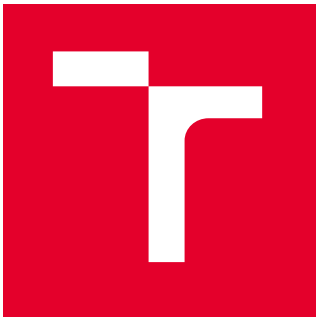
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

DIPLOMOVÁ PRÁCE

Brno, 2023 Bc. Barbora Pomykalová



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING

NÁSTROJ PRO PREDIKCI SMALL RNA V RNA-SEQ DATECH

A TOOL FOR PREDICTION OF SMALL RNA IN RNA-SEQ DATA

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Barbora Pomykalová**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Kateřina Jurečková**
SUPERVISOR

BRNO 2023

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní program **Biomedicínské inženýrství a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Bc. Barbora Pomykalová ***ID:*** 211209

Ročník: 2 ***Akademický rok:*** 2022/23

NÁZEV TÉMATU:

Nástroj pro predikci small RNA v RNA-Seq datech

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma metody a výpočetní nástroje pro identifikaci small RNA u bakterií. 2) Na základě volně dostupných RNA-Seq dat vytvořte vhodný dataset pro predikci small RNA u vybraného bakteriálního genomu. 3) Proveďte predikci small RNA na vytvořeném datasetu pomocí dostupných nástrojů a výsledky srovnejte. 4) Navrhněte vlastní metodu pro predikci small RNA a implementujte ji v libovolném programovacím jazyce. 5) Algoritmus otestujte na vytvořeném datasetu. 6) Proveďte vyhodnocení a diskutujte výsledky.

DOPORUČENÁ LITERATURA:

[1] LEONARD, Simon, Sam MEYER, Stephan LACOUR, William NASSER, Florence HOMMAIS and Sylvie REVERCHON. APERO: a genome-wide approach for identifying bacterial small RNAs from RNA-Seq data. Nucleic Acids Research. 2019, 47(15), e88–e88. ISSN 0305-1048. doi:10.1093/nar/gkz485

[2] STIENS, Jennifer, Kristine B. ARNVIG, Sharon L. KENDALL and Irene NOBELI. Challenges in defining the functional, noncoding, expressed genome of members of the Mycobacterium tuberculosis complex. Molecular Microbiology. 2022, 117(1), 20–31. ISSN 0950-382X. doi:10.1111/mmi.14862

Termín zadání: 6.2.2023 ***Termín odevzdání:*** 22.5.2023

Vedoucí práce: Ing. Kateřina Jurečková

prof. Ing. Valentine Provazník, Ph.D.
předseda rady studijního programu

Abstrakt

Tato diplomová práce se zaměřuje na problematiku detekce small RNA (sRNA) v bakteriálním genomu. Small RNA jsou krátké nekódující transkripty, které hrají klíčovou roli v genové expresi. K dnešnímu dni existuje několik algoritmů zaměřujících se na detekci sRNA z RNA-Sequencing (RNA-Seq) dat, jež mohou být získána z některých sekvenačních platforem. Nejčastěji jsou používány platformy Illumina či Ion Torrent, spadající do nové generace sekvenování, a PacBio s Oxford Nanopore patřící do třetí generace sekvenování.

V této práci byly popsány principy detekce sRNA u volně dostupných nástrojů a následně byl navržen vlastní nástroj pro detekci sRNA – nástroj SEARCHsRNA. Dva z volně dostupných nástrojů – Rockhopper a DETR'PROK, společně s nástrojem SEARCHsRNA, byly otestovány na datech RNA-Seq pro bakterii *Vibrio atlanticus* LGP32.

Klíčová slova

small RNA, RNA-Seq, nekódující RNA, genová exprese, regulace, sekvenační platformy, predikce sRNA

Abstract

This diploma thesis focuses on the detection of small RNA (sRNA) in the bacterial genome. sRNAs are short non-coding transcripts that play a key role in gene expression. To date, there are several algorithms focusing on the detection of sRNAs from RNA-Sequencing (RNA-Seq) data that can be obtained by some of the sequencing platforms. The most frequently used platforms are Illumina and Ion Torrent belonging to the next generation sequencing and PacBio with Oxford Nanopore belonging to the third generation of sequencing.

In this work, the workflow of sRNA detection using freely available tools was described and then an own unique tool for sRNA detection – the SEARCHsRNA tool – was designed. Two open-source software tools – Rockhopper and DETR'PROK, together with newly created tool, were tested on RNA-Seq data for bacteria *Vibrio atlanticus* LGP32.

Keywords

small RNA, RNA-Seq, non-coding RNA, gene expression, regulation, sequencing platforms, prediction of sRNA

Bibliografická citace

POMYKALOVÁ, Barbora. Nástroj pro predikci small RNA v RNA-Seq datech. Brno, 2023. Dostupné také z: https://www.vut.cz/studenti/zav-prace/detail/150860. Diplomová práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce Kateřina Jurečková.

Prohlášení autora o původnosti díla

Jméno a příjmení studenta:	Barbora Pomykalová
VUT ID studenta:	211209
Typ práce:	Diplomová práce
Akademický rok:	2022/23
Téma závěrečné práce:	Nástroj pro predikci small RNA v RNA-Seq datech

Prohlašuji, že svou diplomovou práci jsem vypracovala samostatně pod vedením vedoucí závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

.

V Brně dne: ----------

podpis autora

-
Děkuji vedoucí diplomové práce Ing. Kateřině Jurečkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce.

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic.
-

V Brně dne: ----------

podpis autora

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Přestože jsou small RNA (sRNA) známé již od počátků 70. let 20. století [1], nebyly brány v potaz, dokud nebyla objevena v roce 1983 první potenciální sRNA u *Escherichia coli* a nebyla prokázána její důležitá role v regulaci genové exprese u OmpF proteinu [1], který umožňuje pasivní přenos malých hydrofobních molekul přes membránu [2]. Nalezení a pochopení všech funkčních vlastností sRNA u bakterií může mít velký vliv na průmyslovou a ekologickou sféru (biopaliva) a ve zdravotnictví (terapie – vývoj specifických antibakteriálních léků) [3].

V posledním dvacetiletí se zájem o detekci všech nekódujících RNA (ncRNA), včetně sRNA, rapidně zvýšil, vzhledem k možnosti využití nových sekvenačních platforem poskytujících transkriptomická data s vysokým dynamickým rozsahem a citlivostí [4]. To předchozími metodami možné nebylo. Tento způsob zisku transkriptomických dat rozličnými sekvenačními platformami se souhrnně nazývá RNA-Sequencing (RNA-Seq).

Nejčastěji se pro přípravu RNA-Seq dat využívají platformy sekvenování nové generace (NGS). Mezi tyto platformy patří například celosvětově používaná Illumina [5] a Ion Torrent [6]. Právě sekvenační platforma Illumina poskytuje také přesný postup se speciálním kitem na přípravu dat pro sekvenování malých ncRNA [7].

V posledních letech se začínají využívat i sekvenační platformy třetí generace (TGS), které přináší výhodu sekvenování celých molekul bez fragmentace. To umožňuje kvalitnější a přesnější anotace genomů, které jsou pro predikci ncRNA zásadní.

Samotný zisk kvalitních transkriptomických dat však není jediný problém, který se při detekci sRNA vyskytuje. Po obdržení těchto dat je nutná jejich úprava a analýza tak, aby bylo možné obdržet informace o sRNA. K tomu slouží nástroje, které umí s RNA-Seq daty pracovat a následně vyhodnotit potenciální pozice nových sRNA. Je velmi důležité zmínit, že se jedná pouze o potencionální sRNA, jejichž skutečná přítomnost se musí následně dokázat laboratorními metodami.

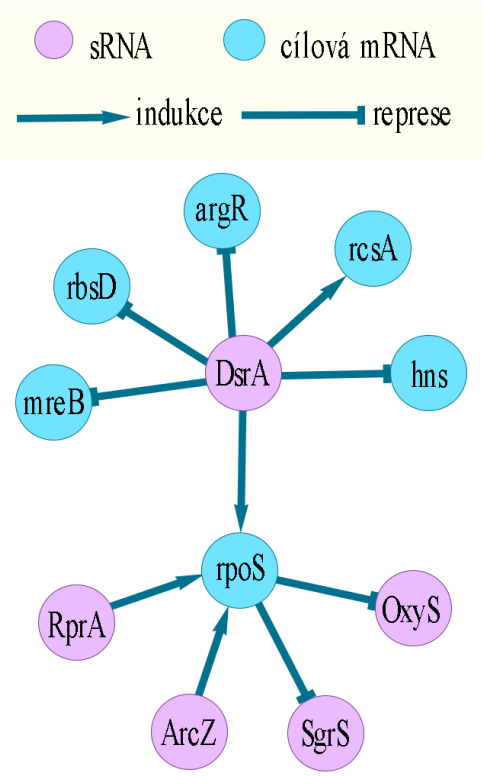
Během posledních 15 let se objevilo několik nástrojů pro predikci sRNA z bakteriálních RNA-Seq dat. Některé však již nefungují nebo nejsou dohledatelné. Mezi stále funkční nástroje patří nástroj Rockhopper [8; 9], DETR'PROK [10], ANNOgesic [11], APERO [12] a Baerhunter [13]. Přestože jsou tyto nástroje volně dostupné a funkční, u některých není možnost je využít na všechna RNA-Seq data, nebo nejsou příliš uživatelsky přívětivé a jejich spuštění je komplikované.

Tato diplomová práce má za úkol popsat principy těchto volně dostupných nástrojů a následně navrhnout a sestavit nový nástroj určený pro predikci sRNA z RNA-Seq dat. Nově vytvořený nástroj – SEARCHsRNA – následně otestovat na vytvořeném datasetu RNA-Seq dat pro bakterii *Vibrio atlanticus* LGP32 a získané výsledky porovnat s výsledky obdrženými z volně dostupných nástrojů – Rockhopper a DETR'PROK.

1. SMALL RNA

Malé nekódující RNA, neboli small RNA (sRNA), jsou krátké ribonukleové kyseliny (RNA), jejichž průměrná délka se pohybuje v rozmezí od 50 do 250 páru bází (pb) [14; 15], ovšem některé literární zdroje uvádí jejich délku až tisíc pb [16]. Je předpokladem, že každá bakterie obsahuje ve svém genomu až stovky těchto RNA [17]. Aktivně se podílejí na regulaci genové exprese [18], například sRNA MicA u bakterie *E. coli* konkrétně ovlivňuje expresi externího proteinu membrány OmpA [19], čímž je schopna ovlivnit patogenitu bakterie [20]. Zmíněná regulace může obecně probíhat na různých úrovních genové exprese – transkripce nebo translace, a také v různých fázích – inicializace, elongace či terminace [3]. Dále může být regulace u některých sRNA podmíněna vnějšími podmínkami, jako je změna prostředí nebo stresové situace [21].

Účinky sRNA na regulaci genové exprese se rozlišují na primární a sekundární. Mezi primární účinky spadají vazby sRNA, které přímo interagují s cílovou mediátorovou RNA (mRNA), resp. s kódující oblastí (CDS) či proteinem (primární cíl) za účelem změny jejich struktury, funkce nebo translace. Small RNA se však mohou vázat i na mRNA genů, které jsou transkripčními faktory pro geny jiné. Tudíž nepřímo ovlivňují expresi i těchto genů (sekundární cíl), což nazýváme sekundární regulací. Kombinací těchto funkčních vlastností sRNA a transkripčních faktorů vznikají velké komplexní regulační sítě, které ovlivňují celkové chování bakterie v daném momentu a daném prostředí [14; 16]. Ukázka dílčí regulační sítě je na Obrázku 1.1.



Obrázek 1.1 Ukázka dílčí regulační sítě pro *E. coli* K12 MG 1655; upraveno [22].

1.1 *cis*- a *trans*- sRNA

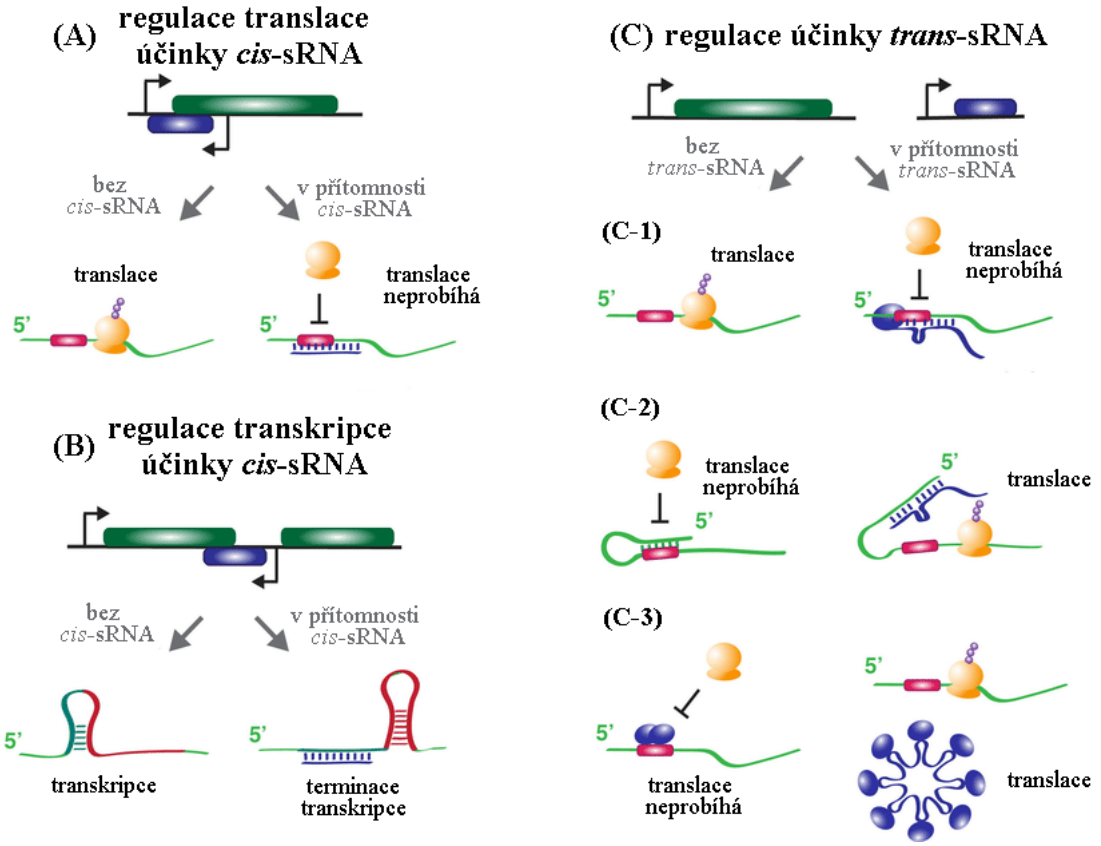
Small RNA je možné rozdělit na dva základní typy: *cis*-encoded small RNA (*cis*-sRNA) a *trans*-encoded small RNA (*trans*-sRNA). *Cis*-encoded sRNA [3], v některých případech nazývány jako antisense RNA (asRNA) [16], jsou v naprosté většině případů dokonale komplementární s cílovou mRNA. Jejich velikost se pohybuje v řádu desítek až tisíců pb. V návaznosti na jejich délku mohou pokrývat nejen celou délku cílového genu, ale také jeho 3' nepřekládané oblasti (3' untranslated region, 3'UTR) a 5' nepřekládané oblasti (5' untranslated region, 5'UTR) [16]. 3'UTR a 5'UTR jsou oblasti, které mohou ohraničovat gen, ale nepodléhají translaci. Příkladem sRNA s těmito vlastnostmi je sRNA AmgR u bakterie *Salmonella enterica,* která dosahuje délky 1200 pb a ovlivňuje genovou expresi proteinu MgtC [23]. *Cis*-encoded sRNA se na chromozomu vyskytují na negativním vlákně ve stejném lokusu jako cílová mRNA [24] a mají schopnost regulovat tuto mRNA primárně i sekundárně.

