvodu bylo rozhodnuto přistoupit ke druhé vlně HTS a automaticky zpracovat všechny kandidáty na absenci standardní telomerové sekvence. Opět byl získán (aktualizovaný[68]) seznam všech vhodných eukaryontních datových souborů a byl vygenerován nový seznam seznam přístupových kódů, a to podle těchto kritérií:

- Pro každý zkoumaný druh bylo zjištěno, zda se v databázi nacházejí jiné vhodné datové soubory od téhož druhu.

[66] Celá tato tabulka má v textovém formátu velikost přes 60 MB.

[67] Tento postup předpokládá alespoň řádovou srovnatelnost jednotlivých datových souborů co do počtu všech čtení s nějakou tandemovou repeticí.

[68] Bylo zjištěno, že během této doby se v něm objevily více než dva tisíce nových druhů, nezahrnutých v původním HTS, ale ty nebyly do jeho druhé fáze zařazeny.

- Jestliže ano, bylo zjišťováno, zda pocházejí ze stejného „bioprojektu“ – jestliže jedna laboratoř provádí jednu „várku“ sekvenací, obvykle je v databázi vedena jako jeden „bioprojekt“. V rámci snahy o získání nezávislých dat byla vždy dávána přednost datovým souborům ze vzájemně různých bioprojektů.
- Pokud nebyly dostupné soubory z jiného bioprojektu, byla ve druhém sledu vybrána data z téhož.
- Vždy byly vybrány maximálně tři nové datové soubory. V mnohých případech ale byly dostupné pouze dva, jeden, případně žádný.

Takto byl vygenerován seznam asi dvou tisíc dalších datových souborů, které byly analyzovány pomocí TRFi/TRMe a přidružených nástrojů stejnými postupy, jaké byly popsány výše.

Jistá modifikace byla provedena pouze v ohledu žádání o výpočetní zdroje – v první vlně HTS byly vždy zpracovávány stejně velké skupiny datových souborů, ve druhé vlně se jejich počet různil, protože spolu byly vždy analyzovány datové soubory jednoho druhu. Proto byly příslušné skripty upraveny tak, aby příkaz `qsub` žádal vždy o tolik procesorů, kolik je aktuálně zpracovávaných datových souborů.

Kapitola 3

Výsledky

V této kapitole bude popsáno, jakých bylo dosaženo výsledků při identifikaci telomerových sekvencí v datech SRA. V souladu se zadáním práce budoou uvedeny jak výsledky HTS, tak i specifičtější výsledky zpracování úžeji vymezené skupiny datových souborů (rostliny čeledi *Solanaceae*).

Na vstupu byl připraven seznam 5152 datových souborů určených ke zpracování. U jisté části však v rámci HTS (zřejmě z technických důvodů) selhal přístup do databáze[69], ty tak nebyly zpracovány a na výstup se tak dostalo 4969 datových souborů. Provedení samotného screeningu trvalo přibližně tři měsíce[70].

3.1 Statistická analýza dat

První vlna HTS vyprodukovala asi pět tisíc datových souborů, z nich 1296 představovalo negativní výsledky, ve kterých žádná známá telomerová sekvence nebyla identifikována v definovaném množství (viz 2.3.4).

Použitá data však nebyla zcela srovnatelná – byly mezi nimi zastoupeny různé sekvenační platformy, různé konkrétní sekvenátory, jednotlivá čtení měla různou délku.

Na základě toho vyvstává otázka, zda pozitivitu či negativitu výsledků nelze asociovat s takovými technickými parametry sběru dat. Cílem nadcházející analýzy je tedy:

1. Zjistit, zda existuje asociace mezi pozitivitou výsledků HTS[71] a příslušnou sekvenační platformou.

2. Zjistit, zda existuje asociace mezi pozitivitou výsledků HTS a příslušným sekvenátorem (v rámci platformy Illumina).

3. Zjistit, zda existuje asociace mezi pozitivitou výsledků HTS a délkou sekvenčních čtení v příslušném datovém souboru.

[69] Takové chyby se objevovaly průběžně během screeningu, nešlo o jedno vymezené období.

[70] V závislosti na intenzitě využívání dostupných výpočetních zdrojů lze celý proces provést i během několika týdnů.

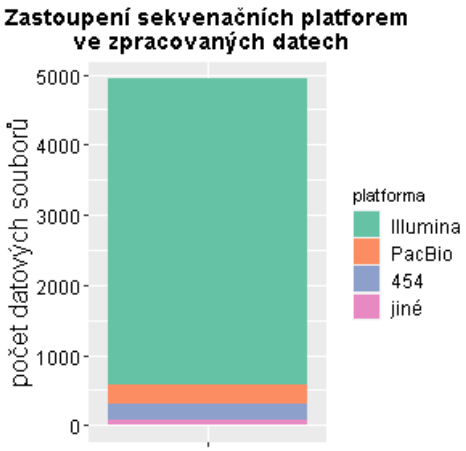
[71] Pozitivitou se myslí, že byla identifikována známá telomerová sekvence. Negativitou se myslí absence známé telomerové sekvence.

Na základě velké datové tabulky „SraRunInfo“ byl vytvořen užší datový soubor, jehož řádky představují zpracované datové soubory. Původní tabulka obsahovala desítky různých proměnných, mnohé z nich však jsou nedostatečně vyplněny či ani nejsou pro tento projekt relevantní. Proto bylo vybráno jen několik nejdůležitějších proměnných (přístupový kód, druh, číslo taxonu, kód bioprojektu, sekvenační platforma, sekvenátor, průměrná délka čtení a binární indikátor pozitivity výsledku).

Ukázka dat:

	acc	druh	TaxID	bioprojekt	platforma	metoda	delka_readu	nalezeno
1	SRR8039422	Panicum hallii	206008	PRJNA495658	ILLUMINA	Illumina HiSeq 2500	302	1
2	SRR8049196	Canis anthus	1707807	PRJNA494815	ILLUMINA	Illumina HiSeq 2500	202	1
3	SRR8049198	Lycaon pictus	9622	PRJNA494815	ILLUMINA	Illumina HiSeq 2500	101	1
4	SRR8064813	Anisolabis maritima	62749	PRJNA497080	ILLUMINA	HiSeq X Ten	300	0
5	SRR8064814	Ara ararauna	9226	PRJNA497082	ILLUMINA	HiSeq X Ten	300	1
6	SRR8064809	Ara glaucogularis	303304	PRJNA497077	ILLUMINA	HiSeq X Ten	300	1
7	SRR8049186	Canis latrans	9614	PRJNA494815	ILLUMINA	Illumina HiSeq 2500	200	1
8	SRR8049189	Cuon alpinus	68730	PRJNA494815	ILLUMINA	Illumina HiSeq 2500	200	1
9	SRR8066610	Canis lycaon	228401	PRJNA496590	ILLUMINA	Illumina HiSeq 2500	200	1
10	SRR8049190	Canis simensis	32534	PRJNA494815	ILLUMINA	Illumina HiSeq 2500	200	1

Mezi zpracovanými daty výrazně převažovala platforma Illumina (4347 souborů), dále PacBio (287 souborů), 454 (234 souborů), na ostatní platformy připadá zbylých 68 souborů. Toto zastoupení je zobrazeno na obrázku 3.1.



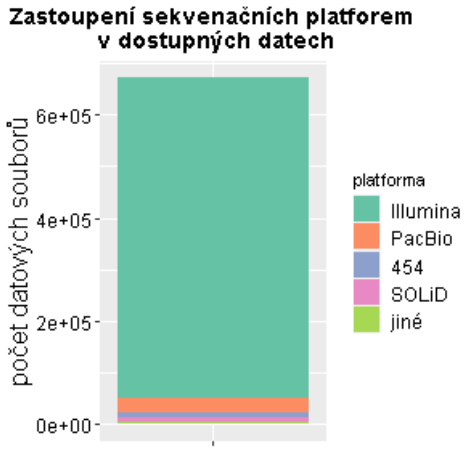
Obrázek 3.1: Počty zpracovaných souborů dle jednotlivých sekvenačních platforem.

Pro srovnání je uvedeno i zastoupení platforem mezi všemi dostupnými daty[72]. Podle datového souboru získaného 24. února 2020 je dostupných 619701 souborů Illumina, 27682 typu PacBio, 11514 typu 454, 7867 typu SOLiD, zbylých je pak 3934. To je zobrazeno na obrázku 3.2.

[72]Rozumí se mezi všemi filtrovanými podle 2.3.1.

Tabulka 3.1: Průměry a mediány délek sekvenčních čtení u nejčastěji zastoupených platforem v rámci zpracovaných dat.

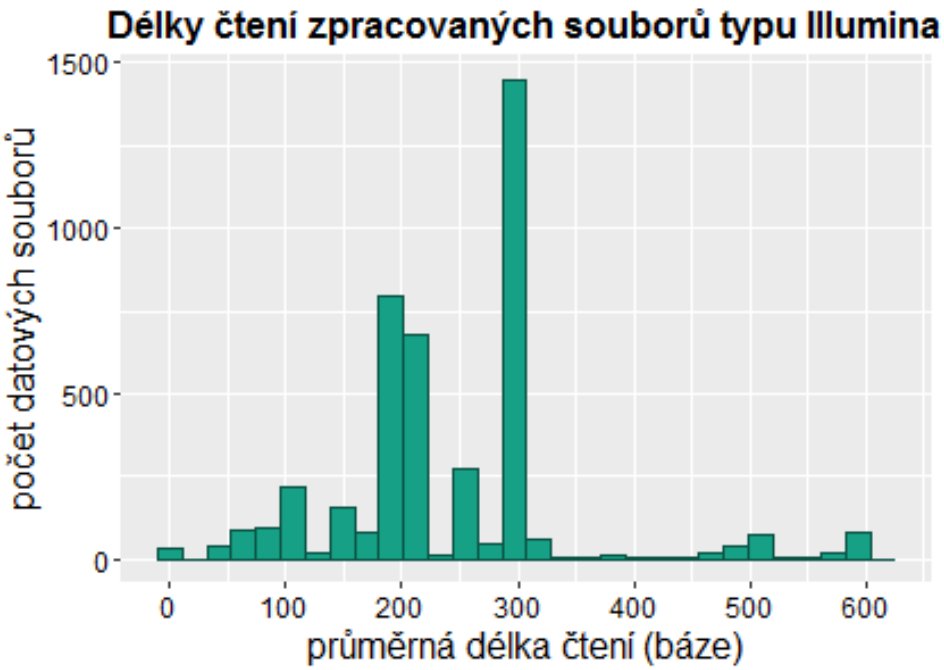
platforma	průměr [bp]	medián [bp]
Illumina	255	208
Pac Bio SMRT	5647	4565
454	542	531
Ion Torrent	209	208



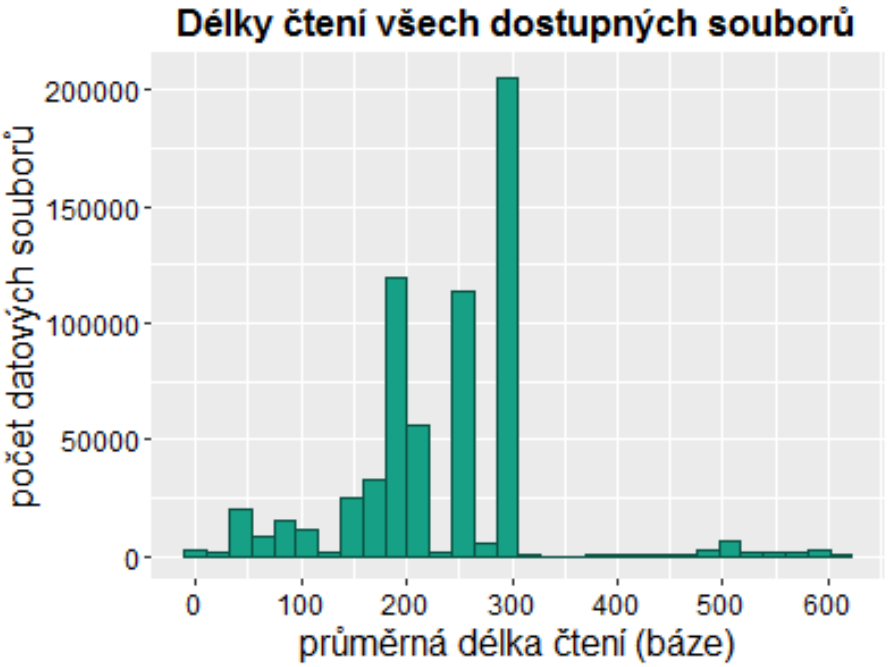
Obrázek 3.2: Počty dostupných souborů dle jednotlivých sekvenačních platforem.

V rámci zpracovaných dat se jednotlivé platformy značně liší vzhledem k průměrné délce sekvenčních čtení. Vzhledem k použitým technologiím vykazují tyto délky polymodální rozložení s maximy okolo hodnot 200 bp, 250 bp a 300 bp. Některé hodnoty (zvláště z TGS) však dosahují i stovek tisíc. Jednotlivé platformy se přitom značně liší v průměrných délkách sekvenčních čtení. Tyto hodnoty jsou shrnuty v tabulce 3.1.

Na obrázku 3.3 je znázorněn histogram délek čtení z platformy Illumina v rámci zpracovaných dat a na obrázku 3.4 je uveden histogram délek sekvenčních čtení u všech dostupných souborů všech platforem. Oba histogramy jsou omezené na rozsah 0 až 600, vyšších hodnot je zanedbatelné minimum.



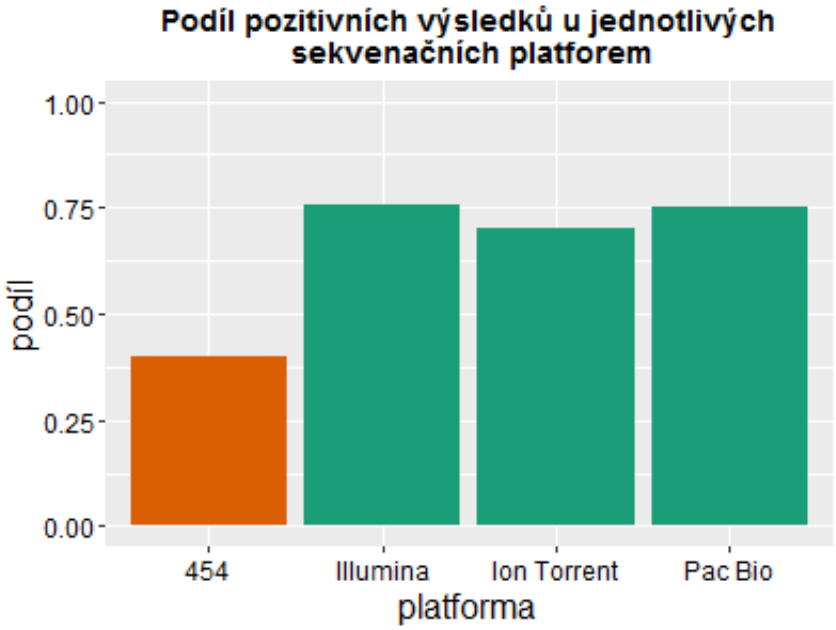
Obrázek 3.3: Histogram délek sekvenčních čtení u zpracovaných souborů z platformy Illumina.



Obrázek 3.4: Histogram délek sekvenčních čtení u dostupných datových souborů všech platforem.

V celkové bilanci představují pozitivní výsledky asi 74 % všech zpracovaných souborů. Byly však identifikovány rozdíly v relativním zastoupení pozitivních výsledků

mezi sekvenačními platformami. Nejnižší míra pozitivních výsledků byla zaznamenána u platformy 454. Viz obrázek 3.5.



Obrázek 3.5: Podíl pozitivních výsledků (souborů se známou telomerovou sekvencí) podle jednotlivých sekvenačních platforem.

Odchylka od očekávaných hodnot četností je patrná již z grafu, byla však i formálně ověřena testem χ^2 na základě kontingenční tabulky 3.2.

Tabulka 3.2: Pozorované četnosti pozitivních a negativních výsledků u jednotlivých platforem.

platforma	pozitivní	negativní
Illumina	3289	1058
Pac Bio SMRT	216	71
454	93	141
Ion Torrent	35	15

Test χ^2 je možné použít, jelikož všechny očekávané četnosti jsou dostatečně vysoké. Předpokladáme rovněž nezávislost jednotlivých pozorování. Nulová hypotéza byla zamítnuta ($p < 0{,}001$), rovněž Fisherův exaktní test vedl k $p < 0{,}001$.

Dále je třeba zkoumat otázku, zda existuje vztah mezi pozitivitou výsledků a průměrnou délkou sekvenčních čtení v daném datovém souboru. Medián délek čtení u pozitivních souborů je 247,5 bp, u negativních 300 bp, což naznačuje značný rozdíl.

Tabulka 3.3: Mediány délek sekvenčních čtení u pozitivních a negativních souborů dle jednotlivých sekvenačních platforem, výsledky příslušných Mannových–Whitneyho testů. Pro rychlé porovnání jsou uvedeny i počty souborů.

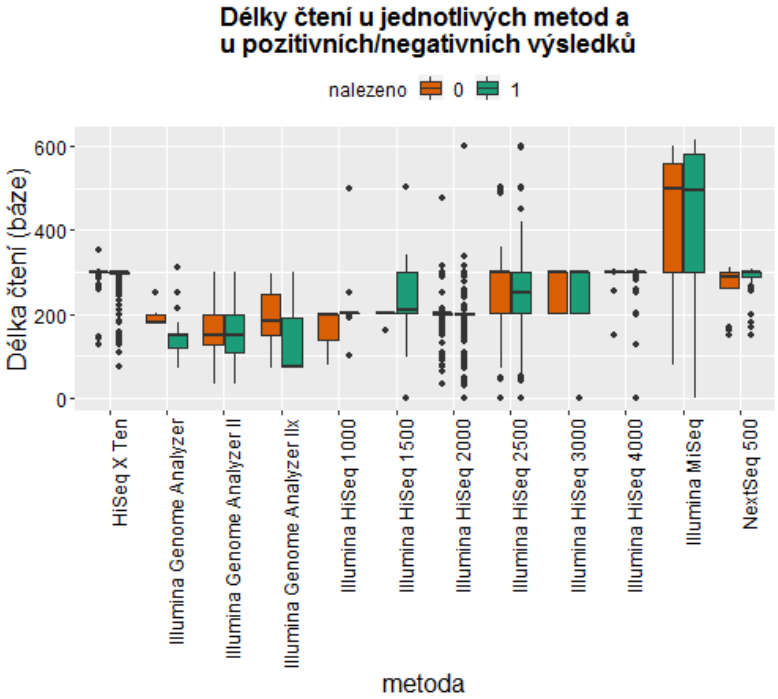
platforma	pozitivní [bp]	negativní [bp]	p	n
Illumina	202	298	$< 0{,}001$	4347
Pac Bio SMRT	4381	6138	0,095	287
454	509	542	0,054	234
Ion Torrent	191	233	0,498	50

Protože příslušná kvantitativní proměnná evidentně nesplňuje předpoklad normality (Shapirův–Wilkův test: $p < 0{,}001$), byly rozdíly testovány pomocí neparametrického Mannova–Whitneyho testu.[73]

Tento rozdíl však ještě nemusí znamenat, že určující proměnnou jsou délky čtení, je třeba rozlišit jednotlivé sekvenační platformy. Mediány dle jednotlivých platforem jsou shrnuty v tabulce 3.3.

Shoda mediánů délek sekvenčních čtení u pozitivních a negativních datových souborů byla zamítnuta jen u platformy Illumina. Budiž podotknuto, že to zřejmě souvisí s nerovným počtem pozorování – souborů typu Illumina je v datech nejvíce.

Dále proto byly filtrovány řádky platformy Illumina a bylo zjišťováno, zda nalezené rozdíly nelze vysvětlit růzností konkrétních sekvenátorů. To lze vizuálně zhodnotit pomocí krabicových grafů – viz obrázek 3.6.

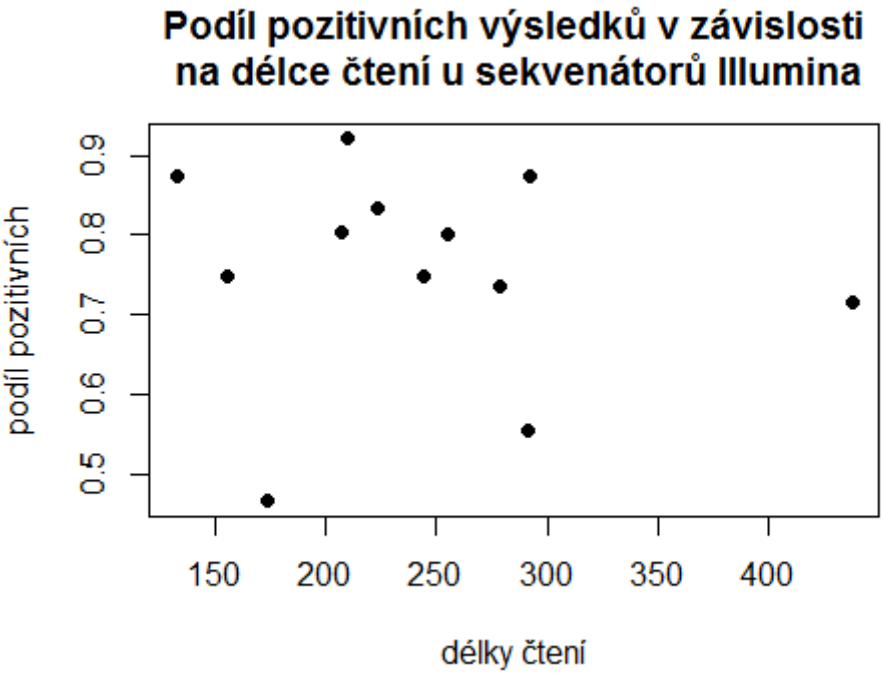


Obrázek 3.6: Délky sekvenčních čtení u pozitivních a negativních souborů – rozdělení dle různých sekvenačních metod platformy Illumina.

[73] Testována je tedy shoda mediánů, nikoli shoda středních hodnot.

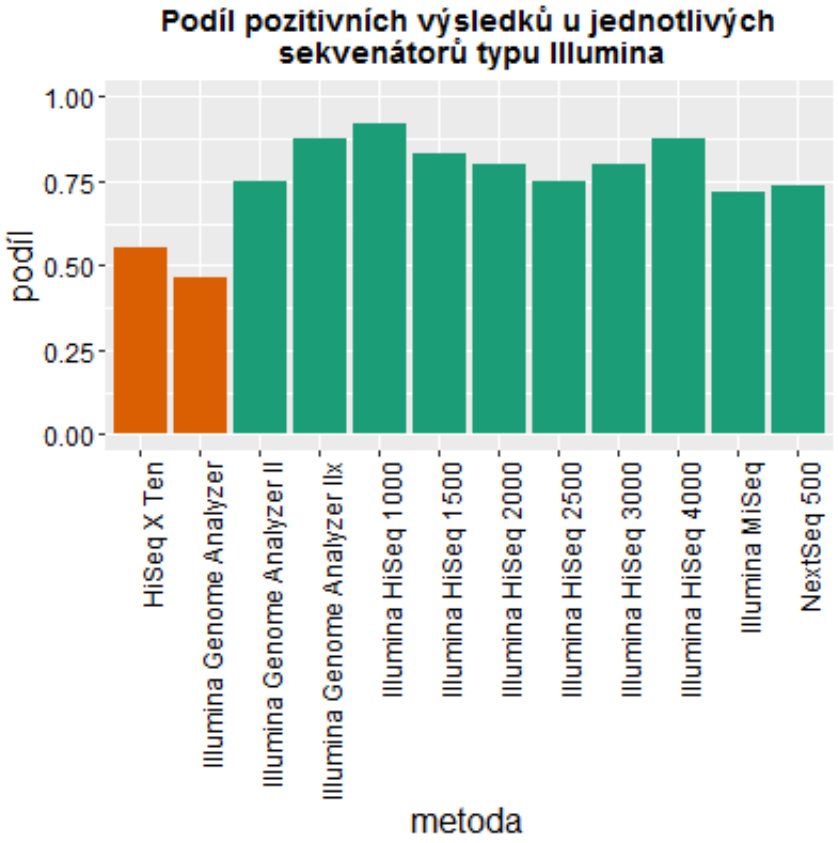
Lze si všimnout až překvapivé shody mediánů u mnohých sekvenátorů (např. HiSeq X Ten, Genome Analyzer II, HiSeq 1000, 1500, 2000, 3000, 4000, MiSeq, NextSeq 500), zatímco v jiných případech rozdíly přetrvávají (Genome Analyzer, Genome Analyzer IIx, HiSeq 2500). Nyní je však již patrné, že pokles podílu pozitivity s nárůstem délky čtení není obecnou vlastností všech datových souborů, ale lze jej připsat konkrétnímu experimentálnímu postupu. V případě zmíněných tří „opticky" nejodlišnějších dvojic byl Mannovými–Whitneyho testy potvrzen rozdíl mediánů (vždy $p < 0{,}001$).

Lze dále uvážit podíl pozitivních výsledků v závislosti na délce sekvenčních čtení (viz obrázek 3.7), z obrázku není patrná žádná závislost, také příslušný lineární model (ve smyslu proložení přímkou) vychází nevýznamný ($p = 0{,}596$).



Obrázek 3.7: Graf zobrazující závislost podílu pozitivních výsledků na průměrné délce sekvenčních čtení u daného sekvenátoru.

Již z uvedeného bodového grafu lze usoudit, že všechny sekvenační metody nevykazují stejný podíl pozitivních výsledků, ještě lépe je to vidět z následujícího grafu 3.8. Z něj je patrné, že nejhorších výsledků je dosahováno u dat pocházejících z Hiseq X Ten a Genome Analyzer, naopak nejlepší výsledky dávají HiSeq 1000, Genome Analyzer IIx či HiSeq 4000. Rozdíl byl potvrzen i Fisherovým testem na základě příslušné kontingenční tabulky ($p < 0{,}001$). Budiž podotknuto, že většina datových souborů pochází ze sekvenátorů HiSeq 2000, resp. HiSeq 2500, u kterých byl podíl pozitivních výsledků 0,80, resp. 0,75.



Obrázek 3.8: Graf zobrazující závislost podílu pozitivních výsledků na konkrétním sekvenátoru typu Illumina.

Z analýzy dat vyplývají tyto závěry:

1. Existuje souvislost mezi pozitivitou výsledků HTS a použitou sekvenační platformou – v případě platformy 454 je míra pozitivních výsledků významně nižší, než jak je tomu u ostatních platforem.

2. V rámci platformy Illumina existují významné rozdíly v podílu pozitivních výsledků mezi jednotlivými sekvenačními metodami.

3. Nebyla zjištěna obecná asociace mezi délkou sekvenčních čtení a pozitivitou výsledků, z vizuální inspekce se zdá, že k tomu dochází jedině v interakci s jiným faktorem, a sice s konkrétní sekvenační metodou. Pokud však již taková interakce nastává, pak se pozitivní výsledky vyskytují spíše u kratších čtení. Korelace mezi délkou čtení a podílem pozitivních souborů nebyla zjištěna ani u různých sekvenátorů platformy Illumina.

3.2 Výsledky HTS

Výsledkem prvního HTS bylo zpracování asi pěti tisíc genomů, identifikace známých telomerových sekvencí (dle 2.3.4) a vytvoření tabulky, v nichž každá dvojice sloupců

odpovídá jednomu genomu (četnosti a sekvence) uspořádané pro každý genom sestupně dle četností. Genomy jsou v tabulce uspořádány dle taxonomie (popsáno v 2.3.6). V této kapitole ovšem nebude jmenovitě uváděn každý zpracovaný druh. Skupiny, v nichž převažují očekávané telomerové motivy, budou popsány pouze souhrnně s uvedením případných výjimek. Pouze taxony s velkou variabilitou výsledků budou popsány podrobněji.

3.2.1 Jednobuněčné eukaryontní organismy

Velká variabilita telomerových motivů byla pozorována u druhu *Plasmodium*. Screening pokryl kolem dvaceti druhů plasmodií (vč. *P. malariae* – `TTAGGG`). U dvanácti z nich byla identifikována „rostlinná“ telomerová sekvence. U šesti druhů byla kromě „rostlinné“ telomerové sekvence nalezena ve vysoké míře i sekvence `TTCAGGG`, u dvou druhů (*P. gallinaceum* a *P. yoelii*) pak pouze tato modifikovaná sekvence. U čtyř druhů byla identifikována kanonická telomerová sekvence, z toho u druhu *P. simiovale* „rostlinná“ spolu s kanonickou.

U druhů *P. gaboni* a *P. reichenowi* nebyla identifikována známá telomerová repetice, u obou se však v hojné míře nachází kandidátní sekvence `AATGG`.

Byly zpracovány tři genomy klíštěnek (*Babesia* – parazité, původci babesióz u zvířat). U dvou (*B. bovis* a *B. divergens* byla nalezena „rostlinná“ telomerová sekvence). U *B. bigemina* nebyla nalezena „rostlinná“ telomerová sekvence, pouze v menší míře `TTTTAGGG`, resp. variabilní repetice délky kolem 12 bp s rovnoměrným zastoupením AT i GC. Stojí za povšimnutí, že ze čtyř zpracovaných genomů *B. divergens* byla pouze v jednom identifikována ve významnějším množství „rostlinná“ telomerová sekvence, zato ve všech čtyřech byla výrazně zastoupena sekvence `TTAAGGATGT` (ve dvou případech druhá nejčetnější).

Byly zpracovány tři druhy rodu *Theileria* (*T. annulata*, *T. orientalis* a *T. parva*). U posledních dvou byla nalezena „rostlinná“ telomerová sekvence, resp. `TTTTAGGG`, pouze u *T. annulata* nebyla v žádném ze čtyř zpracovaných genomů nalezena známá ani kandidátní sekvence.

Různorodé výsledky byly nalezeny v případě rodu *Cryptosporidium*, tyto datové soubory vykazují velkou variabilitu telomerových sekvencí. Bylo zpracováno devět různých druhů tohoto rodu. U čtyř z nich byla nalezena „rostlinná“ telomerová sekvence, ale mnohdy v kombinaci s alternativními motivy `TTTTAGGG`, `TTTAGG`, `TTTGGGG` nebo `TTTTAGG`.

Byly zpracovány druh *Cyclospora cayetanensis* a sedm druhů *Eimeria*, u všech významně zastoupena „rostlinná“ telomerová sekvence. U *E. tenella* v menší míře `TTTGGGG` a u *E. maxima* v menší míře `TTAGGG`.

Byly zpracovány čtyři druhy rodu *Tetrahymena*. U *T. thermophila* a *T. elliotti* byla nalezena pouze `TTGGGG` (v souladu s očekáváním pro tento rod). Ale u *T. borealis* byly nalezeny ve zhruba stejné (a velké) míře motivy `TTGGGG` a `TGGGGG`. Dále u *T. malaccensis* „obratlovčí“ sekvence `TTAGGG`.

U ostatních druhů taxonu Alveolata převažuje „rostlinná“ telomerová sekvence, ta byla nalezena u *Hammondia hammondi*, *Neospora canium*, *Sarcocystis neurona*,

Gregarina niphadrodes, *Besnoitia besnoiti*, *Vitrella brassicaformis*. U *Paramecium tetraurelia* a *Pseudocohnilembus persalinus* byla nalezena sekvence `TTGGGG`.

Sekvence `TTTTGGGG` byla nalezena u *Oxytricha trifallax* a *Sterkiella histriomuscorum*, podobný motiv `TTTGGGG` pak u *Chilodonella uncinata*.

Žádné telomerové sekvence nebyly nalezeny u *Colpodella sp.* a *Azadinium spinosum*, ve druhém případě však zřejmě nešlo o WGS (chyba anotace).

U druhu *Symbiodinium microadriaticum* byla nalezena „rostlinná" sekvence `TTTAGGG`, ovšem v kombinaci `TTTAGGGG` a `TTTTAGGG`. U *Amoebophrya sp.* byly identifikovány pouze dlouhé variabilní motivy jako `CCTTGGGGCCACCTTGGGCCCA`, resp. `TTAGGG` v malé míře.