Na Obrázku 1.2 je vyobrazeno, jakým způsobem mohou působit *cis*-sRNA na své cíle a jakým způsobem je tedy schopna ovlivňovat jejich genovou expresi [3]. V sekci (A) je znázorněno, jaké účinky může mít *cis*-sRNA (znázorněna modře) regulující translaci cílové mRNA (znázorněno zeleně). Small RNA je komplementárně navázána na vazebné místo genu (znázorněno růžově), na kterém ve standardním případě nasedá ribozom (znázorněno žlutě). Vazba sRNA s vazebným místem pro ribozom má za následek, že nemůže dojít k navázání ribozomu na mRNA a tedy nedochází k translaci, která je pro genovou expresi klíčová. V některých případech dochází i k degradaci dané mRNA. V sekci (B) je možné pozorovat, jaké účinky může mít *cis*-sRNA na transkripci genu. Díky přítomnosti sRNA dochází ke změně sekundární struktury cílové mRNA. Ta v její přítomnosti formuje vlásenkovou smyčku, která způsobí, že dochází k předčasné terminaci probíhající transkripce.

Trans-encoded sRNA [3] jsou oproti *cis*-encoded sRNA výrazně kratší a dosahují délky okolo 100 pb. Také nejsou na cílové mRNA navázány zcela komplementárně, neboťu nich dochází ke vzniku mnoha elementů sekundární struktury, zejména pak smyček. Vzniklé elementy se mohou měnit v čase v návaznosti na podmínky prostředí, ve kterém se nachází (změna pH, koncentrace živin či teploty), a mohou tedy působit na více cílových mRNA. Oblast, ve které jsou *trans*-encoded sRNA komplementární s cílovou mRNA, je nazývána seed a dosahuje délky okolo 6 až 8 po sobě jdoucích pb. Mnoho z doposud detekovaných *trans*-sRNA vyžaduje k vazbě na cílovou mRNA chaperonový protein Hfq [16].

Na Obrázku 1.2 (C) lze pozorovat účinky *trans*-sRNA (znázorněna modře) na cílovou mRNA (znázorněna zeleně). V sekci (C-1) lze pozorovat, že pokud se *trans*-sRNA naváže na místo vazby ribozomu, nedojde k translaci tohoto genu. K této vazbě dochází za přítomnosti již zmíněného Hfq chaperonu (znázorněn jako modrý ová). V sekci (C-2) je znázorněno, že *trans*-sRNA může také uvolnit vazebné místo pro ribozom a naopak translaci umožnit. V poslední sekci (C-3) nedochází k translaci,

protože jsou na vazebném místě pro ribozom přítomny RNA-binding proteiny (RBP, znázorněny jako modrý ovál). Ovšem za přítomnosti *trans*-sRNA jsou RBP uvolněny, čímž je průběh translace umožněn. V tomto případě má *trans*-sRNA více částečně komplementárních míst pro RBP. Může tedy zároveň navázat několik RBP a tím uvolnit více vazebných míst pro ribozom.



Obrázek 1.2 Ukázka regulačních účinků sRNA: (A),(B) pro *cis*-sRNA; (C-1,2,3) pro *trans*-sRNA; upraveno [3].

2. RNA-SEQUENCING

K nalezení nekódujících malých RNA v genomu bakterií se využívají některé laboratorní metody, které umožňují sledovat genovou expresi. Mezi nejstarší, finančně nejdostupnější a hojně využívanou metodu, která se používá pro detekci či validaci sRNA, patří metoda Northern blot [25]. Metoda umožňuje kvantifikovat množství určitého transkriptu RNA z genomu na základě jeho známé délky [26]. Mezi další často využívané postupy pro analýzu genové exprese patří metoda sériové analýzy genové exprese (SAGE) [27; 28], DNA microarray [28], reversně transkripční kvantitativní polymerázová řetězová reakce (RT-qPCR) [29] a dropletová digitální PCR (ddPCR) [30].

Přestože pomocí výše zmíněných metod je možné stanovit úroveň genové exprese pro vybrané transkripty, analýza celého transkriptomu možná není. Té bylo dosaženo až metodou RNA-Sequencing (RNA-Seq) využívající sekvenování nové generace pro analýzu většiny přítomné RNA. Z tohoto důvodu je metoda RNA-Seq nazývána jako catch-all metoda [31].

RNA-Seq se také vyznačuje vysokým dynamickým rozsahem pro měření variability úrovně exprese jednotlivých transkriptů [4]. Vysokého dynamického rozsahu je docíleno obdržením obrovského množství čtení během sekvenace [32].

V posledních letech jde také do popředí využití sekvenování třetí generace (TGS) [33], které přináší možnost sekvenování celé molekuly bez fragmentací. Díky těmto metodám je sestavení genomu či případná anotace genů podstatně zjednodušena.

2.1 Obecný pracovní postup přípravy knihovny pro RNA-Seq

V dnešní době existuje více než sto různých protokolů pro přípravu knihovny RNA-Seq [34]. Protokoly se od sebe liší především použitím sekvenačních platforem NGS, které vyžadují rozdílné parametry knihoven RNA-Seq.

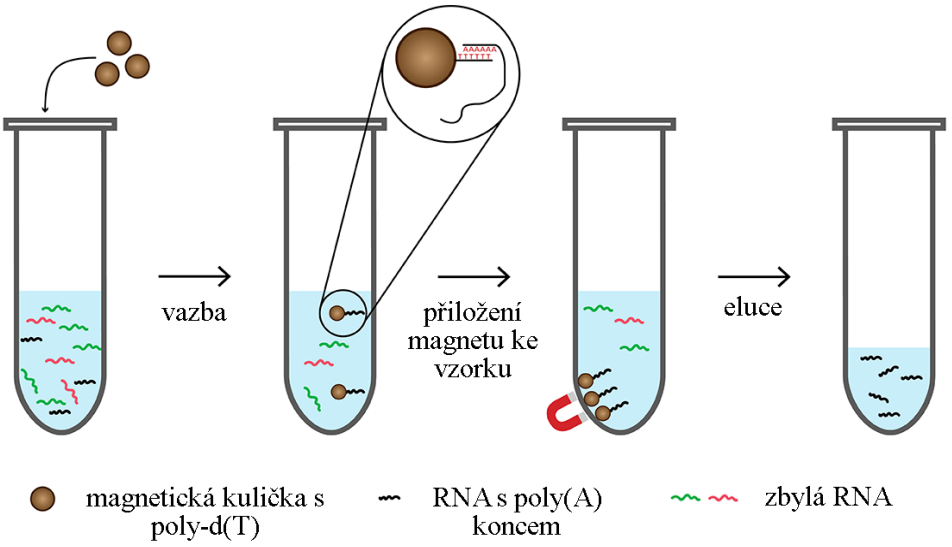
Prvním krokem pro vytvoření finální knihovny RNA-Seq je příprava vzorku obsahující pouze RNA, která má být následně analyzována. Způsob izolace RNA záleží na tom, jestli má být izolována celková RNA (obsahující mediátorovou RNA, ribozomální RNA (rRNA), transferovou RNA (tRNA) a sRNA), nebo pouze její část. Například pro analýzu genové exprese není žádoucí, aby vzorek obsahoval rRNA, která může zastupovat až 85 % z celkové RNA [35]. Naopak je důležitá analýza mRNA, která je ale obsažena ve vzorku pouze ve zhruba pěti procentech celkové RNA [36].

V případě, že je RNA získávána z tkáně, musí být tkáň nejprve mechanicky narušena, např. homogenizátorem. Poté je ke vzorku přidán lyzační roztok, který způsobuje buněčnou lýzi a rozpad proteinů. Dále jsou ke vzorku přidány inhibitory ribonukleáz (např. guanidin isothiokyanát – GITC), aby nedocházelo k degradaci RNA

ribonukleázy (RNázy). RNázy jsou enzymy, které umožňují hydrolyticky štěpit fosfodiesterové vazby RNA [37]. Pro odstranění kontaminujících zbytků buněčných komponent a deoxyribonukleové kyseliny (DNA) se dříve hojně využívala metoda pracující se směsí fenol-chloroformu, která byla ke vzorku přidána. Vzorek byl následně centrifugován a díky tomu došlo k oddělení RNA od zbylých komponent a DNA. Takto je tedy možné izolovat celkovou RNA [38]. Ovšem v poslední době se spíše přistupuje k metodám využívajícím specifické kity pro efektivní a kvalitní izolaci RNA z buněk [39].

Postup extrakce čisté mRNA ze vzorku je rozdílný pro extrakci u eukaryotních a prokaryotních organismů. U obou postupů však vycházíme ze vzorku očištěného od DNA a zbytků proteinů.

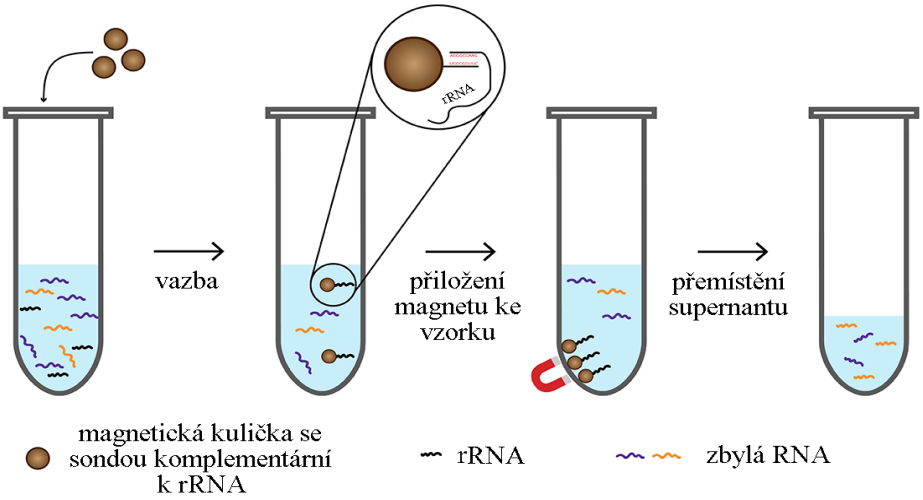
Pro eukaryotické buňky, jejichž mRNA velice často obsahuje poly(A) konec, je možné extrahovat mRNA ze vzorku pomocí magnetických kuliček obalených molekulami poly-d(T). Ty způsobí, že dojde ke komplementárnímu navázaní poly(A) konců mRNA na zmíněné kuličky. Tato metoda je vhodná pouze pro analýzu mRNA s poly(A) koncem při dostatečném množství původního vzorku [40]. Schéma extrakce mRNA s poly(A) koncem je vyobrazeno na Obrázku 2.1.



Obrázek 2.1 Metoda pro extrakci RNA s poly(A) koncem ze vzorku; upraveno [5].

U prokaryot a eukaryot, které neobsahují mRNA s poly(A) konci, se musí pro extrakci mRNA vyžít jiné postupy. Jelikož tyto mRNA nemají stejnou ani obdobnou významnou vlastnost, která by umožnila jednoduchou extrakci, musí být ze vzorku odstraněna ta RNA, která pro následnou analýzu není potřebná. Mezi tyto RNA patří rRNA a tRNA. Nejčastěji dochází k odstranění rRNA, vzhledem k jeho vysokému zastoupení ve vzorku. Odstranění rRNA je prováděno deplecí [35], což umožňuje analyzovat zbylý transkriptom obsahující kódující mRNA, ale i sRNA [34]. Obrázku 2.2 zobrazuje schéma extrakce celé RNA mimo rRNA.

Poté následuje fragmentace extrahovaných RNA. Fragmentace se liší podle zvolené sekvenační platformy. Sekvenační platformy NGS vyžadují krátké fragmenty (Illumina [41], Ion Torrent [42]), naopak sekvenační platformy ze sekvenování TGS využívají schopnosti sekvenace celých molekul komplementární DNA (cDNA) [34]. Umožňují tedy sekvenaci bez fragmentace (Pacific Biosciences [43], Oxford Nanopore 0).



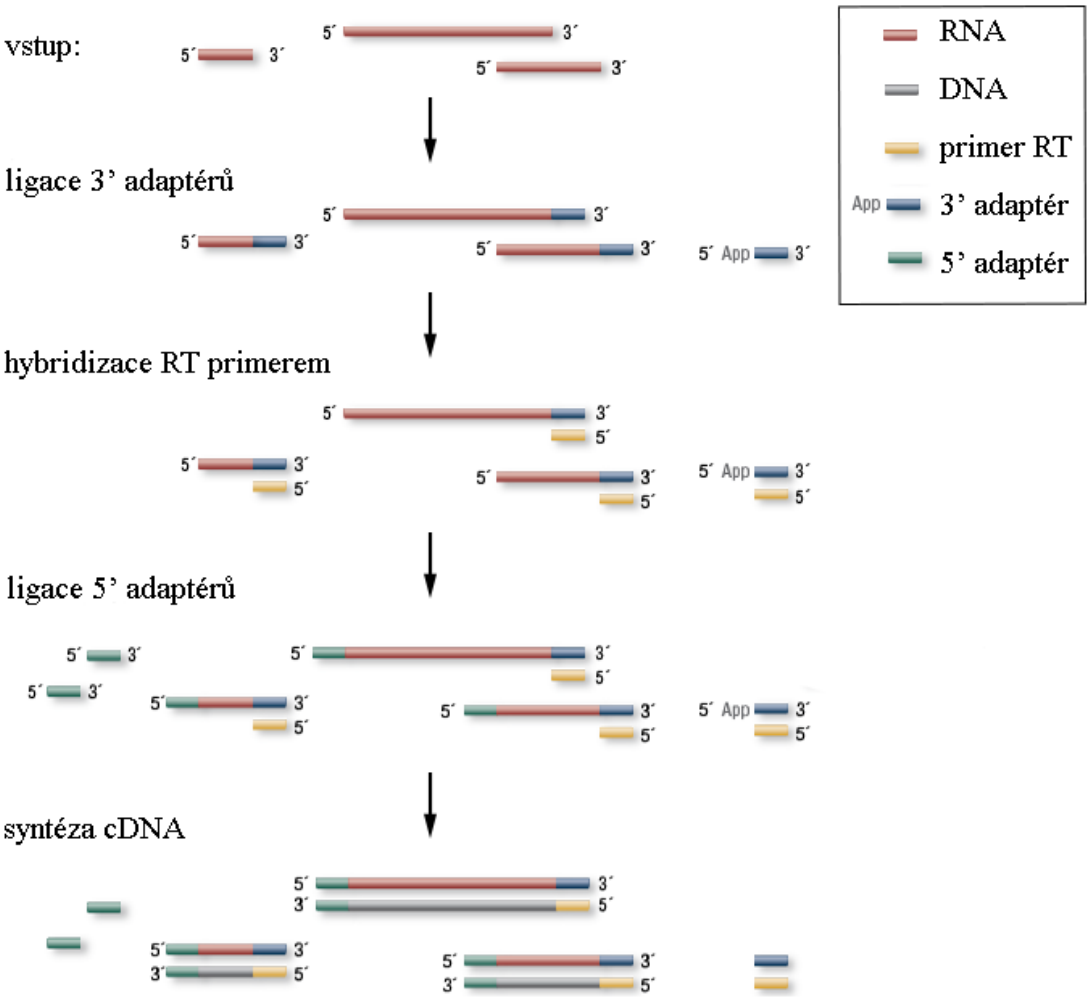
Obrázek 2.2 Metoda pro extrakci celé RNA kromě rRNA; upraveno [5].