Žádná (ani kandidátní) sekvence nebyla nalezena u *Chromera velia*, nejčetnější jsou různé tetranukleotidové mikrosatelity typu `TCTG`. Podobně je na tom *Ichthyophthirius multifiliis*, mezi jeho repeticemi převažují AT-bohaté motivy. U všech čtyř datových souborů však byla v menší míře přítomna sekvence `TTGGGG`.

Bylo zpracováno 13 druhů taxonu Amoebozoa. Převažuje kanonická telomerová sekvence `TTAGGG`. Telomerová sekvence nebyla identifikována u *Entamoeba nuttalli*, *Dictyostelium citrinum*, *Dictyostelium firmibasis* a *Dictyostelium intermedium*. Ani u druhu *Polysphondylium violaceum* nebyla ve větší míře nalezen známý telomerový motiv, pouze v menší míře `TTAGGC`.

Vzorek řas popsaný jako *Cryptophyceae sp.* obsahuje převážně tyto motivy: `TTTAGGG`, `TTTGGGG`, `TGTAGGG`, `TTGAGGG`, `TTTCGGG`, `TTCAGGG` a `TTTATGG`. To naznačuje nečekaně vysokou rozmanitost telomerových motivů v rámci jednoho úzkého taxonu.

Rod *Diplonema* je dle výsledků bez standardních telomerových sekvencí. Byly nalezeny motivy připomínající telomerové sekvence `TTTTTGACCCC` a podobné.

U ostatních druhů taxonu Excavata převažují standardní motivy (`TTAGGG`) včetně *Euglena*, *Leishmania* nebo *Trypanosoma*. Občas se „paralelně" vyskytují sekvence připomínající ty telomerové (`TTCCGC` na 1. místě u *Trypanosoma borreli*, u *Leishmania infantum* na 1. místě `ATTTTTTCGGG`). *Leptomonas seymouri* je dle výsledků bez známé telomerové sekvence. Některé druhy *Trypanosoma* či *Giardia* mají vysoké abundance dlouhých motivů (i před 20 bp) připomínajících „zfúzované" a přitom pozměněné telomerové repetice, např. `TGAACCCTCATTACCTTTACCCCACCA`.

Značná variabilita telomerových motivů byla zaznamenána i v případě taxonu Haptophyta. Převažují kanonické telomerové sekvence, ovšem u rodu *Phaecystis* se vyskytují paralelně s alternativními motivy jako `TTGGGG`, `TTAGGC`, `TTCGGG` a jinými.

3.2.2 Ascomycota

Bylo zpracováno několik set druhů Ascomycota. Žádné známé ani vhodné kandidátní sekvence nebyly nalezeny např. u druhů *Salpingoeca rosetta*, *Leptoxyphium fumago*, *Polychaeton citri*, *Cladosporium herbarum*, *Mycocentrospora sp.* či *Myriangium duriaei*.

U některých druhů se kanonická telomerová sekvence vyskytuje spolu s případnými kandidátními či známými alternativními motivy, např. *Lophium mytilinum*, *Amniculicola lignicola*, *Byssothecium circinans* či *Beverwykella pulmonaria*.

U některých druhů byla identifikována zajímavá kandidátní sekvence `TGGTCG` – např. *Leptosphaeria sp.*, *Alternaria sp.*, *Capronia sp.*, *Stachybotrys sp.*, *Xylaria sp.* V některých případech (např. *Alternaria alternantherae* nebo *Xylaria hypoxylon*) byla přítomna spolu s kanonickou sekvencí.

Kanonické sekvence nebyly nalezeny u mnohých druhů rodu *Aspergillus*: např. *A. campestris*, *A. carbonarius*, *A. awamori*, *A. lacticoffeatus* či *A. costaricaensis*, u posledních tří jmenovaných (a kolem deseti dalších druhů) navíc byla nalezena možná telomerová sekvence `TTTTAGGG`. Stejná nalezena i u dvou analyzovaných druhů rodu *Monascus* nebo u *Thermoascus crustaceus*.

V rámci rodu *Penicilium* byly identifikovány jak kanonické (např. *Penicillium janthinellum*, *Penicillium oxalicum*), tak „rostlinné“ telomerové motivy (např. *Penicillium lanosocoeruleum*, *Penicillium swiecickii*). V případě druhu *Penicilliopsis zonata* byly přítomny oba zmíněné motivy současně.

U druhů *Raffaelea quercivora*, *Diaporthe ampelina* nebo *Grosmannia penicillata* byla nalezena možná telomerová sekvence `TTAGG`, v prvních dvou případech spolu s kanonickou sekvencí, ve třetím případě spolu s kandidátní sekvencí `TGGTCG`. Stejná sekvence (`TTAGG`) pak byla nalezena např. i u druhů *Tuber melanosporum* či *Alloascoidea sp.*

3.2.3 Basidiomycota

Kanonická sekvence `TTAGGG` se opět ukazuje být převažující. Znovu však lze nalézt také značné množství různých alternativ.

U rodu *Amanita* (vč. *A. phalloides*) byla identifikována „rostlinná“ sekvence `TTTAGGG`, kromě *A. muscaria* se sekvencí `TTTTAGGG` – tato sekvence se pak vyskytuje i rodu *Hygrophorus*. U rodu *Termitomyces* byla identifikována sekvence `TAAGGG` ve vyšší míře než kanonická sekvence. U části rodu *Pleurotus* (*P. cystidiosus*, *P. dryinus*, *populinus*, *P. pulmonarius*) nebyla nalezena kanonická sekvence, ale alternativní `TGGTCG`.

U některých druhů (např. *Infundibulicybe gibba*, *Pholiota highlandensis*) byly kanonická a „rostlinná“ sekvence zaznamenány v takřka identickém množství. Podobně u *Elmerina caryae*, *Veluticeps abietina*, *Heliocybe sulcata* ve srovnatelném množství kanonická sekvence a `TTAGG`.

Žádná známá ani vhodná kandidátní sekvence nebyla nalezena u *Anomoloma myceliosum* (v žádném ze 4 zpracovaných datových souborů).

U rodu *Boletus* jsou přítomny kromě kanonických motivů i `TTAGG`, u druhů *B. bicolor* a *B. speciosus* v menší míře i `TGGTCG`.

V případě rodu *Paxillus* (čechratka) v datech převažují „rostlinné“ sekvence, ačkoli i kanonické jsou v menší míře přítomny. U rodů *Pisolithus* a *Rhizopogon* jsou zastoupeny `TTAGG`.

Žádné známé ani kandidátní sekvence nebyly nalezeny mezi dominantními u rodu *Suillus* (klouzek), kanonický či kandidátní motiv byl vždy málo četný, případně chybějící.

U hadovkotvarých hub *Mutinus elegans* a *Phallus impudicus* a u *Sclerogaster hysterangioides* v datech převažuje „rostlinná“ sekvence.

U čtyř z pěti zpracovaných druhů rodu *Auricularia* nad všemi ostatními sekvencemi výrazně dominuje `TTAGG`, kanonická sekvence není přítomna, resp. jen v malém množství.

Neobvyklé dominantní motivy byly identifikovány u *Clavulina sp.* (`AATGG`) a *Tulasnella calospora* (`TAGGGC`, hned „za ní“ `TTAGG`).

Motiv `TTAGG` převažuje nad ostatními kandidátními také u *Postia placenta*, *Adustoporia sinuosa*, *Antrodia serialis*, *Fomitopsis pinicola*, *Piptoporus betulinus*, *Hydnopolyporus fimbriatus*, *Irpex lacteus*, *Trametopsis cervina*, *Laetiporus sulphureus*, *Taiwanofungus camphoratus* a *Thelephora sp.*

„Rostlinná“ sekvence v datech převažuje u chorošotvarých *Phanerochaete carnosa*, *Phanerochaete chrysosporium* a *Phlebiopsis gigantea*. Rovněž tak u *Lentinellus vulpinus*, *Artomyces pyxidatus*, *Auriscalpium vulgare*, *Vararia minispora* a *Lactarius sp.* (ryzec), u něj je navíc v menších zastoupeních přítomna `TGGTCG`, nebo také u *Russula sp.* (holubinka).

U *Lentinus tigrinus* mnohonásobně převažuje motiv `TATAAGGG`, kanonická sekvence zastoupena v menší míře. Kandidátní sekvence nebyly nalezeny u *Udeniomyces sp.*

Zajímavý je rod *Filobasidium*. U čtyř zpracovaných souborů *F. uniguttulatum* byl přítomen dominantní motiv `TTTGGGG`. U *F. elegans* pak byl u dvou zpracovaných souborů výrazně dominantní motiv `TTTACTCCTCGCTAGGGACTTGTGCG`, ale rovněž byl u obou hojně zastoupen motiv `TTAGGGG`, ten byl nalezen i u *F. wieringae*, resp. u druhů *Naganishia albida* či *Naganishia vishniacii*, u rodu *Naganishia* jinak mezi dominantními nebyly přítomny žádné známé motivy.

U *Solicoccozyma aeria* byla nalezena kandidátní sekvence `TTTTGAG` (ve dvou souborech), u *Solicoccozyma phenolicus* pak `TTAGGGGG`. Naopak u *Solicoccozyma terricola* byla zjištěna kanonická telomerová sekvence.

U druhů *Dioszegia crocea* a *Dioszegia cryoxerica* v datech rovněž dominovala sekvence `TTAGGGGG`, stejně tak u *Cryptococcus neoformans*. Také u řady jiných druhů rodu *Cryptococcus* kanonická telomerová sekvence chyběla. U *Cryptococcus depauperatus* ve čtyřech datových souborech převažovala sekvence `TTTTCTAGG`. Stejná sekvence pak byla identifikována i u druhu *Kwoniella europaea*.

U druhů *Phaeotremella skinneri*, *Phaeotremella fagi*, *Saitozyma podzolica*, *Takashimella koratensis*, *Cutaneotrichosporon curvatum*, *Trichosporon asteroides*, *Trichosporon faecale* nebo *Papiliotrema laurentii* v datech dominoval motiv `TTAGGGGG`.

V případě rodu *Apiotrichum* nedominovaly žádné známé motivy, pouze u *Apiotrichum loubieri* kandidátní sekvence `TTGGTG` a u *Apiotrichum veenhuisii* `TTAGGGGG`.

U druhu *Trichosporon asahii* ve všech třech zpracovaných souborech dominoval motiv `TGGACGGGG`. U druhu *Cystobasidium* byl nalezen dominantní motiv `TTTTAGGG`. U druhů *Erythrobasidium hasegawianum* a *Erythrobasidium yunnanense* vždy výrazně dominoval motiv `TTTACAGAGGG`.

Dalším zajímavým rodem je *Malassezia*. Např. u čtyř souborů *M. furfur* nebo u tří souborů *M. japonica* je vždy dominantní `ATTAGG`, podobně u *M. yamatoensis*. S výjimkou *Malassezia vespertilionis*, *Malassezia slooffiae* a *Malassezia cuniculi* pak kanonická sekvence v datech nikdy není mezi dominantními.

3.2.4 Ostatní houby a organismy jim podobné

Skladba nalezených (kandidátních) telomerových sekvencí odpovídá předchozím popsaným skupinám – kanonická sekvence je nejčetnější, ale vyskytují se i mnohé odchylky, případně u mnohých organismů žádné vhodné kandidátní sekvence nebyly nalezeny.

U druhů *Catenaria anguillulae* a *Synchytrium endobioticum* byla zjištěna dominující sekvence `TTTAGG`. U *Spizellomyces palustris* se tato sekvence vyskytuje v množství podobném kanonické sekvenci.

U rodu *Nosema* (hmyzomorka) převažuje v datech motiv `TTAGG`. Podobně je tomu u parazita *Encephalitozoon cuniculi*. Je třeba uvážit, že může jít o kontaminaci.

V případě druhu *Gamsiella multidivaricata* dominují dlouhé sekvenční motivy `AATTCCCTAAACAT` a `GTTTAGGGGATTTC` připomínající „zfúzované“ telomerové motivy.

Žádné kandidátní sekvence nebyly nalezeny u rodů *Haplospora* a *Cunninghamella* (jen u *Cunninghamella elegans* v menší míře „rostlinná sekvence“).

„Rostlinná“ sekvence byla nalezena u rodů *Hesseltinella*, *Circinella*, *Fennellomyces*, *Lichtheimia*, *Phascolomyces*, *Thamnostylum*, *Apophysomyces* nebo *Umbelopsis*. Naopak u rodů *Mucor*, *Actinomucor*, *Pirella*, *Conidiobolus*, *Entomophaga*, *Zoophthora* či *Pilobolus* převažuje sekvence kanonická.

Bez plauzibilních kandidátních sekvencí zůstává rod *Rhizopus* (výjimkou je `TTGTGG` u *Rhizopus stolonifer*). Táž sekvence je přítomna i u *Sporodiniella umbellata*.

U rodu *Mortierella* byla identifikována kandidátní sekvence `TTTAGGGATTT` a u druhu *Kickxella alabastrina* sekvence `TTTTGGGG`.

3.2.5 Ostatní Metazoa

U *Salpa thompsoni* byla přítomna kanonická sekvence jako nejhojnější, ovšem v množství srovnatelném s `TTAGG`.

Kanonickou sekvenci se nepodařilo identifikovat u tasemnice *Taenia solium* a u *Schistocephalus solidus*, u něj byl hojnější spíše motiv `TGGAA`, dále pak u *Schistosoma intercalatum* či *Schistosoma mansoni*. V případě druhu *Tethya wilhelma* byla nalezena kandidátní sekvence `TAGGGC`.

3.2.6 Nematoda

U většiny rodů, např. *Plectus*, *Pristionchus*, *Caenorhabditis*, *Ancylostoma*, *Heterorhabditis*, *Cylicostephanus*, *Strongylus*, *Anisakis*, *Parascaris*, *Anguina*, *Toxocara* a dalších, byl identifikován telomerový motiv `TTAGGC`. Ten, zdá se, u taxonu Nematoda obecně převládá.

Žádné kandidátní sekvence nebyly nalezeny u dvou druhů rodu *Dranunculus*. Může jít o chybu anotace.

Obvyklý motiv `TTAGGC` nebyl nalezen u rodů *Rhabditophanes*, *Parastrongyloides*, *Strongyloides* či *Aphelenchus*. Mezi jejich telomerovými repeticemi dominují relativně dlouhé, variabilní motivy jako např. `ACTTCCTAACCGGTTAACC` u *Aphelenchus avenae*.

Zvláštní výsledky poskytl rod *Heterodera*, v datech dominoval „rostlinný“ motiv `TTTAGGG`.

Plauzibilní kandidátní sekvence nebyly nalezeny u druhů *Haemonchus placei*, *Teladorsagia circumcincta*, *Nippostrongylus brasiliensis* nebo *Trichinella sp.*

3.2.7 Panarthropoda

V rámci členovců je převládajícím telomerovým motivem `TTAGG`.

Nečekané výsledky poskytuje rod *Ixodes* (klíště), u druhů *I. ricinus*, *I. ovatus* a *I. persulacatus* nebyla sekvence `TTAGG` identifikována mezi dominantními. Naopak u *I. scapularis* byla nalezena v relativně vysokém množství. Možnou kandidátní sekvencí u „negativních druhů“ je `TAAGC`.

U některých druhů bez nalezeného motivu `TTAGG` byla identifikována kandidátní sekvence `TGGTCG`, např. *Cheiroseius sp.*, *Penaeus stylirostris*, *Epinotia vertumnana*, *Helicoverpa armigera*, *Ostrinia nubilalis*, *Carpelimus sp.*, *Phlebotomus papatasi* a *Podisus maculiventris*.

Standardní motiv nebyl nalezen u mnohých pavouků, např. *Araneus bicentenarius*, *Nephila clavipes*, *Loxosceles reclusa* a dalších. U těchto druhů dominují dlouhé variabilní motivy jako `TTTTTCCCCAACT`, `ATGAAAATGG` apod.

U druhu *Calanus glacialis* byla identifikována kandidátní sekvence `TTTAGGGG`. U druhů *Lepeophtheirus salmonis* a *Pandarus rhincodonicus* byla identifikována kanonická sekvence `TTAGGG`.

V případě druhu *Chrysodeixis includens* nebyla nalezena sekvence `TTAGG`, ale byl v menší míře nalezen motiv `TTGGG`, jako nejhojnější pak `AGTCCGAGCCA`.

U rodu *Heliconius* (babočka) je většinou přítomna obvyklá „hmyzí“ sekvence, ovšem často je přítomen i motiv `TTTAGG`, někdy (*H. elevatus*) je dokonce četnější než obvyklá sekvence nebo je zastoupen v podobném množství (*H. pardalinus*). U několika druhů (*H. elevatus*, *H. hecale*, *H. pachinus* aj.) je v menší míře přítomen i motiv `TTAGGC`.

U druhu *Ornithoptera priamus* nebyla nalezena „hmyzí“ sekvence, pouze v menší míře „rostlinná“ sekvence.

U druhu *Lionepha erasa* výrazně a konzistentně[74] dominuje motiv `TGGAA`. Naopak u *L. lindrothellus* byl nalezen standardní „hmyzí“ motiv.

V případě rodu *Dendroctonus* (jde o škůdce stromů) konzistentně dominuje sekvenční motiv `TTAGTTTGGG`.

U druhů z taxonu Tenebrionoidea je je běžná telomerová sekvence `TCAGG`, ne však u všech. Např. u *Gridelliopus subsquamosus* byl zjištěn dominantní motiv `TGGAA`, u *Loensus sp.* a *Melambius mideltensis* byla zjištěna simultánní přítomnost kanonické a „rostlinné“ sekvence.

Bez nalezené kandidátní sekvence zůstávají *Oryctes borbonicus*, *resplendens*, *Popillia japonica*, rod *Polystes* a *Anisolabis maritima*.

U některých druhů čeledi *Scarabaeidae* (*Canthidium sp.*, *Onthophagus taurus*, *Pachysoma striatus*) byla nalezena kandidátní sekvence `TTGGG`.

Vhodné kandidátní sekvence nebyly nalezeny u některých dvoukřídlých (Diptera), vč. *Drosophila melanogaster*, ale u jiných druhů, např. *D. anomalata*, *D. ananassae* nebo *D. parapallidosa*, byla nalezena kandidátní sekvence `TTGACC`, stejná sekvence

[74] Konzistentně ve smyslu potvrzení na nezávislých datových souborech.

byla nalezena např. i u *Gnoriste bilineata*. Mezi Diptera s nalezenou „hmyzí“ sekvencí patří např. *Cirrula hians*, *Themira minor* či *Haematobia irritans*.

Žádná kandidátní sekvence nebyla nalezena u rodu *Bombus* (čmelák). U druhů *Nannotrigona testaceicornis* a *Scaptotrigona postica* byla identifikována kandidátní sekvence `TTAGGTTTGGG`.

Standardní „hmyzí“ sekvence nebyly nazeleny u druhů taxonu Parasitoida. Např. u rodů *Eupelmus* a *Eurytoma* je přítomna „rostlinná“ sekvence, u *Eurytoma* navíc `TTCAGGG`, ovšem dominantní jsou dlouhé motivy jako např. `CCCTAAACCCATAAA` u *Eupelmus urozonus*. U rodu *Nasonia* byla nalezena kandidátní sekvence `TTATTGGG`. U druhu *Megaphragma amalphitanum* byla nalezena kandidátní sekvence `TTATGGGG`, u *Aulacidea tavakolii* pak `AATAGGGG`.

U druhu *Gerris buenoi* (bruslařka) byla identifikována kandidátní sekvence `TTAGAGGTGG`, v případě *Photuris sp.* byla zjištěna kandidátní sekvence `TTGGTAGG`, u rodu *Chinavia* pak sekvence `TTGAG`, u *Dichelops melacanthus* `TTAGGTTTGGG`, u Euschistus heros `TTATTGGG`, u rodu *Columbicola* `TTGGG` s výjimkami *C. claytoni* či *C. hoogstraali* bez kandidátní sekvence a *C. exilicornis* s kandidátní sekvencí `TTTGG`.

3.2.8 Lophotrochozoa a Gnathifera

V případě taxonu Lophotrochozoa převažuje kanonická sekvence `TTAGGG`. U druhu *Eisenia fetida* je kanonická sekvence přítomna v množství srovnatelném s množstvím repetic typu `TTTTTAGG`. U rodu *Crassostrea* je kanonická sekvence přítomna spolu s `TTGACC`, u *Pecten maximus* či *Idiosepius paradoxus* pak spolu s „rostlinnou“ sekvencí, u *Haliotis tuberculata* s „hmyzí“ sekvencí, u *Notospermus geniculatus* se sekvencí `TCAAGG`, u *Brachionus calyciflorus* s `TTGGGG`.

U druhu *Placopecten magellanicus* byla zjištěna kandidátní sekvence `TTTGG`. U *Octopus sp.* pak dominují motivy `TTTAGGC` a `TTTTAGG`. U *Adineta vaga* byla nalezena kandidátní sekvence `TGTGGG`. Bez vhodné kandidátní sekvence zůstává rod *Littorina*.

3.2.9 Cnidaria

U taxonu Cnidaria (žahavci) převažuje kanonická telomerová sekvence. V případě druhu *Phymanthus crucifer* převažuje nad kanonickou sekvencí „rostlinná“ sekvence, která je v menší míře přítomna i u řady jiných druhů, např. *Acropora sp.* či *Stylophora pistillata*.

U některých druhů (např. *Chrysaora quinquecirrha*, *Carybdea marsupialis*) je kromě kanonické sekvence přítomna i „hmyzí“ sekvence. U duhu *Kudoa iwatai* byla nalezena kandidátní sekvence `TTTAGG`.

3.2.10 Vertebrata

Obecně lze říci, že u doposud popsaných skupin je variabilita telomerových motivů vyšší než u obratlovců, kteří vykazují značnou konzervovanost kanonického telomerového motivu.

Jistou pozornost může vzbudit *Xiphophorus sp.* (mečovka). Ze sedmi zpracovaných druhů kanonická sekvence dominovala u čtyř. V případě *Xiphophorus cortezi*

byly dominantní dlouhé, variabilní motivy jako `GGATGCGTGTTTGGGGCGTTAGACGGTT`, které by mohly připomínat „zfúzované“ telomerové motivy. U *Xiphophorus malinche* byla kromě kanonické sekvence identifikována i `TTGGGG` o srovnatelné abundanci. U *Xiphophorus nezahualcoyotl* pak nebyla nalezena žádná kandidátní sekvence, ale negativita nebyla povrzena jinými datovými soubory.

Další anomálie se týká rodů *Cyprichromis* (cichlida) a *Xenotilapia*. Nebyly u nich identifikovány kanonické telomerové sekvence, ale byly nalezeny kandidátní sekvence `TGTGGG`.

Kanonické sekvence nebyly nalezeny u dvou souborů rodů *Ariopsis*, dále u *Aspistor*, *Bagre* a jiných z čeledi *Ariidae*, všechny vzorky ale pocházejí z téhož projektu a jde pravděpodobně o chybu anotace.

Kanonická sekvence dále chybí u *Macroscelides proboscideus* a *Dolichotis patagonum*, jde o týž sekvenační projekt a patrně je to opět chyba anotace.

U *Rattus andamanensis* také chybí kanonická sekvence, ale v anotaci je v poznámce napsáno, že má jít o resekvenaci houbového genomu, tudíž nejde o důvěryhodný vzorek.

Nečekané výsledky byly pozorovány u datových souborů opic (*Gorilla*, *Pan*, *Pongo*). Kanonické sekvence nebyly mezi dominantními, případně zcela chyběly. Výrazně dominantním motivem pak byl často `TGGAA` (např. u *Pan paniscus* u (všech) čtyř zpracovaných souborů). Nelze však z toho vyvozovat radikální závěry, např. u *Rhinopitecus bieti* byla v jednom případě také dominantní `TGGAA`, zatímco u dalšího vzorku téhož druhu byla identifikována kanonická sekvence, podobně tomu bylo u vyhynulého koně *Haringtonhippus francisci*. Na druhou stranu, výsledky screeningu jsou v tomto ohledu (dominance `TGGAA` a nízká přítomnost nebo nepřítomnost `TTAGGG`) poměrně konzistentní, data přitom byla vybírána z různých experimentů.

Mnohé negativní výsledky u různých druhů obratlovců pocházejí z bioprojektu označeného PRJEB20997, jedná se patrně o chybně anotovaný bioprojekt.

Poněkud anomální výsledky byly pozorovány u sudokopytníků. Např. *Bos taurus* sice má kanonické telomerové sekvence, ale vždy v relativně malém množství. U rodů *Ovis* nebo *Capra* se v datech vyskytují kanonické telomerové sekvence v abudancích srovnatelných s motivem `TCAAGG`. (Jako dominantní se je u nich přítomen motiv `TTTCCCACGAGGC`.)

Kanonické telomerové motivy se nepodařilo identifikovat u *Lama glama* a *Vicugna pacos*.

U kaloně *Eidolon helvum* byl jako dominantní nalezen „rostlinný“ motiv. Asi jde o kontaminaci jeho rostlinnou potravou.

Netypickým se jeví *Pavo cristatus*, kanonická sekvence je přítomna v nízké míře a ve zpracovaných souborech dominuje motiv `ATTCTATG`.

Kanonická sekvence nebyla nalezena u želvy *Emydura macquarii*, jsou však přítomny motivy `TTGGGG`, `TTTGGGG` nebo `TTGGG` – poslední jmenovaná sekvence je v menší míře přítomna např. i u *Shinisaurus crocodilurus* (krokodýlovec čínský, u něj však byla detekována kanonická sekvence).

V případě *Thamnophis elegans* a *Thamnophis sirtalis* (užovka) byla kromě kanonické sekvence v menší míře zaznamenána i sekvence `TTTAGG`. U *Pseudonaja textilis* (pakobra východní) byla kromě kanonické sekvence v menší míře nalezena `TTGGGG`.

Kanonickou sekvenci se nepodařilo dostatečně zřetelně identifikovat u *Protobothrops flavoviridis*, ve čtyřech zpracovaných souborech se vždy vyskytovala ve velmi malém množství. Nepodařilo se ji identifikovat ani u *Python molurus* či *Aneides lugubris*, u jeho příbuzného *Aneides aeneus* pak přítomna byla, ale nikoli mezi dominantními.

3.2.11 Stramenopila

V této skupině se hojně vyskytuje jak kanonická sekvece, tak i „rostlinná“ sekvence.

Kanonická sekvence se vyskytuje např. u rodů *Nitzschia*, *Fistulifera*, *Phaeodactylum*, *Nannochloropsis*, *Ectocarpus* či *Saccharina*.

„Rostlinná“ sekvence se vyskytuje např. u rodů *Albugo*, *Hyaloperonospora*, *Peronospora*, *Phytophthora* (výjimkou je *Phytophthora nicotianae* u níž kanonická sekvence převažuje nad „rostlinnou“) či *Pythium*

U rodů *Schizochytrium* či *Labyrinthulid* byla nalezena „hmyzí“ sekvence. U rodu *Aphanomyces* byla nalezena kandidátní sekvence `TTTCGGGG`.

Bez vhodné kandidátní sekvence zůstávají rody *Thalassiosira* a *Saprolegnia*.

3.2.12 Chlorophyta

Za převažující lze považovat „rostlinnou“ sekvenci, ovšem míra její konzervovanosti se z výsledků zdá být relativně nízká. Chybí např. u *Chlamydomonas reinhardtii*, u které byly identifikovány kandidátní motivy `TTGGGC` a `TTTTAGGG`. U *Helicosporidium sp.* dominuje „hmyzí“ sekvence. V případě rodu *Prototheca* je přítomna kanonická sekvence, u *Prototheca stagnorum* pak kanonická sekvence spolu s kandidátní sekvencí `TTAGGC`.

3.2.13 Streptophyta

Za nejobvyklejší telomerovou sekvenci lze v rámci tohoto taxonu považovat `TTTAGGG`.

Zajímavé jsou některé druhy rostlin jako *Cycas revoluta* (*Cycadaceae*), *Ephedra altissima* (*Ephedraceae*) či *Gnetum gnemon* (*Gnetaceae*), u kterých se „rostlinná“ sekvence vyskytuje spolu se sekvencí `TTCAGGG`. „Rostlinná“ sekvence nebyla nalezena ani u *Amborella trichopoda* (*Amborellaceae*).

„Rostlinné“ motivy nebyly v dominantním zastoupení identifikovány ani u rostlin čeledi *Cupressaceae* (cypřišovité). U většiny byla (spíše v menší míře) zjištěna kandidátní sekvence `TGGTCG`. U rodu *Juniperus* (jalovec) byly „rostlinné“ motivy nalezeny, ovšem spolu s kanonickými motivy `TTAGGG`, které byly přítomny v ještě větším množství.

Jinou atypickou skupinou je čeleď *Pinaceae* (borovicovité). U rodu *Picea* (smrk) konzistentně dominuje motiv `TAGGC`, „rostlinná“ sekvence je zastoupena v malé míře. V případě rodu *Pinus* (borovice) dominuje standardní „rostlinná“ sekvence, ale motiv `TAGGC` je rovněž přítomen. Výjimkou je *Pinus lambertiana*, u které je „rostlinná“ sekvence přítomna v nízké míře a dominuje motiv `TTAAGG`. Známé telomerové motivy nebyly v dostatečném množství identifikovány ani u druhů *Pseudotsuga menziesii* a *Abies sibirica*.

„Rostlinný“ telomerový motiv nebyl nalezen u některých zástupců čeledi *Araliaceae* (aralkovité) jako např. *Meryta sinclairii* či rod *Panax*.

Netypické jsou rody *Calycadenia*, *Hemizonia* a *Osmadenia* (všechny patří do taxonu *Madieae* čeledi *Asteraceae*), „rostlinná“ sekvence se zdá být nahrazena motivem `TGGTCG`. Žádné kandidátní sekvence nebyly nalezeny ani u rodu *Tragopogon* (kozí brada).

Zvláštními druhy jsou *Camellia sinensis* (čajovník čínský, *Theaceae*), u nějž výrazně převažuje sekvence `TTAGG`, či *Eremocarya lepida* (*Boraginaceae*), u nějž dominuje motiv `TTAGGC`, i když „rostlinná“ sekvence je v menší míře rovněž přítomna. U rodu *Asclepias* (klejicha, *Apocynaceae*) sice dominuje „rostlinná“ sekvence, ovšem v menší míře je často přítomen i motiv `TTCAGGG`, případně `TGGTCG`, `TTAGG` či `TTAGGG`. Kandidátní sekvence nebyly nalezeny u rodu *Psychotria* (*Rubiaceae*), „rostlinné“ sekvence jsou u něj zastoupeny v poměrně malé míře. U druhu *Cassinopsis madagascariensis* (*Icacinaceae*) dominují dlouhé, variabilní motivy jako `AATACCCCTAGTTGGTCAAATTTGACCAA`. Kandidátní sekvence dále nebyly nalezeny u rodu *Anemopaegma* (*Bignoniaceae*).

Určité odchylky byly zaznamenány také u čeledu hluchavkovitých (*Lamiaceae*). U druhu *Ocimum tenuiflorum* dominuje „rostlinná“ sekvence, ovšem v menší míře jsou zastoupeny i motivy `TTTTAGGG`, `TTCAGGG` či `TTTCGGG`. Podobně u druhů *Clerodendrum bungei* a *Rotheca myricoides* je kromě „rostlinné“ sekvence v menší míře zastoupena i sekvence `TTCAGGG`, u *Cornutia pyramidata* kromě „rostlinné“ i `TTAGGG`, u *Genlisea nigrocaulis* kromě „rostlinné“ i `TGGTCG`.

V případě druhu *Jasminum sambac* (*Oleaceae*) je „rostlinná“ sekvence přítomna v relativně malém množství a dominuje sekvence `TTTAGG`.

Žádné vhodné kandidátní sekvence nebyly nalezeny u druhů *Schwalbea americana* a *Lindenbergia philippensis* (*Orobanchaceae*). Atypický je rod *Erythranthe* (kejklířka, *Phrymaceae*), u něhož byl nalezen alternativní motiv `TTTCGGG`.

U druhů *Solanum melongena* a *Solanum scabrum* (*Solanaceae*) dominuje sekvenční motiv `TTCAGGG`, ačkoli u ostatních druhů rodu *Solanum* převažuje standardní „rostlinná“ sekvence. Rostlina *Cestrum elegans* (*Solanaceae*), u níž byla popsána telomerová sekvence `TTTTTTAGGG` tuto sekvence v datech obsahuje, ovšem v relativně malém množství, dominoval naopak motiv `TTTTTAGCAG` [79].