Samotná fragmentace může probíhat buďza pomocí RNáz, které štěpí RNA na přesném místě, nebo za užití alkalických sloučenin [40]. Fragmentace se nejčastěji provádí za zvýšené teploty – okolo 70 °C. Zvýšená teplota je nutná, aby neprobíhala formace sekundárních struktur RNA. K fragmentaci může docházet také až po přepisu RNA do dvouvláknové cDNA (dsDNA) pomocí reverzní transkriptázy (RT). Fragmentace cDNA je pak prováděna pomocí DNA ribonukleáz (DNáz).

Následně jsou na cDNA požadované délky ligovány adaptéry [40]. Adaptéry jsou krátké, synteticky vytvořené oligonukleotidy, které slouží pro navázání cDNA na místo, kde probíhá samotná sekvenace, a slouží jako primery pro polymerázy, které se využívají pro amplifikaci a sekvenování [5; 45]. Existuje několik způsobů ligování adaptérů k cDNA. Nejjednodušší metoda je ligace náhodných hexamerních primerů k fragmentům RNA, ještě před přidáním RT, nebo přímo k fragmentům cDNA. Díky tomuto kroku je ztracena informace o tom, z jakého vlákna daný fragment pochází (non-strand-specific viz kapitola 2.3). Z tohoto důvodu není metoda příliš výhodná a používá se jen velmi zřídka [40].

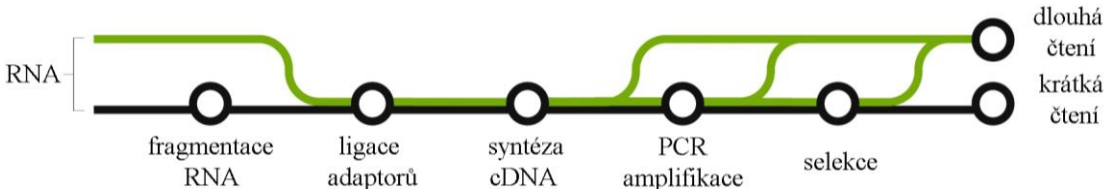
Jestliže je požadována informace o tom, z kterého vlákna výsledné čtení pochází (strand-specific), musí se ligace adaptéru provést tak, že se adaptéry navážou na jednovláknovou RNA (resp. cDNA). Jeden adaptér se naváže na 3‘ konec RNA, kde dochází k hybridizaci adaptéru primerem pro reverzní transkripci. A na 5‘ konec RNA se naváže druhý adaptér. Následnou syntézou pomocí RT získáme cDNA, která obsahuje adaptéry nutné pro sestavení strand-specific knihovny. Schéma výše

uvedeného postupu přípravy cDNA se strand-specific adaptéry je vyobrazeno na Obrázku 2.3.



Obrázek 2.3 Schéma znázorňující přípravu strand-specific knihovny; upraveno [46].

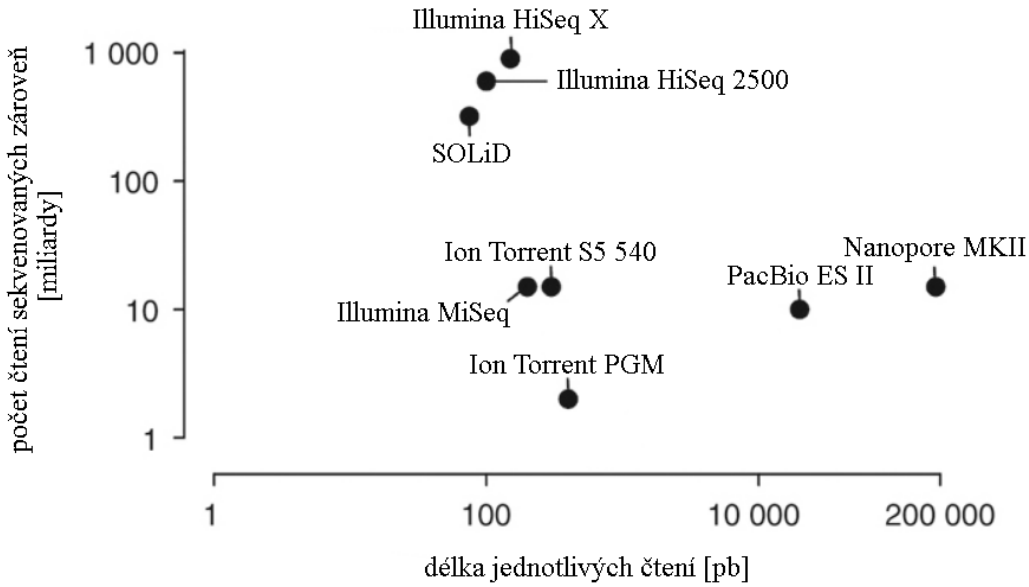
V této fázi jsou cDNA připravena pro amplifikaci, která navýší množství cDNA určené pro sekvenaci [40]. Nejčastěji se využívá emulzní a můstková PCR [5]. U PCR může dojít k nerovnoměrné amplifikaci tím, že se u některých cDNA amplifikace ustálí dříve než u jiných. Z tohoto důvodu musí být provedeny úpravy, jako například využití molekulárních značek před samotnou PCR, které tuto nerovnoměrnost odstraní [40]. Na závěr dochází k odstranění nevyhovujících cDNA (krátkých či dlouhých) a zbylá cDNA jsou připravena pro sekvenaci. Zjednodušené schéma celé přípravy knihovny cDNA určené pro sekvenaci je vyobrazeno na Obrázku 2.4.



Obrázek 2.4 Schéma přípravy knihovny cDNA pro sekvenaci; upraveno [34].

2.2 Sekvenační platformy

Nejčastěji používané sekvenační platformy jsou i nadále platformy NGS Illumina [5] a IonTorrent [6]. Dále jsou často využívány sekvenační platformy, které jsou schopny sekvenovat dlouhá čtení. Tyto platformy spadají do skupiny TGS a patří do nich například platformy Pacific Biosciences (PacBio) nebo Oxford Nanopore. Jedná se o platformy s velkým potenciálem pro budoucí využití [34]. Rozdíly jednotlivých sekvenačních přístupů, které jsou dány především odlišnou délkou a počtem sekvenovaných čtení v jednom běhu, lze pozorovat na Obrázku 2.5.



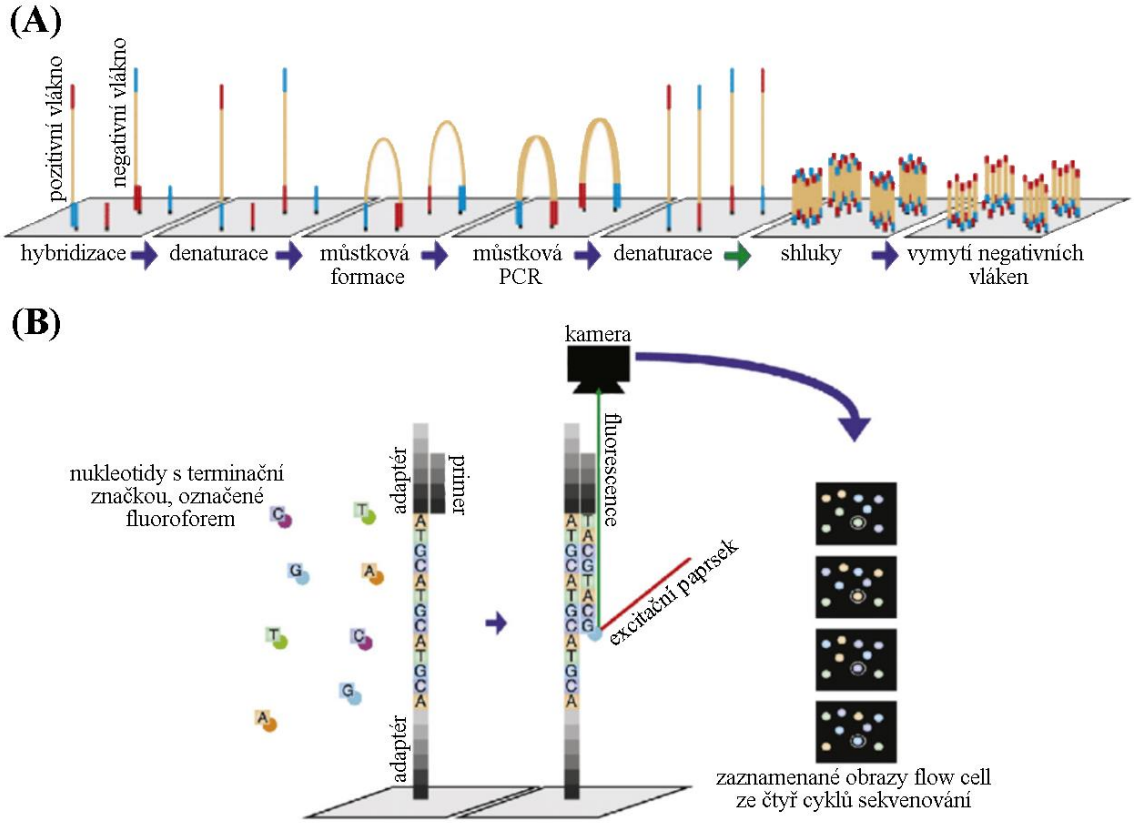
Obrázek 2.5 Graf znázorňující rozdíly mezi sekvenačními platformami v závislosti na délce jednotlivých čtení a počtu sekvenovaných čtení; upraveno [6].

2.2.1 Illumina

Illumina je celosvětově využívaná technologie pro sekvenování. K roku 2020 bylo 82 % bakteriálních genomů z databáze RefSeq sekvenováno právě metodou Illumina [47]. Patří do skupiny NGS a jedná se o technologii sekvenování založené na syntéze. Velikou výhodou této metody je možnost širokého paralelního sekvenování.

Princip Illuminy je vytvoření shluků stejné cDNA pomocí můstkové PCR [48]. Celý proces je znázorněn na Obrázku 2.6 v sekci (A). Nejprve dochází k denaturaci přichystané knihovny a následné hybridizaci fragmentů cDNA na hybridizační povrch, tzv. flow cell. Flow cell je podobná mikroskopickému sklíčku, které je rozděleno do polí a je potaženo dvěma druhy oligonukleotidů, jež jsou komplementární k adaptérům cDNA. Posléze dochází k denaturaci cDNA a následné můstkové PCR. Volný konec cDNA s adaptérem, který prozatím nebyl nikde navázán, je hybridizován k flow cell v těsné blízkosti původně hybridizovaného adaptéru. Tím vzniká tzv. můstková formace cDNA. Následuje PCR, při které dochází k dosyntetizování druhého vlákna můstkové formace a následnou denaturací je vytvořena nová komplementární kopie pro původní

cDNA. PCR takto probíhá ještě 24krát, čímž jsou vytvořeny jednotlivé shluky kopií původních cDNA. Na závěr jsou z flow cell vymyta negativní vlákna, čímž dochází k ponechání pouze pozitivních vláken na flow cell, které tvoří výsledné shluky.



Obrázek 2.6 Sekvenování pomocí sekvenátoru Illumina: (A) příprava shluků pro sekvenaci; (B) sekvenování a detekce signálu; upraveno [48].

Takto vzniklé shluky jsou následně sekvenovány. Sekvenace začíná hybridizací sekvenačních primerů na volný adaptér cDNA. Dochází k postupnému přidávání volných nukleotidů (2'-deoxynukleosid 5'-trifosfátů, dNTP) značených odlišnými fluorofory. Tyto nukleotidy mají inaktivované 3‘ konce pomocí terminačních značek. Díky tomu je docíleno, že se v jednom cyklu může navázat pouze jeden nukleotid do jednotlivého řetězce. Po navázání daného nukleotidu do řetězce je zbytek volných nukleotidů odmyt. Zbylé navázané nukleotidy jsou excitovány pomocí laseru a fluorescenčním detektorem (např. kamerou) je zachycen celý obraz flow cell, na kterém jsou zaznamenány fluorescence způsobené excitovanými fluorofory. Poté dochází k odstranění fluoroforu a terminační značky z nukleotidu a k přidání nových značených nukleotidů, čímž se celý proces opakuje. Počet opakování tohoto cyklu závisí na předem dané délce čtení, která je při sekvenování nastavena. Schéma znázorňující princip snímání obrazů flow cell je zobrazeno na Obrázku 2.6 v sekci (B).

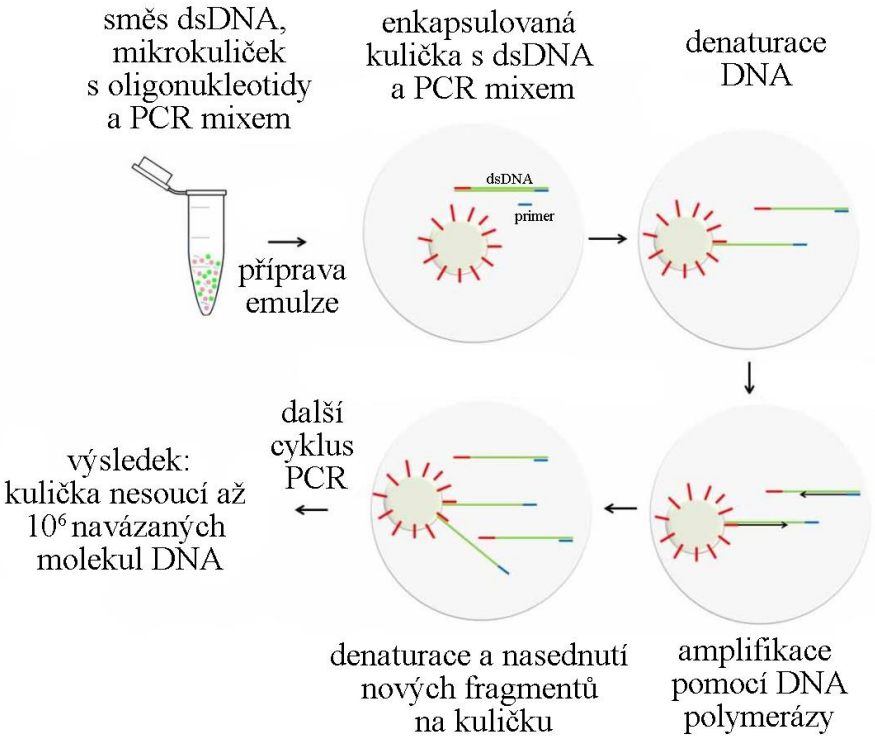
Illumina umožňuje postup zaměřující se na přípravu knihovny s vysokou citlivostí a dynamickým rozsahem pro následnou predikci malých nekódujících RNA [7]. Je tedy

možné zkoumat i diferenciální expresi malých RNA, což je zásadní pro jejich predikci a zejména pro určení jejich funkce. Pro přípravu knihovny pro tuto metodu byl vyvinut speciální kit s názvem *TruSeq Small RNA Library Prep Kit*.

2.2.2 Ion Torrent

Ion Torrent je první sekvenační technologií, která nebyla založena na detekci světelného signálu (nevyužívá ani fluorescenci, ani luminiscenci) [49], ale pro sekvenaci využívá detekci změny pH v prostředí, kterou způsobují uvolněné ionty vodíku z deoxynukleosidtrifosfátů (dNTP) během syntézy druhého vlákna cDNA [50].

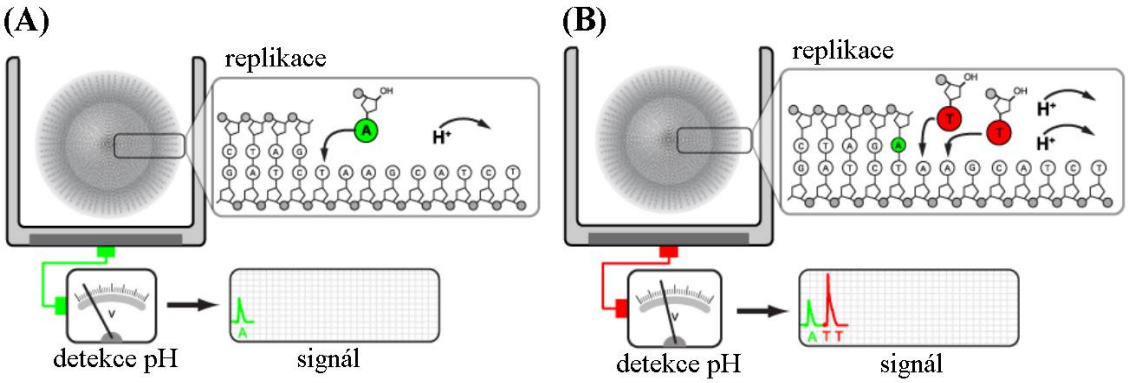
cDNA je syntetizována na dsDNA. Tato dsDNA obsahující specifické adaptéry jsou hybridizovány na mikrokuličky. Tyto mikrokuličky mají na svém povrchu komplementární oligonukleotidy k adaptérům cDNA. Koncentrace cDNA je taková, aby došlo k zachycení pouze jedné molekuly cDNA na jednu kuličku. Na každé z těchto mikrokuliček dochází k emulzní PCR. Emulzní PCR [49; 51] je založena na zapouzdření (enkapsulování) ideálně jednotlivé kuličky do emulze amplifikační tekutiny s olejem. V tomto momentu může docházet k amplifikaci cDNA na kuličkách. K navázané jednovláknové cDNA se dosyntetizuje druhé vlákno a následnou denaturací je odštěpeno. Nově vzniklá vlákna, komplementární k oligonukleotidům na kuličce, nasedají na kuličku, čímž dochází k amplifikaci dané cDNA po celé kuličce. Schéma procesu emulzní PCR je zobrazeno na Obrázku 2.7.



Obrázek 2.7 Schéma emulzní PCR; upraveno [52].

Jakmile jsou kuličky s dostatečným množstvím cDNA připravené, jsou jednotlivě umístěny do mikrojamek obsahujících komplementární polovodič oxidu kovu (CMOS čip) [50]. CMOS čipy jsou senzory vyznačující se schopností přenášet informace z jednotlivých bodů samostatně [53]. Na tomto čipu jsou integrované iontově senzitivní tranzistory (ISFET), které jsou schopny detekovat změnu pH v daném poli. Následně dochází k přidání primeru a DNA polymerázy pro syntézu druhého vlákna cDNA. Poté je přidán jeden z nukleotidů dNTP. Jestliže dochází k navázání tohoto nukleotidu pomocí DNA polymerázy, je uvolněn vodíkový ion H^+, který změní pH roztoku obsaženého v mikrojamce. Tuto změnu je schopný zaznamenat ISFET a změna signálu je převedena do PC, kde je následně vyhodnocena pro daný nukleotid. Jestliže k navázání nukleotidu nedochází, signál se nemění. Po zaznamenání změny signálu jsou zbylé nukleotidy odmyty pryč a celý proces se opakuje s dalšími nukleotidy.

Metoda Ion Torrent má také schopnost detekovat homopolymery. Jestliže se při přidání jednoho z nukleotidů naváže za sebou vyšší počet z těchto nukleotidů, výsledný změněný signál je přímo úměrný vzhledem k počtu takto za sebou navázaných nukleotidů. Princip sekvenování metodou Ion Torrent je vyobrazen na Obrázku 2.8 v sekci (A) pro jeden navázaný nukleotid a v sekci (B) pro dva navázané nukleotidy.



Obrázek 2.8 Sekvenování pomocí Ion Torrent: (A) navázání jednoho nukleotidu do řetězce a detekce změny pH; (B) navázání dvou nukleotidů do řetězce v jednom cyklu a detekce zvýšené změny pH; upraveno [50].

2.2.3 Pacific Biosciences

Společnost Pacific Biosciences vyvinula sekvenační metodu PacBio spadající do třídy TGS, která umožňuje sekvenaci jediné molekuly v reálném čase (single molecule real time, SMRT) [43].

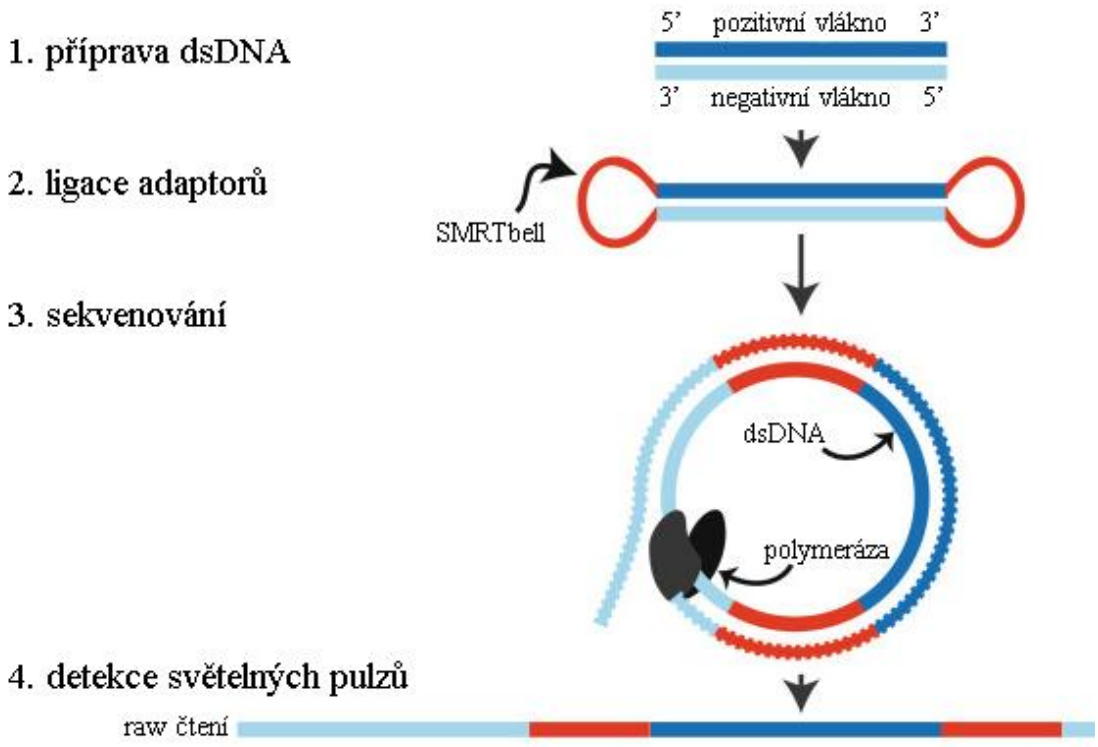
Principem této metody je sekvenace cDNA, která není nijak fragmentovaná. Z původní jednovláknové cDNA musí být připravena dsDNA. Na oba konce dsDNA jsou následně ligovány vlásenkové adaptéry, které zapříčiní vznik kruhové jednovláknové DNA nazývané jako SMRTbell.

SMRTbell je poté nanesena na čip zvaný SMRTcell. Zde se nachází jamky o velikosti 100 nm tzv. zero-mode waveguide (ZMW), jejichž objem má minimální

velikost potřebnou pro detekci světla. Každý čip obsahuje 150 000 těchto jamek. Na každé ZMW je imobilizována jedna polymeráza sloužící pro navázání některého z adaptérů SMRTbell. Jakmile je SMRbell uchycena, jsou do SMRTcell přidány všechny čtyři nukleotidy, které jsou značeny fluorofory lišícími se v generujícím emisním spektru. Pomocí polymerázy dochází k replikaci celé SMRTbell, kdy se postupně do řetězce začleňují fluorescenčně značené nukleotidy. Při jejich začlenění jsou generovány odlišné světelné pulzy, které jsou detekovány a zaznamenány.

Jestliže má polymeráza dostatečně dlouhou životnost, replikace SMRTbell může být provedena vícekrát. Takovouto sekvenací je obdržena kruhová konsensuální sekvence, která poskytuje čtení s lepší přesností.

Během jednoho běhu je využito zhruba 35 000-75 000 ZMW, ve kterých dochází k sekvenaci rozdílných cDNA, jedná se tedy o metodu podporující paralelní sekvenování. Zjednodušené schéma znázorňující sekvenační metodu PacBio je vyobrazeno na Obrázku 2.9.



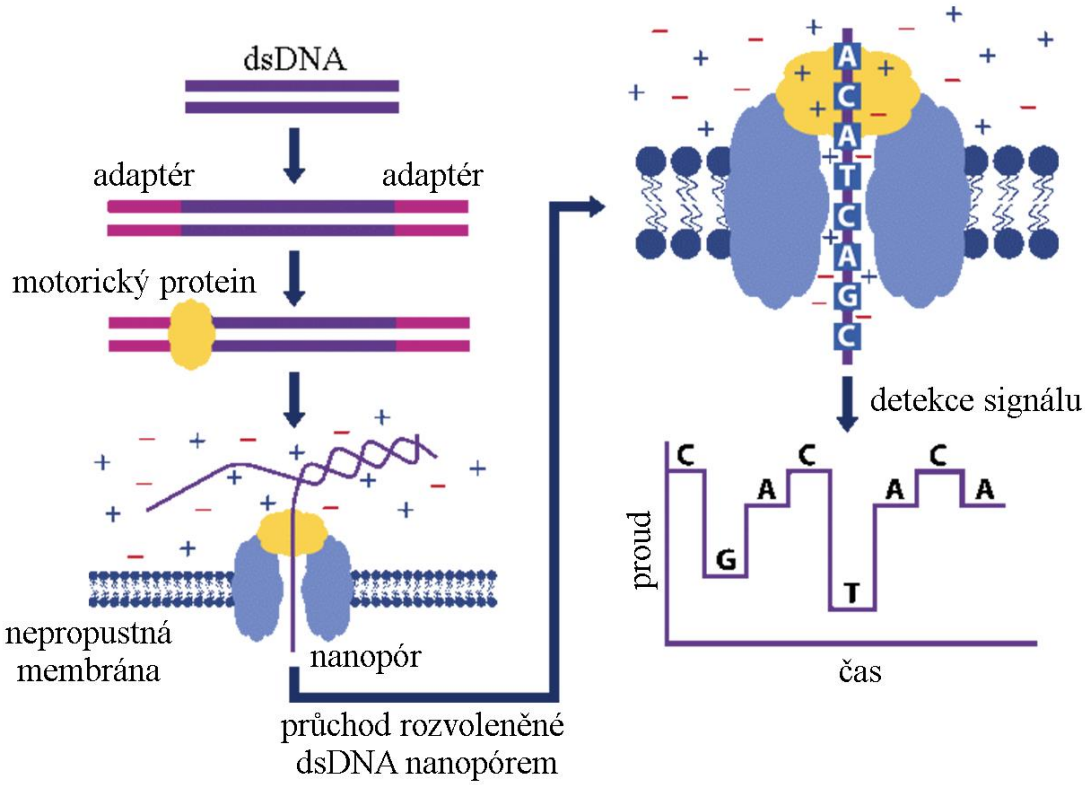
Obrázek 2.9 Schéma sekvenační metody PacBio; upraveno [54].

2.2.4 Oxford Nanopore

Stejně jako PacBio i Oxford Nanopore je sekvenační metoda patřící do TGS, která umí sekvenovat jedinou molekulu DNA v reálném čase 0. Jedná se o první sekvenační metodu, která nevyužívá při samotné sekvenaci replikaci, ale pouze zaznamenává pořadí nukleotidů v cDNA pomocí změny elektrického proudu.

U této metody dochází k sekvenaci dsDNA, na které jsou ligovány dva rozlišné adaptéry. První je zaváděcí adaptér, který je značen písmenem Y. Svůj název nese díky své struktuře připomínající toto písmeno. Na druhý konec je navázán adaptér hairpin, který má vlásenkovou strukturu. Na adaptér Y se následně komplementárně váže motorický protein, který má schopnost rozvolňovat dsDNA.

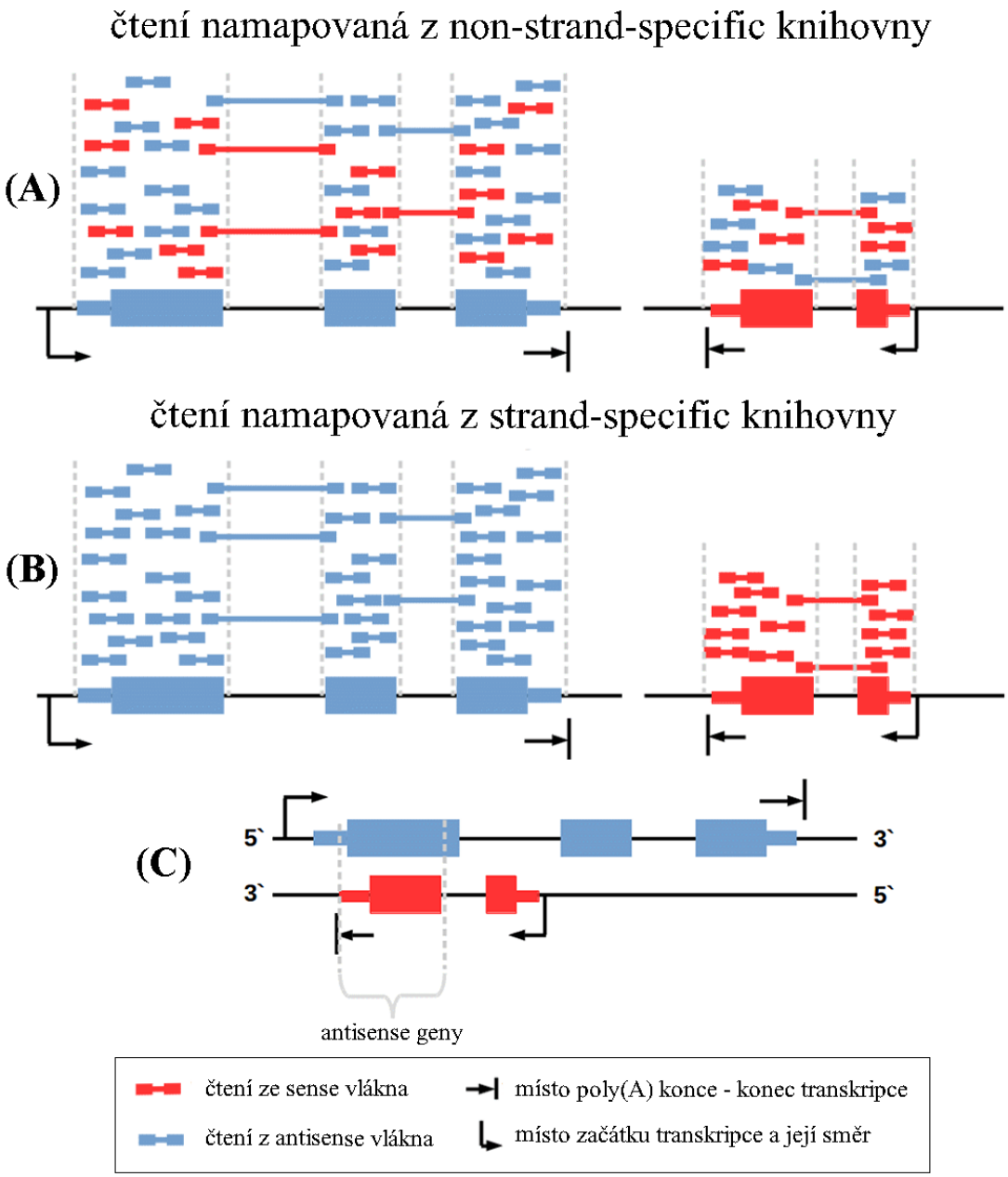
Takto přichystaný vzorek je přidán na flow cellu, která obsahuje zcela nepropustnou membránu s propustnými nanopóry. Motorické proteiny tak přivedou DNA k jednotlivým nanopórům, kde se navážou. Na membránu je přivedeno napětí, které zapříčiní posun DNA napříč nanopórem. Jednotlivé nukleotidy v sekvenci DNA mění iontový proud, který je zaznamenáván senzory uloženými u jednotlivých nanopórů. Tyto senzory jsou schopny zaznamenávat změnu proudu až tisíckrát za jednu sekundu. Zjednodušené schéma výše zmíněné sekvenační metody je znázorněno na Obrázku 2.10.