V případě rodu *Echinocereus* byla jako dominantní zjištěna kanonická telomerová sekvence. U druhu *Aldrovanda vesiculosa* je „rostlinná“ sekvence přítomna v menší míře a je co do počtu převýšena možnými kandidátními sekvencemi `TTAGGG`, `TTGGGG` a `TTTTGGGG`.

Vhodná kandidátní telomerová sekvence nebyla identifikována u druhu *Fagopyrum tataricum* (pohanka tatarská, *Polygonaceae*).

V případě druhů *Leucomphalos mildbraedii*, *Canavalia cathartica* a *Lupinus texensis* (*Fabaceae*) byla nalezena kandidátní sekvence `TGGTCG`, „rostlinná“ sekvence nebyla v daném datovém souboru vůbec přítomna.

Netypickým je rod *Arachnis* (*Fabaceae*), u něhož převažuje „rostlinná“ sekvence, ale v menší míře je přítomna i sekvence `TTTTTAGGG`.

V případě druhů *Chrysolepis sempervirens*, *Fagus crenata* a *Fagus sylvatica* (*Fagaceae*) byla nalezena kandidátní sekvence `TGGTCG`. Stejně je tomu u rodů *Fuscospora*, *Lophozonia*, *Nothofagus* či *Trisyngyne* (*Nothofagaceae*).

Známé telomerové motivy nebyly nalezeny u čeledi *Chrysobalanaceae* (zlatoplodovité), u dvou zástupců (*Chrysobalanus icaco* a *Licania alba*) byla nalezena kandidátní sekvence `TTTATAAGGG`.

„Rostlinná" telomerová sekvence nebyla nalezena u rodu *Euphorbia* (*Euphorbiaceae*), místo ní byly nalezeny kandidátní sekvence `TGGTCG` (*Euphorbia marginata*) a `TTAGGG` (*Euphorbia esula*).

U rodu *Rafflesia* (*Rafflesiaceae*) byla identifikována kandidátní sekvence `TCAGGC`. U druhu *Trema micrantha* (*Cannabaceae*) byla nalezena kandidátní sekvence `TGGTCG`.

V případě rodu *Prunus* (*Amygdaloideae*) byla vždy přítomna „rostlinná" telomerová sekvence, ale u některých druhů (*P. armeniaca*, *P. davidiana*, *P. ferganensis*, *P. persica*, ...) výrazně dominovala sekvence `TTGCC`.

U rodu *Brassica* (*Brassicaceae*) se kromě dominantní „rostlinné" sekvence vyskytovaly i alternativní motivy `TTTCGGG` a `TTTGGGG`.

Kandidátní sekvence `TGGTCG` byla nalezena u *Limnanthes douglasii* (*Limnanthaceae*), *Viviania marifolia* a *Wendtia gracilis* (*Francoaceae*), u rodů *Erodium* a *Geranium* (*Geraniaceae*). Výjimkami jsou *Erodium moschatum* s nalezeným „rostlinným" motivem a *Erodium chrysanthum* se dvěma dominantními motivy – „rostlinnou" sekvencí a sekvencí `TTCAGGG`. Z čeledi *Geraniaceae* byla kandidátní sekvence `TGGTCG` nalezena ještě u *Monsonia vanderietiae* a *Pelargonium capitatum*.

Žádná kandidátní telomerová sekvence nebyla nalezena u *Daphne pseudomezereum* (*Thymelaeaceae*). V případě rodu *Oenothera* (*Onagraceae*) byla „rostlinná" sekvence přítomna v poměrně malém množství a výrazně dominovala sekvence `ATGTTCATCC`. V případě druhu *Pistacia vera* (*Anacardiaceae*) dominovala „rostlinná" sekvence, ale současně s ní byla přítomna i pozměněná sekvence `TTTTAGGG`. Vhodná kandidátní sekvence nebyla nalezena u druhů *Azadirachta indica* (*Meliaceae*) a *Acer miyabei* (*Sapindaceae*). U druhů *Tetrastigma cruciatum* a *T. leucostaphylum* (*Vitaceae*) byla „rostlinná" sekvence přítomna spolu s motivem `TTTTAGGG`.

V případě druhů *Zostera marina* (*Zosteraceae*), *Agapanthus africanus*, *Scadoxus cinnabarinus* (*Amaryllidaceae*), *Agave tequilana*, *Aphyllanthes monspeliensis*, *Asparagus officinalis*, *Dichelostemma ida-maia*, *Lomandra longifolia*, *Sansevieria trifasciata* (*Asparagaceae*), *Haworthia cymbiformis* (*Asphodelaceae*), *Ledebouria cordifolia* (*Hyacinthaceae*) a *Crocus sativus* (*Iridaceae*) byla nalezena kanonická telomerová sekvence.

U rodů *Cymbidium* či *Dendrobium* (*Orchidaceae*) dominuje „rostlinná" telomerová sekvence, ale současně byly nalezeny i odlišné motivy jako `TTTAGG`, `TTTTAGG`, `TTTTAGGG`, `TTAGGG` a jiné. U druhů *Dendrobium falconeri* a *D. moschatum* dokonce kanonická sekvence výrazně převažuje nad „rostlinnou", u druhů *D. lituiflorum* a *D. parishii* jsou zastoupeny v přibližně stejném množství „rostlinná" sekvence a motiv `TTTTAGG`.

V případě druhu *Festuca arundinacea* (*Poaceae*) dominují kanonická a „rostlinná" sekvence, jsou přitom zastoupeny ve srovnatelném množství. „Rostlinný" motiv nebyl v dostatečném zastoupení detekován u druhů *Triticum dicoccoides* a *T. spelta*

(*Poaceae*). U rodu *Sorghum* (*Poaceae*) se kromě dominantního „rostlinného“ motivu vyskytoval i motiv `TTCAGGG`, u rodu *Tripsacum* (*Poaceae*) se kromě „rostlinného“ motivu vyskytoval i motiv kanonický.

Žádné kandidátní telomerové sekvence nebyly identifikovány u čeledi *Zingiberaceae* a u rodu *Fritillaria* (*Liliaceae*).

V případě rodů *Liriodendron* a *Magnolia* (*Magnoliaceae*) se „rostlinná“ telomerová sekvence vyskytovala spolu s alternativním motivem `TTTAGG`.

U druhu *Papaver somniferum* (*Papaveraceae*) byla nalezena kandidátní sekvence `TTCAGGG`.

V případě rodu *Isoetes* (*Isoetaceae*) byla „rostlinná“ sekvence zastoupena minoritně. Kromě ní byly nalezeny kandidátní sekvence `TTAGGG` a `TTTTGGGG`.

3.3 Analýza repetic u rostlin čeledi *Solanaceae*

Dle zadání práce byly v rámci projektu dále zpracována sekvenační data rostlin čeledi lilkovitých (*Solanaceae*). Celkem se jednalo o 112 genomů.[75] Informace o taxonomickém zařazení vzorků byly poskytnuty na úrovni rodu, různé druhy v rámci rodu byly v názvech rozlišeny pouze identifikačními čísly. Data byla analyzována pomocí `TRFi` dle příkazu

```
trf407b.linux64 2 7 7 80 10 50 50 -h -ngs
```

a byly hledány tandemové repetice o minimální délce 3 a minimálním počtu opakování 3.

Nejčastějším telomerovým motivem byla sekvence `TTTAGGG`. U rodu *Salpiglossis* však byl nalezen alternativní motiv `TTCAGGG`. Tentýž motiv byl nalezen u jednoho druhu rodu *Solanum*, dle výsledků HTS by se mohlo jednat o *S. melongena* nebo *S. scabrum*. Navzdory očekávání nebyla u rodu *Cestrum* nalezena jeho telomerová sekvence `TTTTTTAGGG`, u něj a u rodu *Sessea* nebyly nalezeny žádné známé telomerové motivy. U řady zpracovaných souborů byla zjištěna přítomnost jak `TTTAGGG`, tak `TTCAGGG`, přičemž standardní „rostlinný“ motiv zpravidla převažoval. V některých případech však byla jejich množství srovnatelná (*Rahowardiana*, *Nicandra*, *Eriolarynx*, *Dunalia*, *Vassobia*, čtyři druhy *Solanum*). U jednoho druhu *Solanum* a u *Salpichroa* byla nalezena kanonická telomerová sekvence. Jelikož byly v obou případech přítomny i „rostlinné“ motivy, jde pravděpodobně o kontaminaci vzorku, nejspíše „houbovým“ organismem.

[75] Sekvenační data byla poskytnuta ze zahraničního pracoviště v rámci výzkumné spolupráce.

Kapitola 4

Fylogenetická analýza

Vedlejším cílem práce bylo zjistit, zda a nakolik množství tandemových repetic v genomech zjištěná programem TRFi korespondují s evolucí příslušných taxonů, tedy zda bude možné registrovat větší rozdíly mezi evolučně vzdálenějšími druhy a větší podobnost mezi příbuznějšími druhy. Analýza byla provedena v softwaru `R`.

Data byla zpracována pomocí metody nejbližšího souseda (neighbor joining, NJ). Bylo vybráno jedenáct druhů obratlovců, v jejichž genomech byly pomocí TRFi identifikovány tandemové repetice způsobem analogickým dříve provedenému HTS. Šlo o druhy: *Rattus norvegicus*, *Pan troglodytes*, *Corvus cornix*, *Corvus brychyrhynchos*, *Mus caroli*, *Pygoscelis adeliae*, *Parus major*, *Aptenodytes forsteri*, *Pan paniscus*, *Mus musculus* a *Thamnophis elegans*. Kvůli srovnatelnosti bylo za každý datový soubor zpracováno stejné množství (1 milion) sekvenčních čtení. Všechny datasety pocházely ze stejné sekvenační platformy a měly srovnatelnou průměrnou délku čtení (199 až 201).

Před výpočtem matice vzdáleností bylo nutné provést standardizaci dat (druhy jsou pozorování a sekvenční motivy jsou proměnné). Byly vyzkoušeny tři varianty standardizace. Jednak logaritmizace při základu 10 (ve tvaru $\log(x+1)$ kvůli „nulám“ v datech), dále normalizace „min-max“ podle vzorce

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}},$$

kde $x_{\min}$ představuje minimální hodnotu v rámci proměnné x a $x_{\max}$ maximální hodnotu v rámci proměnné x, a za třetí kombinace obojího.

Dále bylo třeba vypočítat matici vzdáleností. Bylo třeba řešit tzv. problém „double-zero“, tedy aby současná nepřítomnost sekvencí mezi dvěma datovými soubory nebyla intepretována jako jejich podobnost.[76] Pro srovnání tedy byla matice vzdáleností vypočtena s využitím eukleidovské metriky, která problém „double-zero“ neřeší, a dále pomocí tzv. Brayovy–Curtisovy nepodobnosti, která tento problém řeší.

Eukleidovská vzdálenost dvou bodů (vzorků) $x = (x_1, \ldots, x_n)^T$ a $y = (y_1, \ldots, y_n)^T$ v n-rozměrném prostoru byla počítána podle vztahu

[76] Ilustrativní příklad problému „double-zero“ z jiné oblasti: Bude-li sledován výskyt tygra, zebry, krokodýla a slona na dvou lokalitách – v ČR a v Arktidě, nelze ze současného nevýskytu příslušných organismů usuzovat na podobnost obou lokalit. Ze současné přítomnosti těchto druhů na nějakých dvou lokalitách však na jejich podobnost usuzovat lze.

$$d_{\text{eukl}}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

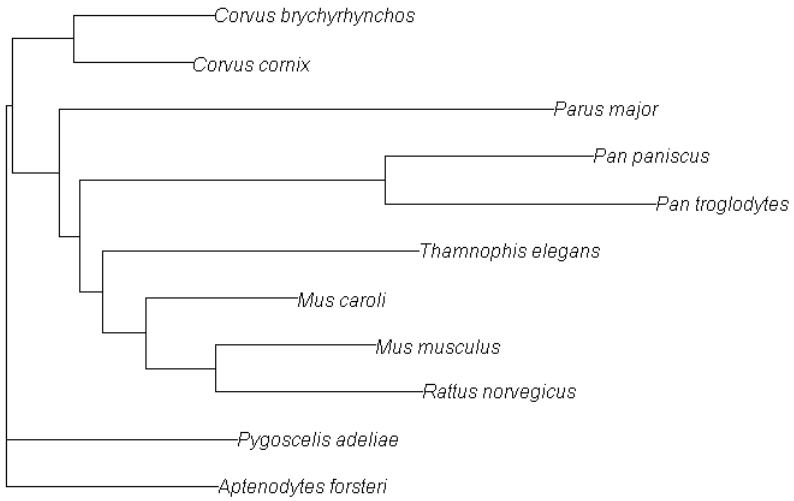
Brayova–Curtisova nepodobnost byla počítána podle vztahu

$$d_{\text{BC}}(x, y) = \frac{\sum_{i=1}^{n}|x_i - y_i|}{\sum_{i=1}^{n}(x_i + y_i)}.$$

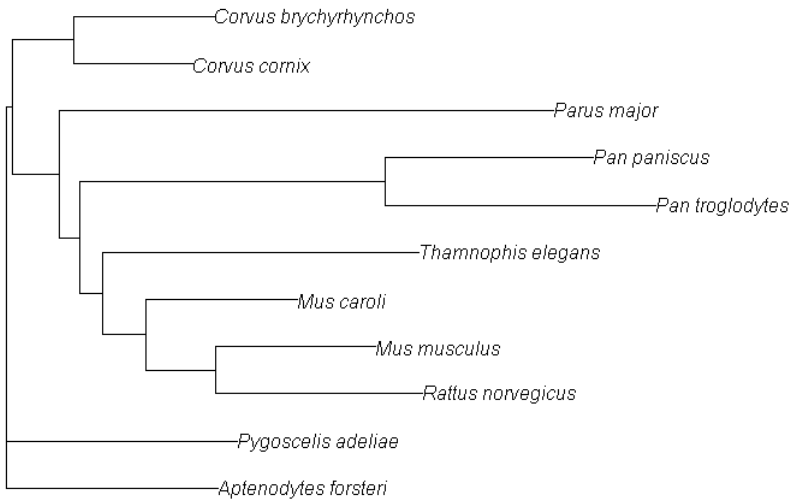
Celkem tak bylo vytvořeno šest různých matic vzdáleností. Na základě nich byly vykresleny fylogenetické stromy metodou NJ (funkce `ape::nj()`). Šest různých fylogenetických stromů je na obrázcích 4.1, 4.2, 4.3, 4.4, 4.5, 4.6.

Celkově výsledky nelze hodnotit jako uspokojivé. Je patrné, že klastrování je schopno poskytovat dostatečně spolehlivé výsledky pouze na úrovni některých relativně příbuzných taxonů, např. oba druhy rodu *Corvus* či *Pan* jsou konzistentně řazeny k sobě, naopak ptáci jako takoví ani savci jako takoví v žádném sestaveném stromu neutvořili monofylum. Ani některé blízce příbuzné druhy (*Mus musculus* a *Mus caroli*) nebyly v žádném z fylogenetických stromů uvedeny jako sesterské.

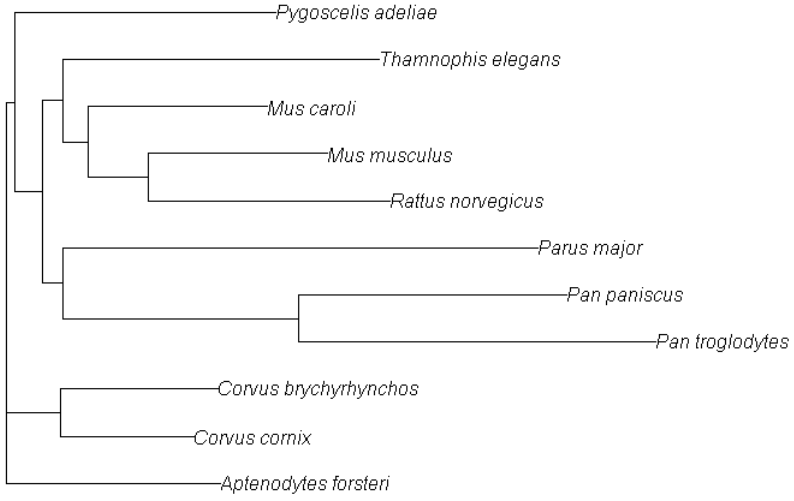
Neprokázalo se tedy, že by bylo možné využít informací z výstupů TRFi jako alternativní metodu k tvorbě fylogenetických stromů. Změny v počtu a složení tandemových repetic v genomech patrně nelze takto přímočaře korelovat s fylogenetickým vývojem. Již dříve bylo sice zjištěno, že tandemové repetice nesou fylogenetický signál, ale k jeho přesnější interpretaci je zapotřebí sofistikovanějších metod [27], [65].



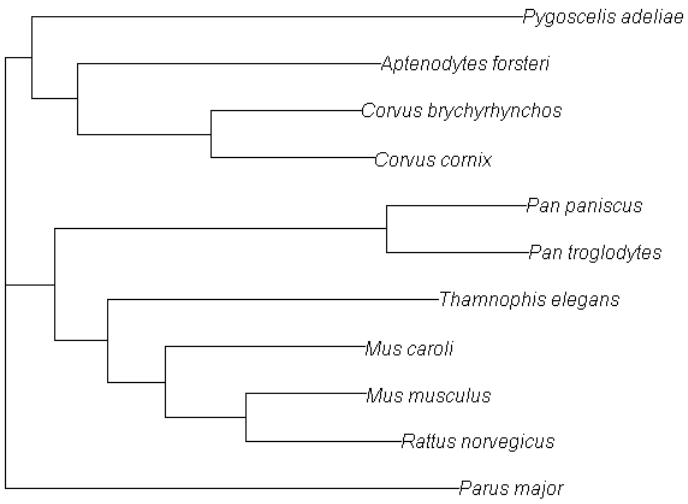
Obrázek 4.1: Fylogenetický strom sestavený pomocí eukleidovských vzdáleností na zlogaritmovaných datech.



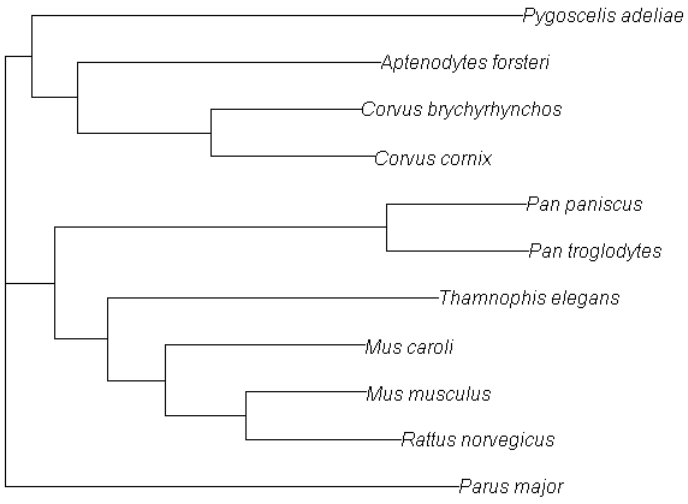
Obrázek 4.2: Fylogenetický strom sestavený pomocí eukleidovských vzdáleností na normalizovaných datech.



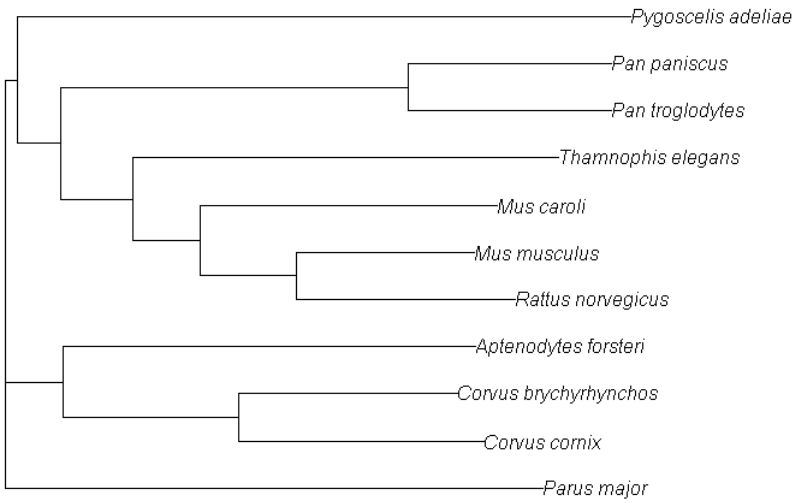
Obrázek 4.3: Fylogenetický strom sestavený pomocí eukleidovských vzdáleností na zlogaritmovaných a poté normalizovaných datech.



Obrázek 4.4: Fylogenetický strom sestavený pomocí Brayových–Curtisových nepodobností na zlogaritmovaných datech.



Obrázek 4.5: Fylogenetický strom sestavený pomocí Brayových–Curtisových nepodobností na normalizovaných datech.



Obrázek 4.6: Fylogenetický strom sestavený pomocí Brayových–Curtisových nepodobností na zlogaritmovaných a poté normalizovaných datech.

Kapitola 5

Diskuze

Celý proces výběru, získání, zpracování a vyhodnocení dat byl spojen s různými obtížemi. Prvním účelem této kapitoly je uvést tyto problémy a pokusit se o zhodnocení jejich závažnosti. Dále bude uvedeno celkové zhodnocení dosažení výsledků.

5.1 Možné příčiny chybných výsledků

Prvním problém souvisí se samotnou kvalitou získaných dat. Data pocházejí od „koncových“ uživatelů, tedy z různých sekvenačních laboratoří. Neexistuje proto žádná záruka kvality provedené sekvenace ani správnosti anotace příslušného datového souboru.

S ohledem na kvalitu sekvenovaného vzorku je třeba upozornit zejména na problém kontaminace „muzejních“ vzorků houbovými organismy, u jiných vzorků pak jde i o zbytky potravy zkonzumované sekvenovaným organismem. Kvůli tomu byly zaznamenány případy falešné pozitivity – např. v genomech hmyzu mohly být zaznamenány „houbové“ telomerové motivy nebo v genomu želvy či kaloně „rostlinné“ telomerové motivy. Konkrétním příkladem může být datový soubor `SRR5062042` – genom plísně *Eutiarosporella tritici-australis* napadající obilí, genom byl izolován z obilných zrn a obsahující hojné množství rostlinných telomerových motivů.

Důležitým zdrojem falešně pozitivních či falešně negativních výsledků jsou i chyby v anotacích. Jako příklad lze uvést datový soubor `SRR5500911`, organismus je veden jako *Boodlea composita* a uvedená anotace je Strategy: WGS, Source: GENOMIC, Selection: RANDOM. Současně je však v popisu experimentu uvedeno, že se jedná o sekvenaci chloroplastové DNA. Uvedená anotace je tedy vzhledem k provedení experimentu nesprávná, zejm. Selection: RANDOM mělo být nahrazeno adekvátějším výrazem vystihujícím povahu provedené sekvenace.

Důvodem falešné pozitivity může být kromě kontaminace či jiné chyby v přípravě knihovny také přítomnost ITS. Jejich abundance bude patrně nižší než u pravých telomerových sekvencí. Je však těžké rozlišit, zda málo zastoupená známá telomerová sekvence je skutečnou telomerovou sekvencí podhodnocenou kvůli chybě v přípravě dat, zda jde o ITS nebo o kontaminaci.

Mezi výsledky byly nalezeny i případy falešné negativity „maskované“ falešnou pozitivitou. Jako příklad lze uvést datový soubor šimpanze (*Pan paniscus*)

ERR032779, při prvním screeningu nebyla identifikována standardní telomerová sekvence, ale motiv TGGAA jako nejhojněji zastoupený. Při druhém běhu screeningu byly analyzovány další tři soubory (ERR032310, SRR740808, ERR032964), v nich již byla identifikována kanonická sekvence, ale ne mezi dominantními motivy. Na prvním místě byl vždy motiv TGGAA. Přítomnost standardní telomery však byla u šimpanzů experimentálně ověřena, proto se spíše jedná o chyby v přípravě vzorků než o „velký objev“ [61].

Chyby v anotacích se dále týkají i sekvenačních platoforem, např. datový soubor SRR3303112 má pocházet z platformy Illumina (HiSeq 2000), ale průměrná délka sekvenčních čtení přesahuje 24 tisíc bp. Pravděpodobně tedy pochází z jiné sekvenační platformy.

Problém falešné negativity rovněž souvisí s použitou sekvenační platformou. Nejméně „vhodná“ je platforma 454, u níž byla zaznamanána nejvyšší míra negativních výsledků. Důvod může souviset např. s vysokou chybovostí sekvenací homopolymerních úseků. Ostatní rozšířené platformy (Illumina, Ion Torrent, Pac Bio) poskytují srovnatelné výsledky co do poměru pozitivních a negativních výsledků. Lze si klást otázku, zda tato míra pozitivity závisí na délce sekvenčních čtení. Z logiky věci tomu tak musí být, protože repetitivní motivy dlouhé např. 30 bp nemohou být identifikovány v sekvenčních čteních dlouhých např. jen 50 bp. Rovněž zpracování dat ukázalo jistou souvislost mezi podílem pozitivních výsledků a délkou čtení (i u relativně krátkých sekvenčních motivů), ale z toho nelze přímočaře usuzovat na kauzální vztah, protože zavádějícím faktorem může být (a nejspíš také je) použitý sekvenátor, nejde tedy o artefakt algoritmu TRFi, alespoň ne tehdy, pokud délka čtení několikanásobně přesahuje délku vyhledávaného sekvenčního motivu.

Je také třeba podotknout, že provedený screening má „orientační“ povahu, a zvláště u negativních výsledků, včetně těch s kandidátní telomerovou sekvencí, nemůže sloužit jako nástroj spolehlivého potvrzení přítomnosti či nepřítomnosti telomerové sekvence, ale spíše jako nástroj předběžné bioinformatické explorace schopné naznačit slibné směry dalšího výzkumu.

5.2 Problémy spojené s prováděnými výpočty

Další problematická oblast, kterou je třeba zmínit, je samotné získávání dat a provádění výpočtů. Stažení dat pomocí nástroje SRA Toolkit může selhat (jsou staženy prázdné soubory), aniž by o tom program informovat chybovým hlášením. Vzhledem k relativně velkému objemu stahovaných dat bylo toto stahování prováděno pomocí dávkových úloh, ale to vedlo k malé efektivitě využití rezervovaných výpočetních kapacit (na druhou stranu, spouštět tyto úlohy na čelním uzlu nebylo přijatelné vzhledem k relativně dlouhé době stahování). Rovněž dávkové úlohy spouštějící TRFi nevyužívaly rezervované zdroje optimálně (využití procesoru mnohdy v „malých desítkách“ procent). Při prvním běhu screeningu bylo vždy zpracováváno 10 datových souborů současně, ale každý z nich se zpracovával různou dobu, úloha proto neskončila, dokud nebyl zpracován poslední z nich.

5.3 Souhrnná interpretace dosažených výsledků

Telomerové sekvence jsou obvykle prezentovány jako sekvenčně konzervované úseky genomu. Z tohoto důvodu bylo překvapením, že z pěti tisíc zpracovaných genomů v rámci prvního běhu HTS bylo bezmála 1300 bez dominantně zastoupené známé telomerové sekvence (tj. asi $\frac{1}{4}$ výsledků).

Při druhém běhu HTS byl celkový počet negativních výsledků asi 260, ovšem toto číslo není zcela směrodatné, protože jednak ne všechny negativní výsledky bylo možné ověřit na jiných datových souborech (nebyly vždy dostupné) a jednak identifikovaný známý telomerový motiv překročivší příslušný „práh“ pozitivity nemusí být ještě pravou telomerovou sekvencí.

Samotná pozitivita a negativita výsledků navíc sama neurčuje, zda se jedná o „zajímavý“ výsledek, protože ten může mít i povahu záměny jedné známé telomerové sekvence za jinou známou sekvenci. Takové případy nebyly druhým během HTS pokryty, protože byly hledány a nalézány postupně při ručním zpracování fylogeneticky uspořádaných dat, zatímco „ověřovací“ screening proběhl již dříve a byl založen na plně automatizované práci s daty.

Souhrnně je možné říci, že jedinou „velkou“ skupinou organismů, u níž byla v rámci HTS pozorována silná konzervovanost telomerových motivů, jsou obratlovci (*Vertebrata*). U jednobuněčných eukaryot, „bezobratlých“, „houbových“ organismů a „nižších“ i „vyšších“ rostlin byla zaznamenána nečekaně vysoká četnost různých modifikací běžných telomerových motivů nebo i prostých negativních výsledků, tj. nenalezení žádné kandidátní sekvence.

O variabilitě telomerových motivů svědčí i nalezené rozdíly mezi telomerami napříč různými druhy v rámci jednoho rodu, jako je tomu např. u rodů *Plasmodium*, *Cryptosporidium*, *Tetrahymena*, *Penicilium*, *Prunus* či *Solanum*. Z toho lze vyvodit závěr, že telomerové sekvence se neliší jen mezi „rozsáhlými“ a evolučně vzdálenými taxony, ale i mezi taxony blízce příbuznými.

Občasná přítomnost více různých telomerových motivů může vést k vyslovení hypotézy, že u některých organismů se může současně vyskytovat více různých funkčních telomerových sekvencí, příkladem takových organismů by mohly být některé rostliny čeledi *Solanaceae*, v úvahu připadají např. *Capsicum annulum*, *Rahowardiana sp.* či *Nicandra sp.* Podobná situace byla zjištěna i v genomech některých hub (např. *Infundibulicybe gibba*, *Pholiota highlandensis*) nebo u zástupců hmyzu (např. babočky *Heliconius evenatus* a *H. pardinalis*), ale v těchto případech jde o kanonickou a „rostlinnou“ sekvenci, resp. o `TTAGG` a `TTTAGG`, které se navíc liší délkou repetitivního motivu.

Závěr

Hlavním cílem této bakalářské práce bylo provést plošný screening telomerových sekvencí ve veřejně dostupných datech NGS uložených v databázi SRA. Práce cílila na nalezení neobvyklých telomerových sekvencí – ať už zcela nových kandidátních motivů, nebo motivů již známých, ale netypických pro příslušný taxon. Vedlejšími cíli práce pak bylo zjistit, zda lze ze stanovených abundancí tandemových repetic v genomech jednoduše sestrojit smysluplný fylogenetický strom metodou NJ, a dále analyzovat poskytnutá sekvenační data rostlin čeledi *Solanaceae*.

Práce byla provedena v prostředí MetaCentra, screening byl implementován na bázi dříve vzniklých softwarových nástrojů Tandem Repeats Finder a Tandem Repeats Merger a bylo zpracováno po jednom genomu od každého tehdy dostupného sekvenovaného eukaryontního druhu.

Bylo provedeno vyhledání známých telomerových motivů v nalezených repetitivních motivech, a pokud nebyla nalezena žádná telomerová sekvence, negativita výsledků byla ověřována na jiném datovém souboru, byl-li takový k dispozici.

Zpracovaná data byla automatickým způsobem seřazena dle taxonomie a bylo provedeno jejich vizuální zhodnocení s cílem nalézt kandidáty na neobvyklé telomerové sekvence. Výsledkem zpracování pěti tisíc genomů bylo nalezení několika desítek druhů, u nichž je k dispozici kandidát na novou či jinak neobvyklou telomerovou sekvenci. Přibližně sedm set druhů zůstalo jako „negativní“ výsledky bez kandidátní telomerové sekvence přítomné v dostatečném množství (ať už očekávané, nebo neobvyklé), většina z toho jsou však pravděpodobně výsledky falešně negativní.

Na základě četností nalezených tandemových repetic byl dále sestaven fylogenetický strom s využitím různých metod normalizace dat a výpočtu matice vzdáleností, ale výsledky nebyly uspokojivé. Nepotvrdilo se, že by popsaným způsobem bylo možné sestavovat spolehlivé fylogenetické stromy.

Hlavní přínos práce spočívá v rozšíření poznatků o pestrosti a variabilitě telomerových motivů napříč systémem eukaryot; a této souvislosti lze závěrem lze konstatovat následující.