Obrázek 2.10 Schéma sekvenační metody Oxford Nanopore; upraveno [55].

2.3 Strand-specific a non-strand-specific knihovna

U RNA-Seq knihoven existují dva základní typy: stranded a non-stranded knihovna [56; 57]. Jako stranded je nazývána taková knihovna, které nese informaci, z jakého vlákna dané čtení pochází. Tato informace může být obdržena zvolením odlišných adaptérů, které jsou ligovány k 3‘ a 5‘ koncům RNA a nesou tedy informaci o tom, ze kterého vlákna pochází.



Obrázek 2.11 Srovnání non-strand specific a strand-specific dat: (A) non-strand-specific knihovna; (B) strand-specific knihovna; (C) výhoda strand-specific knihovny; upraveno [58].

Na rozdíl od stranded knihovny, non-stranded knihovna tuto informaci neobsahuje. Díky tomu ve výsledné analýze dochází k překryvu čtení z negativního vlákna na pozitivní vlákno, čímž může dojít ke zkresleným výsledkům. Názorná ukázka je vyobrazena na Obrázku 2.11. V části (A) je zobrazena non-stranded knihovna, kde se mapují čtení vůči referenci chybně a do výsledků je zanesena chyba. V sekci (B) lze pozorovat stranded knihovnu, kde jsou čtení k referenci mapována správně – pozitivní čtení k pozitivnímu vláknu a negativní čtení k negativnímu vláknu. Hlavní výhodu můžeme pozorovat v sekci (C), kde jsou znázorněny geny, které jsou na rozdílných

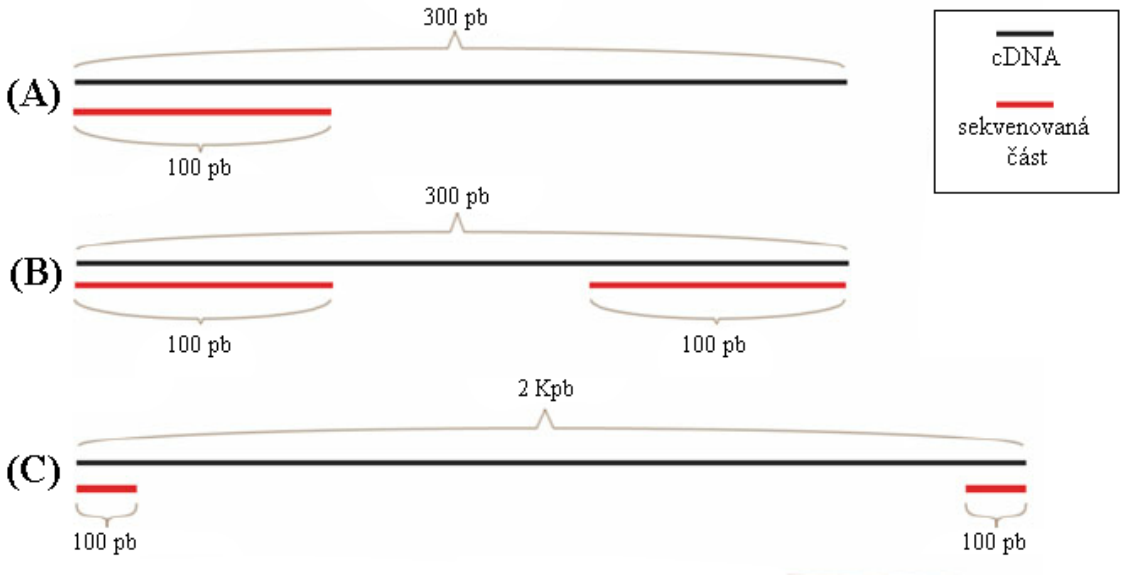
vláknech a bez stranded dat by nebylo možné určit, která čtení náleží kterému genu a tím pádem by došlo k nesprávnému vyhodnocení namapovaných čtení.

Posledním podtypem stranded knihovny (second-strand knihovny) je knihovna nazývaná reversly stranded (first-strand knihovna) [59]. Reversly stranded knihovna je získána tak, že z původní RNA je syntetizována komplementární cDNA, která je poté sekvenována. Díky tomu jsou obdržená čtení vůči původním transkriptům komplementární (tzn. směr čtení je opačný vůči směru transkriptu, ze kterého pochází). Kdežto stranded data, jak je uvedeno výše, obsahují čtení, jejichž směr odpovídá směru transkriptu.

2.4 Single-end, paired-end a mate-end data

Při sekvenování mohou být knihovny cDNA sekvenovány různými způsoby a tedy vznikají různé typy výsledných čtení [60]. Prvním typem jsou single-end data, která jsou získána tak, že konkrétní molekula cDNA je sekvenována pouze z jednoho konce, viz Obrázek 2.12, sekce (A).

Dalším typem jsou paired-end data. Tato data obsahují stejnou část, jako u single-end dat, ale dochází k sekvenaci i z druhého konce cDNA. Tím je zapříčiněno zkrácení maximální vzdálenosti mezi dvěma čteními v genomu, viz Obrázek 2.12, sekce (B). Posledním typem je metoda mate-end, která je velmi podobná typu paired-end. Liší se v délce původní molekuly cDNA, která je sekvenována. U metody mate-end jsou molekuly cDNA delší, viz Obrázek 2.12 sekce (C).



Obrázek 2.12 Typy obdržených čtení: (A) single-end data; (B) paired-end data, (C) mate-end data; upraveno [60].

3. DOSTUPNÉ NÁSTROJE PRO PREDIKCI sRNA

3.1 Rockhopper

Rockhopper je volně dostupný software pro analýzu bakteriálních RNA-Seq dat implementovaný v programovacím jazyku Java pod licencí GNU GPL [8]. Byl vyvinut již v roce 2013 a řadí se tak mezi nejstarší nástroje pro analýzu těchto dat za účelem zjistit pozice všech nekódujících RNA v genomu bakterie [9]. V roce 2015 byl upraven, aby dával ještě přesnější výsledky jak z hlediska citlivosti, tak i z hlediska specificity, a aby bylo možné sestavovat transkripty *de novo* [8].

Rockhopper je schopný najednou pracovat s více daty z různých experimentů tak, aby mohl zachytit diferenciální genovou expresi mezi těmito experimenty. Vstupními daty Rockhopperu je genom v datovém formátu FASTA, anotace genomu v datovém formátu PTT nebo RNT a čtení z jednotlivých experimentů v jednom z uvedených datových formátů: FASTAQ, QSEQ, FASTA, SAM nebo BAM. [61]

Nástroj v prvním kroku provede zarovnání čtení vůči genomu a normalizaci počtu namapovaných čtení. Následně na základě počtu transkriptů určuje polohu hledaných nekódujících RNA (5'UTR, 3'UTR, sRNA) a pomocí analýzy diferenciální genové exprese identifikuje pozice operonů. Všechny tyto výsledky souhrnně vizualizuje pomocí Integrated Genomics Viewer (IGV), což je software určený pro vizualizaci geonomických dat. Výstupem celého procesu jsou tři soubory ve formátu TXT. První z těchto souborů (*Summary*) obsahuje obecné informace o počtech jednotlivých detekovaných nekódujících RNA, druhý (*transcripts*) obsahuje podrobnější informace o genech a nově nalezených transkriptech a poslední soubor (*operons*) obsahuje informace o detekovaných operonech [9]. Stručný postup celé analýzy je vyobrazen na Obrázku 3.1.

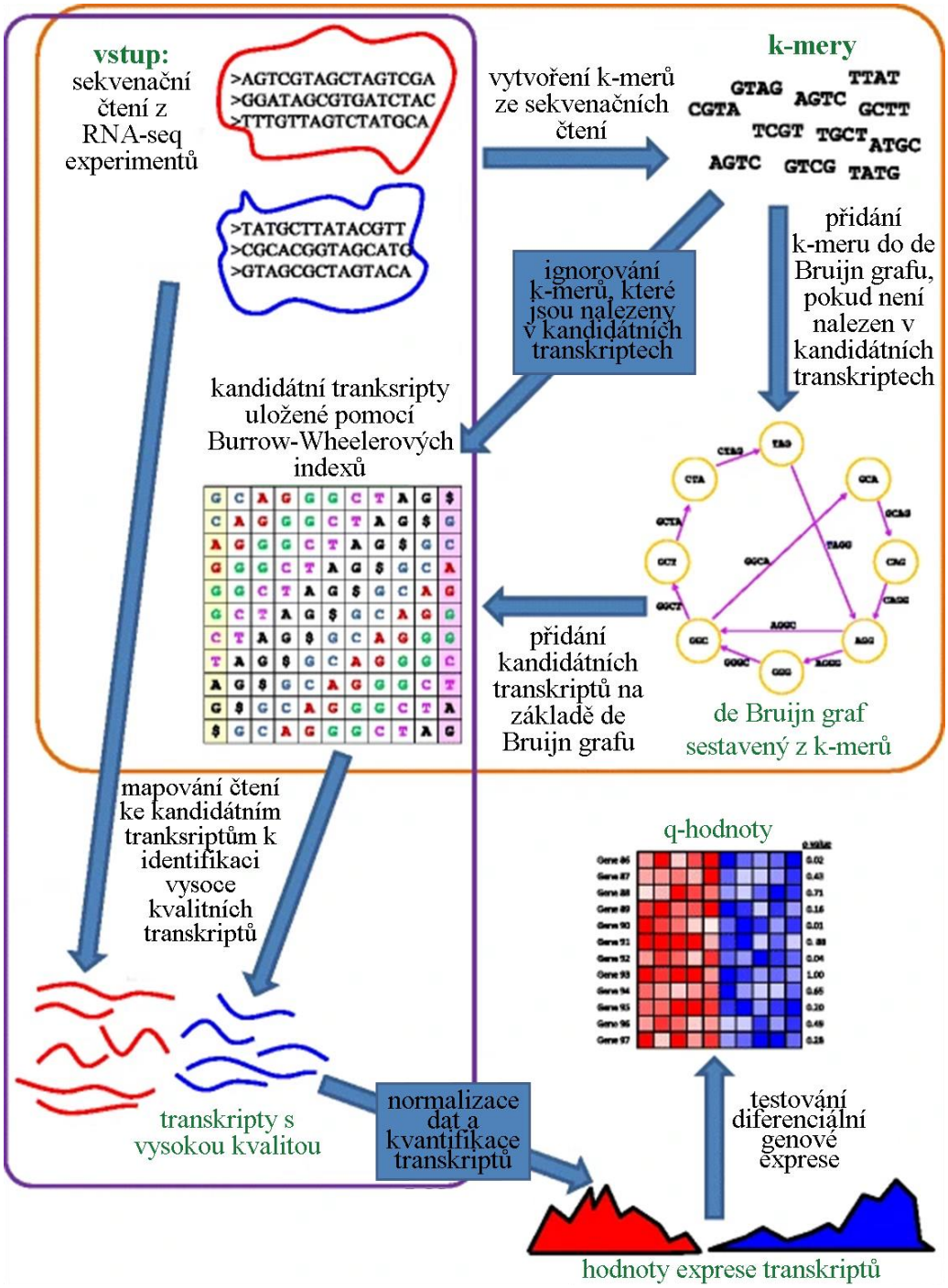
3.1.1 Skládání transkriptů *de novo* a zarovnání čtení k referenčnímu genomu

Rockhopper umožňuje skládat čtení *de novo* nebo zarovnávat k referenci. Vstupní čtení mohou být typu paired-end či single-end a mohou být stranded nebo non-stranded.

Sestavení *de novo* je provedeno tak, že jsou nejprve z dostupných čtení vytvořeny k-mery (výchozí nastavení: délka 25 pb). Tyto k-mery jsou posléze spojovány za využití de Bruijnových grafů a Burrow-Wheelerovy transformace. Toto spojení je ojedinělé právě pro Rockhopper a díky tomu nabízí relativně časově nenáročné *de novo* skládání transkriptů [8].

Mapování čtení k referenci je možné pouze s dostupnou genomovou sekvencí. Rockhopper umožňuje automaticky vyhledat a stáhnout reference, které jsou volně dostupné v GenBank databázi [61]. Samotné zarovnání čtení je provedeno za pomocí volně přístupného nástroje Bowtie2. Jedná se o velice rychlý nástroj vyznačující se i svou vysokou senzitivitou a přesností [62]. Bowtie2 využívá k indexování v genomu

full-text (FM) indexy, které jsou založeny na Burrow-Wheelerově transformaci [63]. Po vytvoření FM indexu může být čtení, plně se shodující s referenčním genomem, zarovnáno. Jestliže se dané čtení neshoduje s genomem, je pomocí Smith-Watermanova algoritmu rozšířeno a zarovnáno (může dojít i k vícenásobnému zarovnání) tak, aby výsledné zarovnání obsahovalo maximální povolený počet neshod vůči referenci. Tato hodnota je určena z Phred skóre, získaného při sekvenování. Zvolená sekvenační metoda tedy ovlivňuje výpočet pro výsledné zarovnání. [9]



Obrázek 3.1 Schéma postupu detekce softwaru Rockhopper; upraveno [8].

3.1.2 Identifikace nekódujících RNA

Jak již bylo zmíněno, Rockhopper je nástroj pro detekci nekódujících RNA. Tyto detekce jsou založeny na základě normalizovaných hodnot pokrytí čtení. Nejprve dochází k nalezení tzv. semen, které mají určitou délku (výchozí nastavení: 10 pb) a pokrytí, které je vyšší než prahová hodnota udávající průměrné pokrytí napříč celým genomem.

Následně jsou tato semena rozšiřována oběma směry na základě Bayesovské statistiky. Je určeno semeno *s* a region z genomu *r* takový, že alespoň jeden nukleotid sdílí se semenem *s*. Cílem je určit, zda region *r* náleží stejnému transkriptu jako semeno *s* na základě Bayesovské statistiky:

$$p(C|x_r) = \frac{p(C)p(x_r|C)}{p(x_r)}, \tag{0.1}$$

kde x_r udává počet čtení mapujících se do regionu *r* a *C* je závislá proměnná ($C = \{c_{r\leftrightarrow s}, c_{r|s}\}$) se dvěma různými výsledky. Varianta $c_{r\leftrightarrow s}$ označuje, že region *r* náleží stejnému transkriptu jako semeno *s*, $c_{r|s}$ naopak. Jsou tedy určeny dvě pravděpodobnosti – pravděpodobnost $p(c_{r\leftrightarrow s}|x_r)$ udávající, že region *r* je součástí stejného transkriptu jako semeno *s*, a pravděpodobnost $p(c_{r|s}|x_r)$, že není. První z těchto pravděpodobností je založena na Poissonově rozložení čtení mapovaných k semeni *s* za předpokladu, že jsou vzorkována rovnoměrně a nezávisle. Druhá z těchto pravděpodobností je dána geometrickým rozložením pozadí všech čtení mapujících se antisense vůči kódujícím proteinům dle anotace.

Zda dojde k přidání regionu *r* k semeni *s* nebo nikoliv, spočívá v tom, která z výše uvedených pravděpodobností má vyšší hodnotu. Jestliže dojde k situaci, že se takto rozšířená semena přesahují, dojde k jejich sloučení v semeno nové. Původní semena zaniknou. [9]

3.1.3 Nastavení parametrů uživatelem

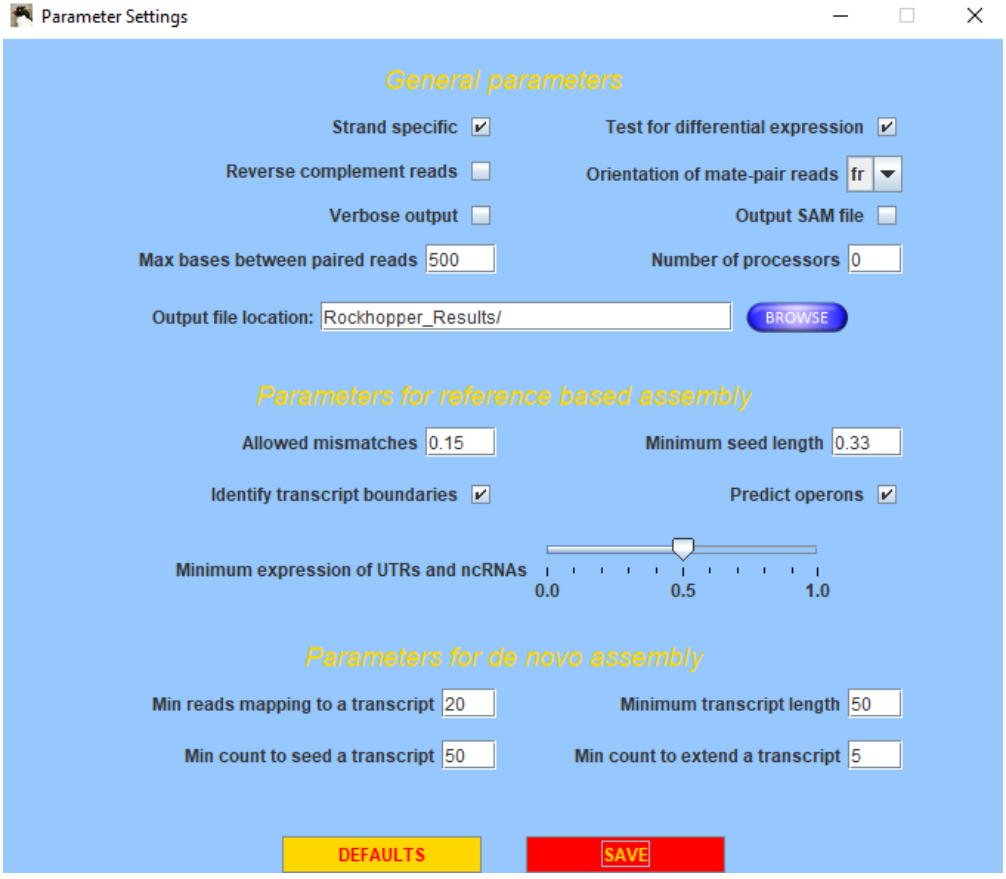
Uživatel má v prostředí Rockhopper možnost nastavit několik proměnných [61]. Všechny tyto proměnné lze upravit ve vyskakovacím okně *Parameter Settings*, viz Obrázek 3.2.

V první sekci *General parameters* si uživatel volí základní parametry analýzy. Jsou zde voleny informace, zda uživatel používá data ze sekvenování, která jsou specifická či nespecifická pro dané vlákno a jestli má single-end, paired-end či mate-end data. Dále je možné zatrhnout volbu pro výpočet diferenciální genové exprese, pro výpis podrobnějších výsledků, nebo pro zápis výsledků do SAM souboru. Také se zde volí složka, do které jsou výstupní data ukládána. Posledním parametrem v této sekci je možnost určit Rockhopperu, kolik procesorů v počítači může pro výpočty použít. Doporučeno je nastavení na hodnotu 0, kdy Rockhopper sám určuje, kolik procesorů využije.

Druhá sekce *Parameters for reference based assembly* nastavuje parametry pro zarovnání k referenci a obsahuje možnost nastavení parametru neshody, kdy si uživatel volí, jaká je maximální přípustnost odlišnosti zarovnaného čtení vůči referenci (do neshody se započítávají i povolené mezery a jsou uvedeny v procentech vůči délce daného čtení) a parametr minimální délky semene (procento počtu přesně zarovnaných nukleotidů z délky čtení). Uživatel si také může zvolit, zda chce, aby byly predikovány operony a jestli chce detekovat a analyzovat transkripty, které zatím nejsou anotované. Tato nastavení hrají zásadní roli ve výpočetní rychlosti. Poslední parametr této sekce je parametr minimální exprese UTR a nekódujících RNA (ncRNA), který ovlivňuje specificitu a senzitivitu detekce (čím vyšší hodnota, tím vyšší specificita, ale nižší senzitivita).

Poslední část *Parameters for de novo assembly* je zaměřena pouze pro nastavení *de novo* skládání. Je zde možné nastavit minimální počet přesně namapovaných čtení pro daný transkript, aby byl tento transkript uvažován, dále minimální délku sestaveného transkriptu, minimální počet k-merů pro dané semeno a minimální počet stejných k-merů pro sloučení dvou semen.

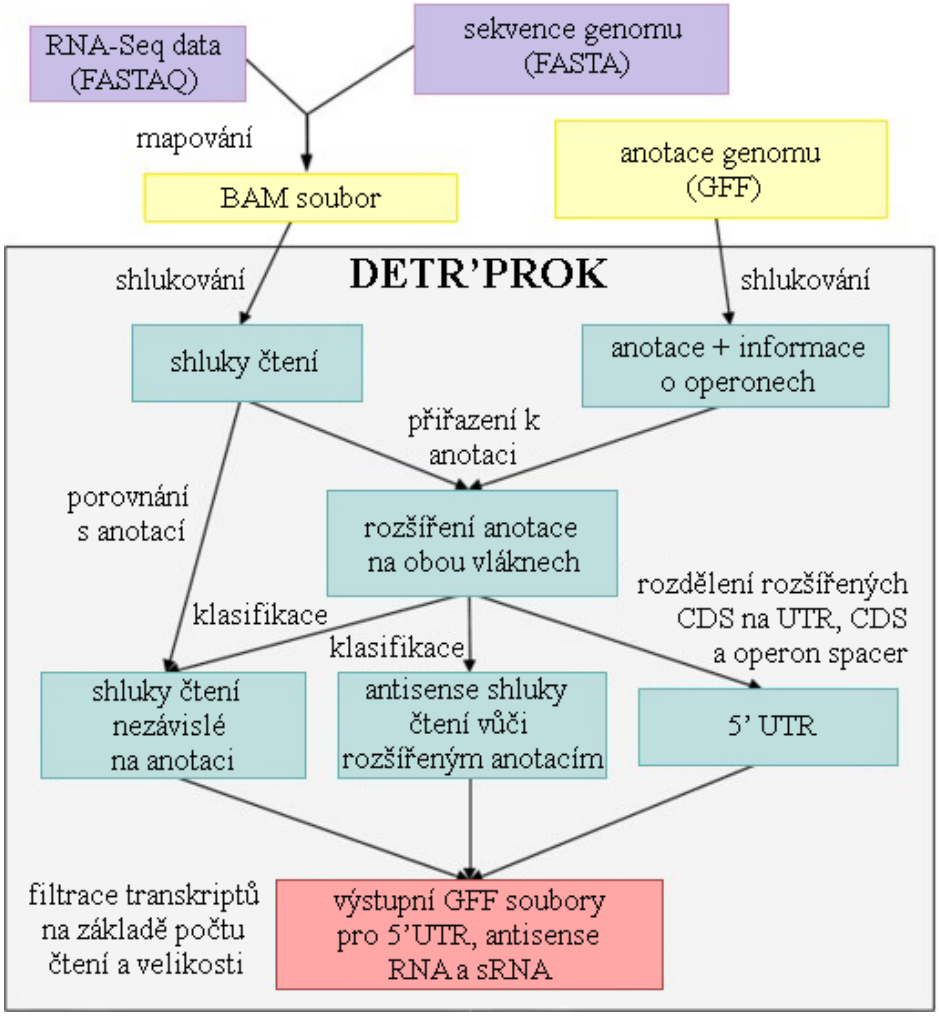
Samozřejmě je možno zvolit, zda bude analýza prováděna *de novo* nebo s referencí, kolik experimentů proběhlo a kolik sekvenačních dat v jednotlivých experimentech bylo získáno. Tato nastavení se provádí již v hlavním okně Rockhopperu.



Obrázek 3.2 Ukázka vyskakovacího okna *Parameter Settings* programu Rockhopper.

3.2 DETR'PROK

Další z metod zabývající se detekcí nekódujících RNA je DETR'PROK, který byl stejně jako Rockhopper uveden již v roce 2013 [10]. Analýza dat metodou DETR'PROK je založena na dostupných nástrojích, které jsou spojeny pomocí platformy Galaxy. Webová platforma Galaxy je open source určená pro práci s velkým množstvím dat nejen v genomice, ale také proteomice a metabolomice. Galaxy má tři hlavní cíle – aby analyzační postupy byly všem dostupné, byly reprodukovatelné a získané výsledky byly publikovatelné a transparentní [64].



Obrázek 3.3 Zjednodušené schéma postupu analýzy programu DETR'PROK; upraveno [10].

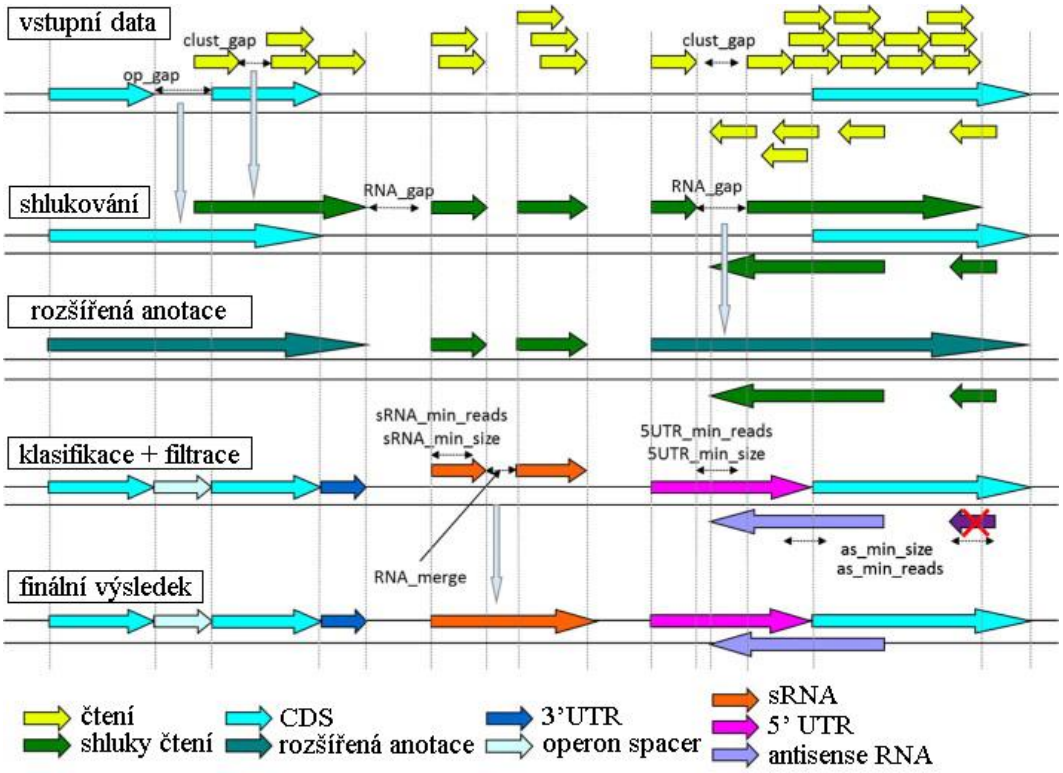
3.2.1 Princip detekce DETR'PROK

Jedná se o výpočetní algoritmus, do kterého musí vstupovat stranded data z RNA-Seq ve formátu BAM a anotace ve formátu GFF. Poté algoritmus shlukuje tato vstupní data, rozšiřuje dostupné anotace na základě namapovaných čtení z BAM souboru a predikuje nekódující RNA – 5'UTR, sRNA, asRNA. Predikce těchto RNA je provedena

ve čtyřiceti krocích. Většina z těchto kroků má základy ze sady nástrojů S-MART pro práci s bioinformatickými daty. Všechny tyto kroky jsou volně přístupné a editovatelné, aby si je uživatel mohl upravit dle potřeb [10]. Zjednodušené schéma pracovního postupu DETR'PROK lze vidět na Obrázku 3.3, kde jsou ve fialových obdélnících znázorněna data, která jsou potřebná pro vytvoření vstupních dat, jež jsou zobrazena žlutě. V zelených obdélnících jsou pak znázorněny jednotlivé pracovní kroky. Výstupní soubor v datovém formátu GFF je vyobrazen červeně, tento soubor obsahuje informace o detekovaných 5'UTR a sRNA (včetně antisense RNA).

3.2.2 Nastavení parametrů uživatelem

Na začátku kódu si může uživatel nastavit několik parametrů. Tyto parametry jsou přímo určené k aktivnímu nastavení uživatelem vzhledem k použitým datům a očekávaným výsledkům. Funkce jednotlivých parametrů v algoritmu je znázorněna na Obrázku 3.4.



Obrázek 3.4 Funkce parametrů v algoritmu DETR'PROK; upraveno [10].

Prvním parametrem, který se nastavuje, je parametr *feature_list*, který odpovídá seznamu transkriptů, jež budou z dostupné anotace využity (např. CDS, rRNA, tRNA). U tohoto parametru je důležité zkontrolovat soubor obsahující anotace pro určení správných zkratek pro dané transkripty (v různých anotacích se mohou lišit).

Následně se uvažuje nastavení maximálních mezer pro čtení (*clust_gap*), CDS (*op_gap*) a potenciálních RNA (*RNA_gap*). Nastavením těchto parametrů dochází ke sloučení/ponechání blízkých čtení, CDS a následně navzájem blízko detekovaných

transkriptů vůči sobě. Možnost zvolení těchto hodnot je vhodné vzhledem k různorodosti hustoty genů v genomu napříč bakteriálními druhy a kvůli odlišné hloubce sekvenování při rozlišných experimentech.