1. Výsledky práce ukázaly nečekaně velké množství výjimek, odchylek a jiných abnormalit u telomerových sekvencí napříč systémem eukaryot. Telomerové motivy se tak ukazují být ze sekvenčního hlediska měnlivějšími, než se očekávalo a předpokládalo.

2. Výsledky práce naznačují, že telomerové sekvence se mohou poměrně často lišit i mezi fylogeneticky příbuznými taxony, např. i mezi různými druhy v rámci jednoho rodu.

3. Na základě relativně častých nálezů více různých telomerových motivů přítomných ve srovnatelném množství v rámci jediného genomu lze vyslovit hypotézu, že u některých organismů může existovat více různých funkčních telomerových sekvencí.

Seznam obrázků

Seznam použité literatury

[1] ANDERSON, Stephen. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.* 9(13): 3015—3027. doi: 10.1093/nar/9.13.3015

[2] ANSORGE, Wilhelm, Brian S. SPROAT, Josef STEGEMANN a Christian SCHWAGER, 1986. A non-radioactive automated method for DNA sequence determination. *Journal of Biochemical and Biophysical Methods* 13(6): 315–323, doi:10.1016/0165-022X(86)90038-2

[3] ARI, Sule a Muzaffer ARIKAN, 2016. Next-Generation Sequencing: Advantages, Disadvantages, and Future v *Identification of Gene Families Using Genomics and/or Transcriptomics Data*: 109–135, doi: 10.1007/978-3-319-31703-8_5

[4] AZZALIN, Claus M., Patrick REICHENBACH, Lela KHORIAULI, Elena GIULOTTO a Joachim LINGNER, 2007. Telomeric Repeat–Containing RNA and RNA Surveillance Factors at Mammalian Chromosome Ends. *Science.* 318(5851): 798–801. doi: 10.1126/science.1147182

[5] BANDARIA, Jigar N., Peiwu QUIN, Veysel BERK, Steven CHU a Ahmet YILDIZ, 2016. Shelterin Protects Chromosome Ends by Compacting Telomeric Chromatin. *Cell.* 164(4): 735–746. doi: 10.1016/j.cell.2016.01.036

[6] BANFALVI, Gaspar a Gabor NAGY, 2011. Interphase Chromosome Structures of Human Cells. *DNA and Cell Biology.* 30(12): 1007–1009. Dostupné z: doi: 10.1089/dna.2011.1288

[7] BÄR, Christian a Maria A. BLASCO, 2016. Telomeres and telomerase as therapeutic targets to prevent and treat age-related diseases. *F1000Research* 5. doi: 10.12688/f1000research.7020.1

[8] BEHJATI, Sam a Patrick S. TARPEY, 2013. What is next generation sequencing? *Arch Dis Child Educ Pract* 98(6): 236-–238. doi: 10.1136/archdischild-2013-304340

[9] BEJARANO, Leire, Giuseppe BOSSO, Jessica LOUZAME, Rosa SERRANO, Elena GÓMEZ-CASERO, Jorge MARTÍNEZ-TORRECUADRADA, Sonia MARTÍNEZ, Carmen BLANCO-APARICIO, Joaquín PASTOR a Maria A. BLASCO, 2019. Multiple cancer pathways regulate telomere protection. *EMBO Mol Med* 11:e10292, doi:10.15252/emmm.201910292

[10] BENSON, Dennis A., Mark CAVANAUGH, Karen CLARK, Ilene KARSCH-MIZRACHI, David J. LIPMAN, James OSTELL a Eric W. SAYERS, 2013. GenBank. *Nucleic Acids Res.* 41(Database issue): D36—D42. doi: 10.1093/nar/gks1195

[11] BENSON, Gary, 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*27(2): 573–580.

[12] BENSON, Gary, 2016. Webové stránky programu Tandem Repeats Finder. https://tandem.bu.edu/trf/trf.html

[13] BENTLEY, David R., Shankar BALASUBRAMANIAN, (...) a Anthony J. SMITH, 2008 *Nature* 456(7218): 53-–59. doi: 10.1038/nature07517

[14] BIESSMANN, Harald, James M. MASON, Kristian FERRY, Marie D'HULST, Katrin VALGEIRSDOTTIR, Karen L. TRAVERSE, Mary-Lou PARDUE, 1990. Addition of telomere-associated HeT DNA sequences \heals" broken chromosome ends in *Drosophila*. *Cell* 61(4): 663–73, doi:10.1016/0092-8674(90)90478-w

[15] BLACKBURN, Elisabeth a Joseph Grafton GALL, 1978. A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. *J. Mol. Biol.* 120(1):33-–55.

[16] BOLZÁN, Alejandro D. Interstitial telomeric sequences in vertebrate chromosomes: Origin, function, instability and evolution, 2017. *Mutation Research.* 773: 51–65. doi: 10.1016/j.mrrev.2017.04.002

[17] CANARD, Bruno a Robert S. SARFATI, 1994. DNA polymerase fluorescent substrates with reversible 3'-tags. *Gene* 148: 1–6

[18] CASACUBERTA, Elena, 2017. Drosophila: Retrotransposons Making up Telomeres. *Viruses.* 9(7): 192. doi: 10.3390/v9070192

[19] CASIENS, Sherwood, 1998. The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.* 32: 339-–77, doi: 10.1146/annurev.genet.32.1.339

[20] COLLINS Kathleen, 2006. The biogenesis and regulation of telomerase holoenzymes. *Nat Rev Mol Cell Biol.* 7(7): 484—494. doi: 10.1038/nrm1961

[21] COREY, David R., 2009. Telomeres and Telomerase: From Discovery to Clinical Trials. *Chem. Biol.* 16(12): 1219–1223. doi: 10.1016/j.chembiol.2009.12.001

[22] CUBILES, María D., Sonia BARROSO, María I. VANQUERO-SEDAS, Alicia ENGUIX, Andrés AQUILERA a Miguel A. VEGA-PALAS, 2018. Epigenetic features of human telomeres. *Nucleic Acids Res.* 46(5): 2347-–2355. doi: 10.1093/nar/gky006

[23] CUSANELLI, Emilio a Pascal CHARTRAND, 2015. Telomeric repeat-containing RNA TERRA: a noncoding RNA connecting telomere biology to genome integrity. *Front Genet.* 6: 143. doi: 10.3389/fgene.2015.00143

[24] CVRČKOVÁ, Fatima, 2006. Úvod do praktické bioinformatiky. Academia.

[25] DENG, Zhong, Julie NORSEEN, Andreas WIEDMER, Harold RIETHMAN a Paul M. LIEBERMAN, 2009. TERRA RNA Binding to TRF2 Facilitates Heterochromatin Formation and ORC Recruitment at Telomeres. *Mol Cell.* 35(4): 403-–413. doi: 10.1016/j.molcel.2009.06.025

[26] DERELLE, Romain, Guifré TORRUELLA, Vladimír KLIMEŠ, Henner BRINKMANN, Eunsoo KIM, Čestmír VLČEK, B. Franz LANG a Marek ELIÁŠ, 2015. Bacterial proteins pinpoint a single eukaryotic root. *Proc Natl Acad Sci USA.* 112(7): 693-–699. doi: 10.1073/pnas.1420657112

[27] DODSWORTH, Stephen, Mark W. CHASE, Laura J. KELLY, Ilia J. LEITSH, Jiří MACAS, Petr NOVÁK, Mathieu PIEDNOEL, Hanna WEISS-SCHNEEWEISS a Andrew R. LEITCH, 2015. Genomic Repeat Abundances Contain Phylogenetic Signal. *Syst Biol.* 64(1): 112–126. doi: 10.1093/sysbio/syu080

[28] DRMANAC, Radoje, Andrew B. SPARKS, (...) a Clifford A. REID, 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science.* 327(5961): 78–81. doi: 10.1126/science.1181498.

[29] FAJKUS, Petr, Vratislav PEŠKA, Zdeňka SITOVÁ, Jana FULNEČKOVÁ, Martina DVOŘÁČKOVÁ, Roman GOGELA, Eva SÝKOROVÁ, Jan HAPALA a Jiří FAJKUS, 2016. *Allium* telomeres unmasked: the unusual telomeric sequence (CTCGGTTATGGG)n is synthesized by telomerase. *Plant Journal* 85(3), 337–347, doi: 10.1111/tpj.13115, dostupné z: https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.13115 (odkaz funkční k 7. září 2019)

[30] FRYDRYCHOVÁ, Radmila, Petr GROSSMANN, Pavel TRUBAC, Magda VÍTKOVÁ a František MAREC, 2004. Phylogenetic distribution of TTAGG telomeric repeats in insects. *Genome* 47(1): 163–178, doi:10.1139/g03-100

[31] FULNEČKOVÁ, Jana, Tereza ŠEVČÍKOVÁ, Jiří FAJKUS, Alena LUKEŠOVÁ, Martin LUKEŠ, Čestmír VLČEK, B. Franz LANG, Eunsoo KIM, Marek ELIÁŠ, Eva SÝKOROVÁ, 2013. A Broad Phylogenetic Survey Unveils the Diversity and Evolution of Telomeres in Eukaryotes. *Genome Biol Evol.* 5(3): 468—483, doi: 10.1093/gbe/evt019, dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3622300/ (odkaz funkční k 7. září 2019)

[32] GONG, Peng, Hua WANG, Jingsong ZHANG, Yudong FU, Zhengmao ZHU, Jinmiao WANG, Yu YIN, Haiying WANG, Zhongcheng ZHOU, Jiao YANG, Linlin LIU, Mo GOU, Ming ZENG, Jinghua YUAN, Feng WANG, Xinghua PAN, Rong XIANG, Sherman M. WEISSMAN, Feng QI a Lin LIU, 2019.

Telomere Maintenance-Associated PML Is a Potential Specific Therapeutic Target of Human Colorectal Cancer. *Transl Oncol.* 12(9): 1164-–1176. doi: 10.1016/j.tranon.2019.05.010

[33] GREIDER, Carolyn W. a Elizabeth H. Blackburn, 1985. Identification of a specific telomere terminal transferase activity in tetrahymena extracts. *Cell.*43(2 Pt 1): 405–13. doi: 10.1016/0092-8674(85)90170-9

[34] GRIFFITH, Jack D., Laurey CORNEAU, Soraya ROSENFIELD, Rachel M. STANSEL, Alessandro BIANCHI, Heidi MOSS a Titia DE LANGE, 1999. Mammalian Telomeres End in a Large Duplex Loop. *Cell* 97, 503—514, doi: 10.1016/s0092-8674(00)80760-6, dostupné z: https://www.cell.com/cell/fulltext/S0092-8674(00)80760-6?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867400807606%3Fshowall%3Dtrue (odkaz funkční k 7. září 2019)

[35] HAYFLICK, Leonard, 1965. The Limited in Vitro Life Time of Human Diploid Cell Strains. *Experimental Cell Research.* 37: 614–636. doi: 10.1016/0014-4827(65)90211-9, dostupné z:https://www.researchgate.net/publication/9269319_The_Limited_in_Vitro_Lifetime_of_Human_Diploid_Cell_Strains (odkaz funkční k 6. září 2019)

[36] HAYFLICK, Leonard a Paul Moorhead, 1961. The serial cultivation of human diploid cell strains. *Exp Cell Res.* 25: 585–621. doi: 10.1016/0014-4827(61)90192-6

[37] HEATHER, James M. a Benjamin CHAIN, 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics* 107(1): 1-–8. doi: 10.1016/j.ygeno.2015.11.003, dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4727787/ (odkaz funkční k 21. září 2019)

[38] HOLLEY, Robert W., J. T. MADISON a A. A. ZAMIR, 1964. A new method for sequence determination of large oligonucleotides. *Biochem. Biophys. Res. Commun.* 17: 389—394.

[39] HUANG, Yu-Feng, Sheng-Chung CHEN, Yih-Shien CHIANG, Tzu-Han CHEN a Kuo-Ping CHIU, 2012. Palindromic sequence impedes sequencing-by-ligation mechanism. *BMC Syst Biol.* 6(2): S10, doi: 10.1186/1752-0509-6-S2-S10

[40] HYMAN, E. D., 1988. A new method of sequencing DNA. *Anal Biochem.* 174(2): 423–36. doi: 10.1016/0003-2697(88)90041-3

[41] CHEN, Fei, Mengxing DONG, Meng GE, Lingxiang ZHU, Lufeng REN, Guocheng LIU a Rong MU, 2013. The History and Advances of Reversible Terminators Used in New Generations of Sequencing Technology. *Genomics Proteomics Bioinformatics* 11(1): 34-–40, doi: 10.1016/j.gpb.2013.01.003

[42] CHEN, Yong, 2019. The structural biology of the shelterin complex. *Biol Chem.* 400(4): 457–466. doi: 10.1515/hsz-2018-0368.

[43] ICHIKAWA Y., Y. NISHIMURA, H. KURUMIZAKA, a M. SHIMIZU, 2015. Nucleosome organization and chromatin dynamics in telomeres. *Biomol Concepts.* 6(1): 67–75. doi: 10.1515/bmc-2014-0035.

[44] JEON, Yoon-Seong, Sang-Cheol PARK, Jeongmin LIM, Jongsik CHUN a Bong-Soo KIM, 2015. Improved pipeline for reducing erroneous identification by 16S rRNA sequences using the Illumina MiSeq platform. *Journal of Microbiology* 53(1): 60-–69. doi: 10.1007/s12275-015-4601-y

[45] KARSCH-MIZRACHI, Ilene, Yasukazu NAKAMURA a Guy COCHRANE, 2012. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* 40(Database issue): D33-–D37. doi: 10.1093/nar/gkr1006

[46] KIRCHER, M. a J. KELSO, 2010. High-throughput DNA sequencing – concepts and limitations.*Bioessays.* 32(6): 524–36. doi: 10.1002/bies.200900181.

[47] KLOBUTCHER, L. A., M. T. SWANTON, P. DONINI a D. M. PRESCOTT, 1981. All gene-sized DNA molecules in four species of hypotrichs have the same terminal sequence and an unusual 3' terminus. *Proc Natl Acad Sci USA* 78(5): 3015–9. doi:10.1073/pnas.78.5.3015

[48] KODAMA, Yuichi, Martin SHUMWAY a Rasko LEINONEN, 2012. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*40(Database issue): D54-–D56. doi: 10.1093/nar/gkr854

[49] KOUBKOVÁ, L., B. VOJTĚŠEK, R. VYZULA, 2014. Sekvenování nové generace a možnosti jeho využití v onkologické praxi. *Klin Onkol* 27: S61–S68.

[50] LAIRD, Diana J. a Irving WEISSMAN, 2004. Telomerase maintained in self-renewing tissues during serial regeneration of the urochordate Botryllus schlosseri. *Dev Biol.* 273(2): 185–94. doi: 10.1016/j.ydbio.2004.05.029

[51] DE LANGE, Titia, 2005. Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev.* 19(18): 2100–10, doi: 10.1101/gad.1346005

[52] LEJNINE, Serguei, Vladimir L. MAKAROV a John P. LANGMORE, 1995. Conserved nucleoprotein structure at the ends of vertebrate and invertebrate chromosomes. *Proc. Natl. Acad. Sci. USA* 92: 2393–2397, doi: 10.1073/pnas.92.6.2393, dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC42490/pdf/pnas01484-0609.pdf (odkaz funkční k 6. září 2019)

[53] LIU, Dan, Matthew S. O'CONNOR, Jun QIN, Zhou SONGYANG, 2004. Telosome, a Mammalian Telomere-associated Complex Formed by Multiple Telomeric Proteins. *J Biol Chem.* 279(49): 51338–42, doi: 10.1074/jbc.M409293200

[54] MARGULIES, Marcel, (...) a Jonathan M. ROTHBERG, 2005. Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature.* 437(7057): 376—380. doi: 10.1038/nature03959

[55] MAXAM, Allan M. a Walter GILBERT, 1977. A new method for sequencing DNA. *Proc. Nati. Acad. Sci. USA* 74(2): 560–564, doi: 10.1073/pnas.74.2.560, dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC392330/pdf/pnas00024-0174.pdf (odkaz funkční k 21. září 2019)

[56] DE MAAYER, Pieter, Angel VALVERDE a Donald COWAN, 2014. The Current State of Metagenomic Analysis, v knize *Genome Analysis: Current Procedures and Applications*: str. 197, ISBN: 978-1-912530-20-5.