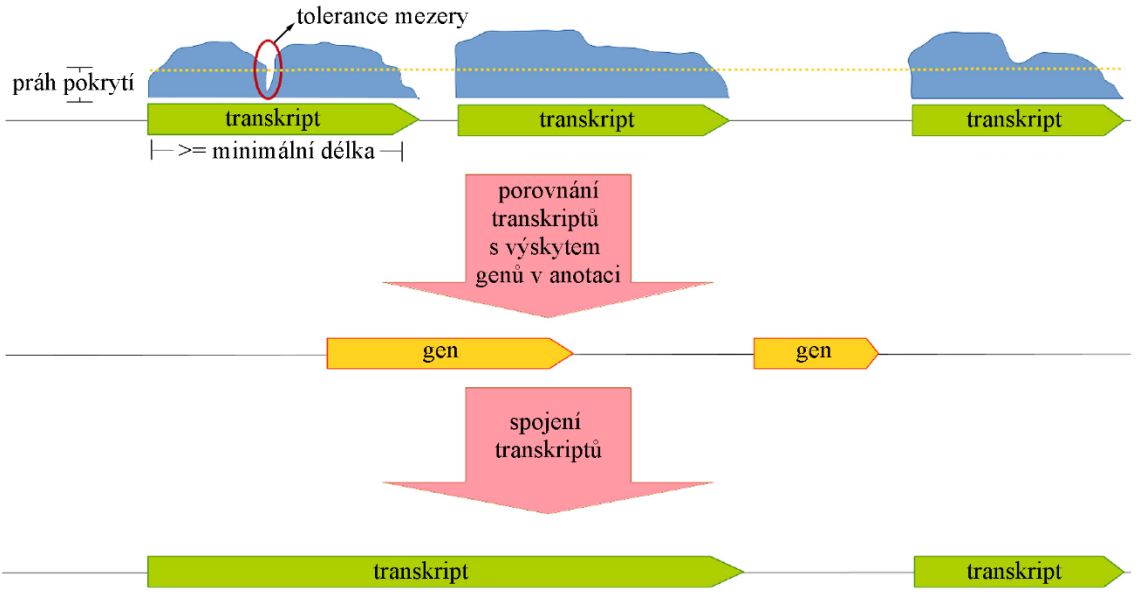
Dalším parametrem je *RNA_merge* udávající maximální vzdálenost, aby se nově detekované transkripty spojily v jeden. Poslední parametry udávají minimální počet čtení na nový transkript (*sRNA_min_reads, asRNA_min_reads 5'UTR_min_reads*) a jeho minimální délku (*sRNA_min_size, asRNA_min_size, 5UTR_min_size*).

3.3 ANNOgesic

Nepříliš starým způsobem detekce sRNA je metoda ANNOgesic, která byla poprvé uvedena v roce 2018 [11]. ANNOgesic je nástroj, který byl vyvinut v programovacím prostředí Python 3 a je volně dostupný jako open source. Jedná se o metodu, která byla vyvinuta na zpracování RNA-Seq a diferenciálních RNA-Seq (dRNA-Seq) dat tak, aby bylo detekováno co možná nejvíce funkčních transkriptů. ANNOgesic je schopný detekovat operony, geny (CDS), tRNA, rRNA, sRNA, 3' a 5' UTR konce, riboswitche, místa pro začátek transkripce a další. Algoritmy pro detekci jednotlivých částí jsou rozděleny do jednotlivých modulů, které mohou být různě propojovány.

Vstupními daty pro ANNOgesic mohou být informace o pokrytí z RNA-Seq ve formátu WIGGLE nebo zarovnaná čtení ve formátu BAM. Některé z modulů vyžadují navíc také anotaci v datovém formátu GFF3. Výstupem je pak nová anotace v datovém formátu GFF3 obsahující informace o detekovaných transkriptech, soubory v datovém formátu CSV obsahující informace z jednotlivých modulů a obrázky. Pro zobrazení nové anotace může být použit software IGV.

3.3.1 Princip detekce ANNOgesic

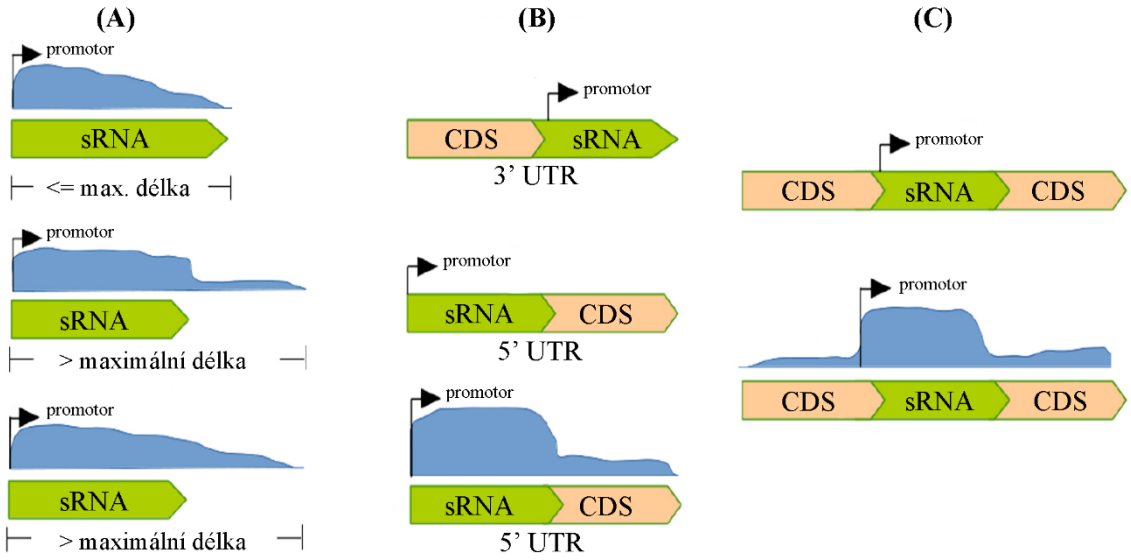


Obrázek 3.5 První část detekce sRNA u ANNOgesic; upraveno [11].

V prvním kroku celého procesu jsou nalezeny potenciální transkripty. Jak lze vidět na Obrázku 3.5, dochází k detekci transkriptů na základě pokrytí namapovaných čtení. Minimální práh pro pokrytí si uživatel nastavuje sám, stejně tak maximální toleranci mezery, která má nižší pokrytí než je daný práh, a minimální délku transkriptů. Poté dochází ke spojování transkriptů vzhledem ke znalostem získaných z anotace. Například dva transkripty mohou být spojeny do jednoho, a to v případě, kdy tyto transkripty překrývají stejný gen.

Samotná detekce sRNA u ANNOgesic je v modulu s názvem *sRNA*. Nejprve algoritmus porovnává potencionální sRNA transkripty s databází BSRD [65], která obsahuje experimentálně potvrzené bakteriální sRNA. Jestliže dojde ke shodě, potenciální úsek je prohlášen za sRNA. Pokud ke shodě nedojde, porovnává se transkript s databází Národního centra pro biotechnologické informace (NCBI), jestli se nejedná o neredundantní protein. Pokud nedojde k nalezení transkriptu ani v jedné z těchto databází, musí transkript splňovat určité podmínky, aby mohl být prohlášen za sRNA. První z podmínek je přítomnost promotoru pro daný úsek, vhodná sekundární struktura transkriptu a délka mezi 30 až 500 pb. Takto jsou detekovány sRNA, které se nacházejí v intergenovém prostoru (*trans*-sRNA) a nebo antisense vůči genu (*cis*-sRNA). Tři ukázky takovýchto detekovaných sRNA jsou na Obrázku 3.6 v sekci (A).

ANNOgesic však počítá s tím, že se sRNA mohou nacházet i v oblasti 3‘ a 5‘ UTR konců anebo přímo ve struktuře určitého genu. I tyto potenciální sRNA musí obsahovat svůj promotor a vyskytovat se v genu s vyšším pokrytím než samotný gen (v pokrytí genu musí být prudký pokles, nárůst, nebo obojí). Detekce sRNA v oblasti 3‘ a 5‘ UTR je vyobrazena na Obrázku 3.6 v sekci (B) a detekce v rámci genu je vyobrazena v sekci (C).



Obrázek 3.6 Způsoby detekce sRNA u ANNOgesic: (A) detekce *trans*-sRNA a *cis*-sRNA; (B) detekce v oblasti 3‘ UTR a 5‘ UTR, (C) detekce ve struktuře genu; upraveno [11].

3.4 APERO

Jedním z novějších přístupů detekce sRNA v bakteriích je metoda APERO. Ta byla poprvé uvedena v roce 2019 [12] a je volně dostupná v programovacím prostředí R (využívající balíčky Rsamtools, Reshape2 a Snowfall) nebo přes platformu Galaxy.

APERO, jako všechny ostatní výpočetní nástroje, využívá pro detekci data získaná pomocí RNA-Seq metod. Je primárně určen pro zpracování paired-end dat, která byla získána z krátkých nefragmentovaných RNA. Není tedy vhodný pro všechny RNA-Seq data, což je jeho hlavní limitací. Také se od většiny z dostupných nástrojů liší způsobem detekce sRNA. Detekce není založena na zjištění pokrytí namapovaných čtení k referenci, ale na detekci stabilních 5‘ konců sRNA a následném rozšíření těchto nalezených úseků ve směru 5‘→ 3‘.

3.4.1 Princip detekce APERO

Vstupem do algoritmu APERO je soubor s namapovanými čteními ve formátu BAM. Tato data jsou vyfiltrována tak, aby obsahovala pouze paired-end čtení se správnou orientací a umístěním vůči sobě. Algoritmus je rozdělen do dvou hlavních částí. Jak už bylo zmíněno výše, v první části se zaměřuje na detekci 5‘ konců mapovaných čtení a v druhé části pak dochází k iterativnímu prodlužování ve směru 5‘ → 3‘.

Na Obrázku 3.7 je znázorněno schéma pracovního postupu APERO. Sekce (A) znázorňuje první část algoritmu. Detekce 5‘ konců je založena na nalezení oblastí, které mají vysoký výskyt začátků čtení a jsou tedy označovány jako tzv. počáteční oblasti. Maximální šířku této počáteční oblasti w_{max} může uživatel libovolně nastavit. Algoritmus následně analyzuje počáteční oblasti o šířce $1 \leq w \leq w_{max}$ a pro konkrétní pozice v genomu a konkrétní počáteční oblast vyhodnocuje pokrytí vůči sousedním oblastem (toto pokrytí je vypočteno pouze z počátečních pozic čtení). Za sousední oblasti jsou uvažovány ty oblasti, které sousedí s počáteční oblastí a jsou f dlouhé ($1 < f < d$). Vzdálenost d určuje maximální vzdálenost sousední oblasti od původní počáteční oblasti. Také hodnotu maximální vzdálenosti d si volí uživatel sám a ovlivňuje tím prostorovou rozlišovací schopnost algoritmu (v jaké nejmenší vzdálenosti od sebe se mohou vyskytovat dvě různé sRNA). Pokud je v sousedství několik počátečních oblastí se stejnou mírou pokrytí, jejich šířka se zužuje a je vybrána ta počáteční oblast, která je nejužší a přitom má stejnou hodnotu pokrytí. Výstupem této části je soubor obsahující seznam centrálních poloh počátečních oblastí, jejich poloviční délky w, informaci, ze kterého vlákna pochází, a počet čtení, které v této oblasti začínají.

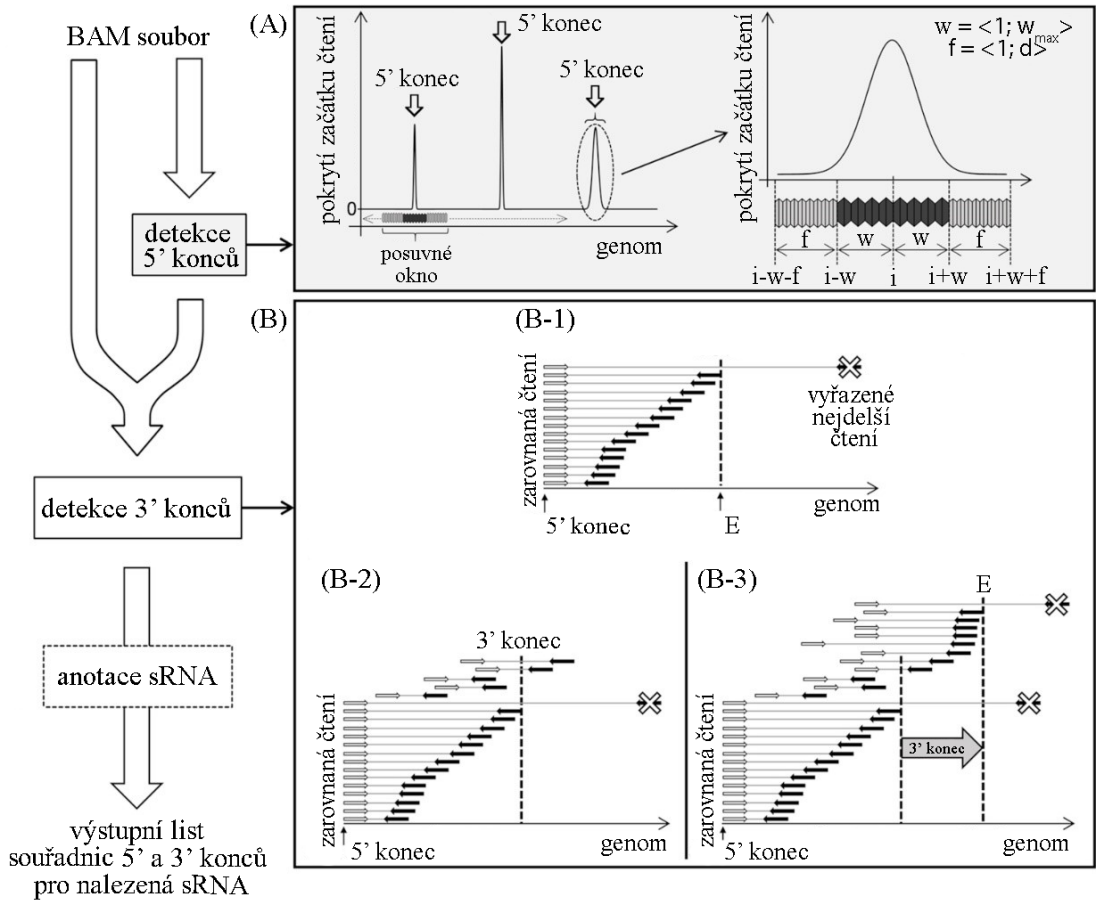
Na Obrázku 3.7 v sekci (B) je znázorněna druhá část algoritmu. Zde dochází k iterativnímu prodlužování počátečních oblastí (5‘ konců transkriptů) ve směru 5‘ → 3‘. Nejprve je vyfiltrováno 1 % nejdelších čtení spadajících do dané startovací oblasti. Ze zbylých čtení je vybráno nejdelší čtení a jeho 3‘ konec je nazván jako potenciální

3‘ konec, označen jako *E* (B-1). Následně je vypočten poměr pokrytí $F_{start,E}$ pro potenciální konec *E*:

$$F_{start,E} = \frac{C_E}{\frac{C_{start,E}}{L_{start,E}}}, \tag{0.2}$$

kde C_E je pokrytí na pozici *E*, $C_{start,E}$ je počet čtení mezi počáteční oblastí a koncem *E*, $L_{start,E}$ je vzdálenost mezi počáteční oblastí a koncem *E*. Jestliže hodnota $F_{start,E}$ je menší než F_{min}, je pozice *E* nazvána jako 3‘ konec (B-2). V opačném případě dochází k prodloužení analyzovaného transkriptu a určení nového potenciálního konce *E* (B-3). Transkript je prodloužen o čtení, která dosahovalá alespoň částečného překryvu s původním transkriptem. Celý postup ze sekce (B) je iterativně opakován do doby, než $F_{start,E}$ není menší než F_{min}. Hodnotu F_{min} si uživatel nastavuje sám a zásadně tím ovlivňuje délku výsledných sRNA.