[57] MASON, James M., Thomas A. RANDALL a Radmila Capkova FRYDRYCHOVA, 2015. Telomerase lost? *Chromosoma.* 125(1): 65-–73. doi: 10.1007/s00412-015-0528-7

[58] McCLINTOCK, Barbara, 1939. The Behavior in Successive Nuclear Divisions of a Chromosome Broken at Meiosis. *Proc. Natl. Acad. Sci. USA*: 405–416. doi:10.1073/pnas.25.8.405, dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1077932/ (odkaz funkční k 6. září 2019)

[59] McCLINTOCK, Barbara, 1941. The Stability of Broken Ends of Chromosomes in Zea Mays. *Genetics* 26(2), 234–282. Dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1209127/ (odkaz funkční k 6. září 2019)

[60] McKNIGHT, Thomas D. a Dorothy E. SHIPPEN, 2004. Plant Telomere Biology. *Plact cell* 16(4): 794-–803. Dostupné z: doi:10.1105/tpc.HistPersp

[61] MEYNE, Julianne, Robert L. Ratliff a Robert K. Moyzis, 1989. Conservation of the human telomere sequence (TTAGGG)n among vertebrates. *Proc Natl Acad Sci USA* 86(18): 7049-53. doi: 10.1073/pnas.86.18.7049, dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC297991/pdf/pnas00285-0228.pdf (odkaz funkční k 7. září 2019)

[62] MRAVINAC, Brankica, Nevenka MEŠTROVIĆ, Vladimir V. ČAVRAK a Miroslav PLOHL, 2011. TCAGG, an alternative telomeric sequence in insects. *Chromosoma* 120, 367–376. doi:10.1007/s00412-011-0317-x

[63] MULLER, Hermann Joseph, 1938. The Remaking of Chromosomes. *Collecting Net* 13: 181—198.

[64] NAKANO, Michihiko, Jun KOMATSU, Shun-ichi MATSUURA, Kazunori TAKASHIMA, Shinji KATSURA a Akira MIZUNO, 2003. Single-molecule PCR using water-in-oil emulsion. *Journal of Biotechnology* 102(2): 117–124

[65] NOVÁK Pavel, Pavel NEUMANN a Jiří MACAS, 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics.* 11: 378. doi: 10.1186/1471-2105-11-378

[66] NOVAK, P., P. NEUMANN, J. PECH, J. STEINHAISL a J. MACAS, 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. *Bioinformatics* 29: 792–793.

[67] NYRÉN, Pål, 1987. Enzymatic method for continuous monitoring of DNA polymerase activity. *Anal Biochem.* 167(2): 235–8. doi: 10.1016/0003-2697(87)90158-8

[68] NYRÉN, Pål a A. LUNDIN, 1985. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal Biochem.* 151(2): 504–9. doi: 10.1016/0003-2697(85)90211-8

[69] OKA, Y. S., S. SHIOTA, S. NAKAI, Y. NISHIDA a S. OKUBO, 1980. Inverted terminal repeat sequence in the macronuclear DNA of *Stylonychia pustulata*. *Gene* 10(4): 301–6. doi:10.1016/0378-1119(80)90150-x

[70] OKAZAKI, Satoshi, Kozo TSUCHIDA, Hideaki MAEKAWA, Hajima ISHIKAWA, Haruhiko FUJIWARA, 1993. Identification of a pentanucleotide telomeric sequence, (TTAGG)n, in the silkworm *Bombyx mori* and in other insects. *Mol Cell Biol.* 13(3): 1424–32, doi: 10.1128/mcb.13.3.1424, dostupné z https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/8441388/ (odkaz funkční k 7. září 2019)

[71] OLOVNIKOV, Alexej Matvejevič, 1973. A theory of marginotomy. The incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon. *J. Theor. Biol.* 41(1), 181–190. Dostupné z: https://www.researchgate.net/publication/293811225_Principle_of_marginotomy_in_the_synthesis_of_polynucleotides_at_a_template (odkaz funkční k 6. září 2019)

[72] PALM, Wilhelm a Titia DE LANGE, 2008. How Shelterin Protects Mammalian Telomeres. *Annu Rev Genet.* 42: 301–34. doi: 10.1146/annurev.genet.41.110306.130350.

[73] PARDUE Mary-Lou, P. G. DEBARYSHE, 2000. Drosophila telomere transposons: genetically active elements in heterochromatin. *Genetica.* 109(1-2): 45–52.

[74] PARDUE Mary-Lou, P. G. DEBARYSHE, 2008. Drosophila telomeres: A variation on the telomerase theme. *Fly (Austin).* 2(3): 101–10. doi: 10.4161/fly.6393

[75] PAREEK, Chandra Shekhar, Rafal SMOCZYNSKI a Andrzej TRETYN, 2011. Sequencing technologies and genome sequencing. *J Appl Genet.* 52(4): 413-–435. doi: 10.1007/s13353-011-0057-x

[76] PARK, Vicki Murtif, Karen M. GUSTASHAW a Theresa M. WATHEN, 1992. The Presence of Interstitial Telomeric Sequences in Constitutional Chromosome Abnormalities. *Am. J. Hum. Genet.* 50: 914–923.

[77] PEŠKA, Vratislav. Neobvyklé telomery. Brno, 2011. Disertační práce. Přírodovědecká fakulta Masarykovy univerzity. Vedoucí práce Eva SÝKOROVÁ.

[78] PEŠKA, Vratislav, 2017. Výlet na konec genomu 1. Jak se kopírují telomery. *Živa* 2: 53–57.

[79] PEŠKA, Vratislav, Eva SÝKOROVÁ, Jiří FAJKUS, 2008. Two faces of *Solanaceae* telomeres: a comparison between *Nicotiana* and *Cestrum* telomeres and telomere-binding proteins. *Cytogenet Genome Res* 122(3–4): 380-–387, doi:10.1159/000167826000167826

[80] PEŠKA, Vratislav, Petr FAJKUS, Miloslava FOJTOVÁ, Martina DVOŘÁČKOVÁ, Jan HAPALA, Vojtěch DVOŘÁČEK, Pavla POLANSKÁ, Andrew R. LEITCH, Eva SÝKOROVÁ, Jiří FAJKUS, 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(4): 644–54. doi: 10.1111/tpj.12839, dostupné z: https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.12839 (odkaz funkční k 7. září 2019)

[81] PODLEVSKY, Joshua D., Christopher J. BLEY, Rebecca V. OMANA, Xiaodong QI, Julian J. L. CHEN, 2007. The Telomerase Database. *Nucleic Acids Res.* 36, D339–343, doi: 10.1093/nar/gkm700, článek dostupný z. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2238860/, databáze dostupná z: http://telomerase.asu.edu (odkazy funkční k 7. září 2019)

[82] PORRECA, Gregory J., Jay SHENDURE a George M. CHURCH, 2006. Polony DNA sequencing. *Curr Protoc Mol Biol.* kapitola 7: část 7.8. doi: 10.1002/0471142727.mb0708s76.

[83] PORRO, Antonio, Sascha FEUERHAHN, Julien DELAFONTAINE, Harold RIETHMAN, Jacques ROUGEMONT a Joachim LINGNER, 2014. Functional characterization of the TERRA transcriptome at damaged telomeres. *Nat Commun.* 5: 5379. doi: 10.1038/ncomms6379

[84] RICHARDS, E. J., F. M. AUSUBEL, 1988. Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. *Cell.* 53(1): 127–36. doi: 10.1016/0092-8674(88)90494-1

[85] RONAGHI, Mostafa, Mathias UHLEN a Pål NYRÉN, 1998. A sequencing method based on real-time pyrophosphate. *Science.* 281(5375): 363–365. doi: 10.1126/science.281.5375.363

[86] ROTHBERG, Jonathan M., Wolfgang HINZ, (...) a James BUSTILLO, 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature.* 475(7356): 348–52. doi: 10.1038/nature10242.

[87] Roumelioti, Fani-Marlen, Sotirios K. SOTIRIOU, Vasiliki KATSINI, Maria CHIOUREA, Thanos D. HALAZONETIS a Sarantis GAGOS, 2016. Alternative lengthening of human telomeres is a conservative DNA replication process

with features of break-induced replication. *EMBO Rep.* 17(12): 1731-–1737. doi: 10.15252/embr.201643169

[88] SAHARA, K., F. MAREC, W. TRAUT, 1999. TTAGG telomeric repeats in chromosomes of some insects and other arthropods. *Chromosome Res.* 7(6): 449–60.

[89] SANDELL, L. L. a V. A. ZAKIAN, 1993. Loss of a yeast telomere: arrest, recovery, and chromosome loss. *Cell.* 75(4): 729–39, doi: 10.1016/0092-8674(93)90493-a

[90] SANGER, F., S. NICKLEN a R. COULSON, 1977. *Proc. Nati. Acad. Sci. USA* 74(12): 5463–5467, doi: 10.1073/pnas.74.12.5463

[91] SHENDURE, Jay, Gregory J. PORRECA, Nikos B. REPPAS, Xiaoxia LIN, John P. McCUTCHEON, Abraham M. ROSENBAUM, Michal D. WANG, Kun ZHANG, Robi D. MITRA a George M. CHURCH, 2005. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* 309(5741): 1728–1732.

[92] SHENDURE, Jay a Hanlee JI, 2008. Next-generation DNA sequencing. *Nature Biotechnology* 26(10): 1135–1145, doi: 10.1038/nbt1486

[93] SHOKRALLA, Shadi, Jennifer L. SPALL, Joel F. GIBSON a Mehrdad HAJIBABAEI, 2012. Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology* 21(8): 1794–805. doi: 10.1111/j.1365-294X.2012.05538.x, dostupné z: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-294X.2012.05538.x# (odkaz funkční k 28. září 2019)

[94] SCHALCHIAN-TABRIZI, Kamran, Marianne A. MINGE, Mari ESPELUND, Russell ORR, Torgeir RUDEN, Kjetill S. JAKOBSEN a Thomas CAVALIER-SMITH, 2008. *PLoS One.* 3(5): e2098. doi: 10.1371/journal.pone.0002098, dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2346548/ (odkaz funkční k 7. září 2019)

[95] SCHECHTMAN, M. G., 1990. Characterization of telomere DNA from Neurospora crassa. *Gene.* 88(2): 159–65. doi: 10.1016/0378-1119(90)90027-o

[96] SCHEIBE, Marion, Nausica ARNOULT, DennisKAPPEI, Frank BUCHHOLZ, Anabelle DECOTTIGNIES, Falk BUTTER a Matthias MANN, 2013. Quantitative interaction screen of telomeric repeat-containing RNA reveals novel TERRA regulators. *Genome Res.* 23(12): 2149–2157. doi: 10.1101/gr.151878.112

[97] SCHMUTZ, Isabelle a Titia DE LANGE, 2016. Shelterin. *Current Biology* 26(10): R397-9. doi: 10.1016/j.cub.2016.01.056

[98] SFEIR, Agnel a Titia DE LANGE, 2012. Removal of shelterin reveals the telomere end-protection problem. *Science.* (New York, N.Y.) vol. 336, 6081: 593–7. Dostupné z: doi:10.1126/science.1218498

[99] De SILANES, Isabel López, Osvaldo GRANA, Maria Luigia De BONIS, Orlando DOMINGUEZ, David G. PISANO a Maria A. BLASCO, 2014. Identification of TERRA locus unveils a telomere protection role through association to nearly all chromosomes. *Nat Commun.* 5: 4723. doi: 10.1038/ncomms5723

[100] SITOVÁ, Zdeňka. Tandem Repeats Merger. Repozitář GitHub: https://github.com/zdenkas/tandem-repeats-merger

[101] STADEN, R., 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* 6(7): 2601—2610. doi: 10.1093/nar/6.7.2601

[102] VAN STEENSEL, Bas, Agata SMOGORZEWSKA a Titia DE LANGE, 1998. TRF2 protects human telomeres from end-to-end fusions. *Cell.* 92(3): 401–13, doi: 10.1016/s0092-8674(00)80932-0

[103] TURCATTI, Gerardo, Anthony ROMIEU, Milan FEDURCO a Ana-Paula TAIRI, 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.* 36(4): e25, doi: 10.1093/nar/gkn021

[104] UDROIU, Ion a Antonella SGURA, 2020. Alternative Lengthening of Telomeres and Chromatin Status. *Genes* 11(1): 45. doi:10.3390/genes11010045

[105] VICTORELLI, Stella, João F. PASSOS, 2017. Telomeres and Cell Senescence - Size Matters Not. *EBioMedicine.* 21: 14–20. doi: 10.1016/j.ebiom.2017.03.027

[106] VÍTKOVÁ, Magda, Jiří KRÁL, Walther TRAUT, Jan ZRZAVÝ, František MAREC, 2005. The evolutionary origin of insect telomeric repeats, (TTAGG)n. *Chromosome Research* 13: 145–156, doi: 10.1007/s10577-005-7721-0

[107] WATSON, Jim, 1972. Origin of concatemeric T7 DNA. *Nat. New Biol.* 239(94), 197–201. Dostupné z: https://wellcomelibrary.org/item/b19846927#?c=0&m=0&s=0&cv=0&z=0.0339%2C0.1336%2C0.9511%2C0.6418 (odkaz funkční k 6. září 2019)

[108] WEBB, Christopher J. a Virginia A. ZAKIAN, 2016. Telomerase RNA is more than a DNA template. *RNA Biol.* 13(8): 683-–689. doi: 10.1080/15476286.2016.1191725

[109] WEINERT, T. A. a L. H. HARTWELL, 1988. The RAD9 gene controls the cell cycle response to DNA damage in Saccharomyces cerevisiae. *Science* 241: 317–322, doi: 10.1126/science.3291120

[110] WEISS, H. a H. SCHERTHAN, 2002. *Aloe spp.* – plants with vertebrate-like telomeric sequences. *Chromosome Res.*10(2): 155–64.

[111] WYATT, Haley D. M., Stephen C. WEST a Tara L. BEATTIE, 2010. InTERTpreting telomerase structure and function. *Nucleic Acids Res.*38(17): 5609—5622. doi: 10.1093/nar/gkq370

[112] XIN, Huawei, Dan LIU a Zhou SONGYANG, 2008. The telosome/shelterin complex and its functions. *Genome Biol.* 9(9): 232, doi: 10.1186/gb-2008-9-9-232, dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2592706/ (odkaz funkční k 7. září 2019)

[113] XIONG, Jin, 2006. Essential Bioinformatics. Cambridge University Press, New York.

[114] YEHEZKEL, Shiran, Yardena SEGEV, Evani VIEGAS-PÉQUIGNOT, Karl SKORECKI a Sara SELIG, 2008. Hypomethylation of subtelomeric regions in ICF syndrome is associated with abnormally short telomeres and enhanced transcription from telomeric regions, 2008. *Hum. Mol. Genet.* 17: 2776-–2789. doi: 10.1093/hmg/ddn177

[115] ZGLINICKI, Thomas von, 2002. Oxidative Stress Shortens Telomeres. *Trends Biochem Sci.* 27(7): 339–44.

[116] ZHANG, Qi, Nak-Kyoon KIM a Juli FEIGON, 2011. Architecture of human telomerase RNA. *Proc Natl Acad Sci* 108(51): 20325-–20332. doi: 10.1073/pnas.1100279108

[117] ZHAO Shuang, Feng WANG a Lin LIU, 2019. Alternative Lengthening of Telomeres (ALT) in Tumors and Pluripotent Stem Cells. *Genes (Basel).* 10(12): 1030. doi: 10.3390/genes10121030

[118] ZVEREVA, M. I., D. M. SHCHERBAKOVA a O. A. DONTSOVA, 2010. Telomerase: Structure, Functions, and Activity Regulation. *Biochemistry (Mosc)* 75(13): 1563–83 doi: 10.1134/S0006297910130055.

[119] CESNET, zájmové sdružení právnických osob. https://wiki.metacentrum.cz/wiki/

[120] Vývojový tým softwaru SRA Toolkit. Soubor nástrojů a knihoven pro data ze SRA. http://ncbi.github.io/sra-tools/

[121] Wikimedia Commons, 2008. Schematics of the arrangement of proteins in the telosome/shelterin complex. Dostupné z: https://commons.wikimedia.org/wiki/File:Telosome.png (odkaz funkční k 21. září 2019)

[122] https://www.ncbi.nlm.nih.gov/Taxonomy