Takto jsou určeny souřadnice 5‘ konců a 3‘ konců všech potenciálních sRNA, které jsou zapsány do výstupního souboru v datovém formátu TXT. Tento soubor obsahuje podrobné informace o nově detekovaných transkriptech (3’UTR, 5’UTR, sRNA včetně rozlišené antisense RNA) pomocí algoritmu APERO.



Obrázek 3.7 Zjednodušené schéma postupu analýzy programu APERO; upraveno [12].

3.5 Baerhunter

Dalším z nových přístupů detekce sRNA a UTR u bakterií je Baerhunter [13] uvedený taktéž v roce 2019. Jedná se o volně dostupný algoritmus implementovaný v programovacím prostředí R. Jako předchozí metody, i Baerhunter vyžaduje na svém vstupu zarovnaná čtení vůči referenci v datovém formátu BAM a soubor s anotací v datovém formátu GFF3. Baerhunter je vhodný pro stranded data, jak pro single-end, tak i pro paired-end data, což umožňuje širší možnost využití. Samotní autoři uvádějí, že Baerhunter není nástroj pro predikci přesných počátků a konců sRNA, ale že se jedná o detekci založenou na hledání potenciálních výskytů sRNA. Přesné pozice by měly být určeny experimentálně. Další výhodou algoritmu Baerhunter je možnost analyzovat více BAM souborů zároveň, což umožňuje další možnosti zpracování, jako například diferenciální analýzu genové exprese.

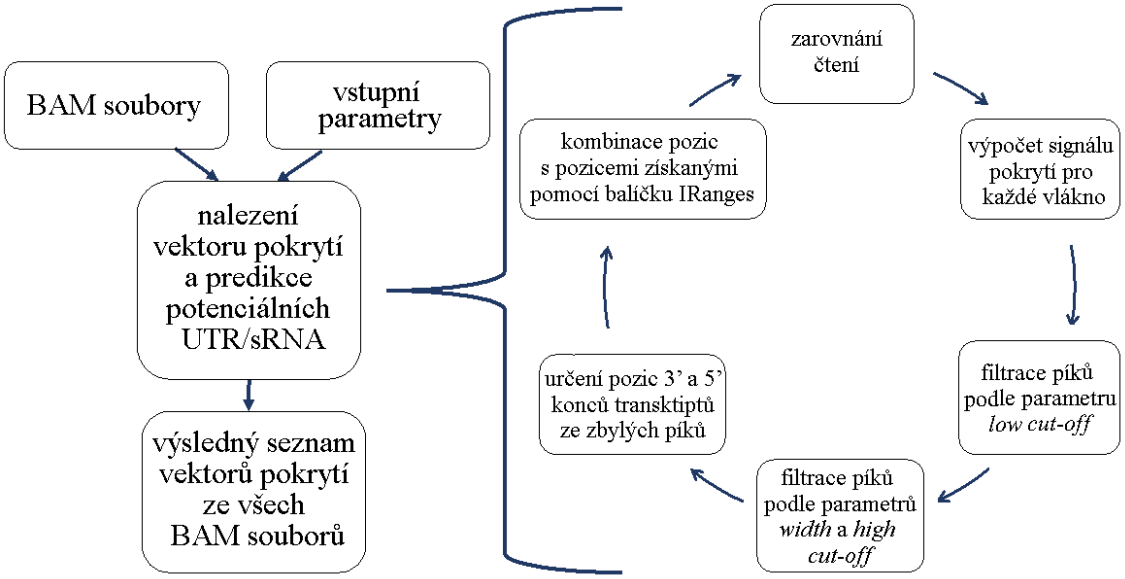
3.5.1 Princip detekce Baerhunter

Celý algoritmus začíná načtením všech analyzovaných BAM souborů obsahujících zarovnaná čtení vůči referenci. Nejprve dochází k výpočtu signálu pokrytí pro pozitivní i negativní vlákno. Následně dochází k vytvoření vektorů pokrytí. Vektor pokrytí představuje řadu po sobě jdoucích genomických intervalů s konstantním pokrytím. Každý z těchto vektorů pokrytí je zkontrolován na základě minimálního pokrytí, které musí obsahovat (*low-cut-off*). Takto vybrané vektory určují píky v signálu pokrytí. Následně jsou tyto detekované píky filtrovány na základě minimální délky (*width*) a konstantního pokrytí (*high-cut-off*). Prahové hodnoty *low-cut-off, width* a *high-cut-off* uživatel nastavuje sám při spouštění algoritmu a hrají v detekci významnou roli.

Získané píky napříč celým genomem se kombinují s píky obdrženými pomocí balíčku IRanges R [66]. Celý proces se zopakuje pro všechny vstupní BAM soubory. Výstupem je soubor sloučených vektorů pokrytí z jednotlivých BAM souborů. Tento proces je vyobrazen na Obrázku 3.8.

Jakmile je získán seznam všech vektorů pokrytí, dochází k určení a uložení jejich pozic v genomu. Následně jsou načteny informace z anotace genomu ve formátu GFF3 a dochází k porovnání anotovaných úseků (kromě anotací již detekovaných sRNA/UTR u daného organizmu) s nalezenými pozicemi vektorů pokrytí. Detekované vektory pokrytí, které se v plné délce překrývají s anotovaným úsekem, jsou odstraněny. Pokud se detekované vektory pokrytí překrývají s anotovaným genem pouze částečně, jejich části, které se s genem nepřekrývají, jsou izolovány a na základě délky jsou označeny jako potenciální UTR. Vektory pokrytí, které se s anotovanými úseky vůbec nepřekrývají, jsou označeny jako potenciální sRNA.

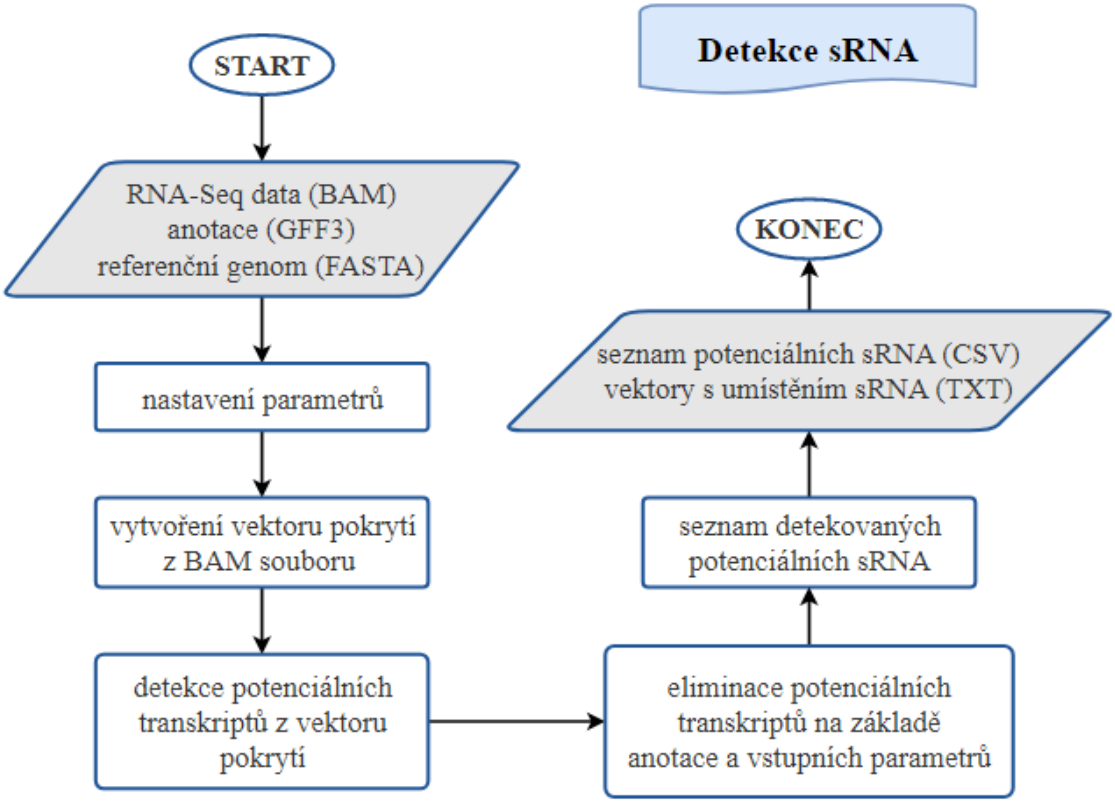
Výsledné potenciální UTR a sRNA jsou přidány ke vstupní anotaci a uloženy do nové anotace v souladu se standardy datového formátu GFF3. Nově nalezené transkripty mohou být ještě filtrovány na základě úrovně jejich exprese pomocí hodnoty TPM (počet transkriptů na milion bází).



Obrázek 3.8 Schéma detekce sRNA a UTR u algoritmu Baerhunter; upraveno [13].

4. NÁVRH A IMPLEMENTACE NÁSTROJE PRO PREDIKCI SRNA

Pro tuto diplomovou práci byl navrhnut uživatelsky přívětivý nástroj pro detekci potencionálních sRNA z RNA-Seq dat (SEARCHsRNA). Nástroj byl implementován v programovacím prostředí R a pro svou funkci využívá několik volně dostupných balíčků funkcí. Mezi tyto balíčky patří balíček Rsamtools (balíček spadající do projektu Bioconductor) [67], balíček seqinr [68], který umožňuje správu sekvenačních dat a balíček dplyr [69], který umožňuje zjednodušenou manipulaci s daty. Celý nástroj je volně dostupný na: https://github.com/xpomyk03/SEARCHsRNA.git. Zjednodušený pracovní postup nástroje SEARCHsRNA je nastíněn na Obrázku 4.1 a detailněji popsán v kapitole 4.3.



Obrázek 4.1 Zjednodušené schéma principu detekce nástroje SEARCHsRNA pro nalezení potenciálních sRNA.

4.1 Použitá data

Pro potřeby této práce byla využita data z RNA-Seq pro bakterii *Vibrio atlanticus* LGP32 a pro bakterii *Clostridium beijerinckii* NRRL B-598. Data z obou bakterií byla použita v průběhu návrhu, implementace a kontroly funkčnosti jednotlivých bloků

nástroje SEARCHsRNA. Na závěr byl tento nástroj otestován pouze na datech bakterie *Vibrio atlanticus* LGP32 (viz kapitola 5).

4.1.1 *Vibrio atlanticus* LGP32

Bakterie *Vibrio atlanticus* (dříve nazývaná *Vibrio splendidus*) je gramnegativní bakterie spadající do rodu *Vibrio.* Bakterie tohoto rodu se nachází v mořské vodě, v mořských sedimentech na pobřežích a v ústí řek. Některé z nich mají patogenní účinky na měkkýše, ploutvonožce i savce [70; 71]. Konkrétně u kmene *Vibrio atlanticus* LGP32 byly prokázány patogenní účinky na ústřici *Crassostrea gigas*, které mohou vést až k jejímu úhynu [72].

Jako referenční genom byl využit genom bakterie *Vibrio atlanticus* složený ze dvou chromozomů. První z těchto chromozomů dosahuje délky 3 299 303 pb a jeho referenční sekvence je volně dostupná z databáze GenBank pod přístupovým kódem NC_011753.2. Druhý chromozom dosahuje délky 1 675 515 pb a jeho referenční sekvence je volně dostupná z databáze GenBank pod přístupovým kódem NC_011744.2.

RNA-Seq data bakterie *Vibrio atlanticus* LGP32 byla původně publikována a využita při testování predikčního nástroje DETR'PROK [10]. Sekvenační běh je dostupný pod kódem SRR836420. Čtení o délce 38 pb, byla získána sekvenační metodou Illumina GA-IIX. Tato čtení byla předzpracována – provedením kontroly kvality a odstřižení adapterů. Čtení byla namapována k referenci pomocí nástroje Bowtie, který je dostupný na webové platformě Galaxy. Výsledný dataset obsahuje dva soubory v datovém formátu BAM, každý pro jeden chromozom. Soubory byly staženy z volně dostupných doplňkových dat pro predikční nástroj DETR'PROK.

4.1.2 *Clostridium beijerinckii* NRRL B-598

Bakterie *Clostridium beijerinckii* NRRL B-598 je aerobní bakterie spadající do rodu *Clostridium* [73]. Jedná se o bakterii bez patogenních účinků, která má schopnost produkovat butanol.

Jako referenční genom pro bakterii *Clostridium beijerinckii* NRRL B-598 byl využit genom o délce 6 186 993 pb, který je volně dostupný z databáze GenBank pod přístupovým kódem CP011966.3.

Pro tuto práci byla využita RNA-Seq data bakterie *Clostridium beijerinckii* NRRL B-598, která se v databázi NCBI Sequence Read Archive (SRA) vyskytují pod přístupovým kódem SRP033480. Čtení, o délce 75 pb, byla získána pomocí sekvenační platformy Illumina NextSeq. Získaná čtení byla opět předzpracována – došlo k odstřižení adapterů a odstranění zbytkové rRNA. Čtení byla následně namapována k referenčnímu genomu a komprimována do BAM formátu.

4.2 Vstupní data a parametry

Vstupními daty pro nástroj SEARCHsRNA jsou referenční sekvence pro daný organizmus v datovém formátu FASTA, anotace pro daný organizmus v datovém formátu GFF3 a stranded nebo reversly stranded RNA-Seq data v datovém formátu BAM. Sekvenačních dat v datovém formátu BAM může být během jednoho spuštění programu použito více, důležité je, aby obsahovala unikátní název a byla mapována ke stejnému referenčnímu genomu. Je tedy možné pomocí tohoto nástroje během jednoho spuštění hledat sRNA z více RNA-Seq dat (z více replikátů).

V nástroji je použito několik volitelných parametrů, které uživatel může nastavit. Nastavení těchto parametrů probíhá při spuštění funkce `search_sRNA()` (viz kapitola 4.3.2). Prvním parametrem je parametr pod názvem `type_of_data` (datový typ `str`). Tento parametr udává, zda se jedná o stranded (`type_of_data = 'Stranded'`) nebo reversly stranded data (`type_of_data = 'ReverslyStranded'`). Defaultně je tento parametr nastaven na hodnotu pro stranded data.

Následující tři parametry ovlivňují chování samotné detekce potenciálních sRNA a defaultně jsou všechny tyto parametry nastaveny na proměnnou `NULL`. K finálnímu automatickému nastavení těchto tří parametrů dochází až během samotné detekce. Parametr `threshold_coverage_steepness` (datový typ `int`) ovlivňuje detekci 5‘ a 3‘ konců všech potenciálních transkriptů obsažených v genomu. Uživatel nastavuje tzv. míru změny v pokrytí. Tato míra změny pokrytí je udávána v počtu čteních, o která se musí pokrytí změnit, aby daná oblast byla považována za 5‘ a 3‘ konec transkriptu. Např. pokud bude za tento parametr nastavena hodnota 10, jako 5‘ a 3‘ konce budou považovány pouze oblasti, u kterých dochází ke změně v pokrytí o 10 čtení. Doporučený rozsah tohoto parametru je od 2 do 10, pro experimentální účely i vyšší.

Parametr `threshold_coverage_min` (datový typ `int`) udává hodnotu počtu namapovaných čtení, která musí minimálně dané místo obsahovat, aby bylo možné jej uvažovat jako 5‘ a 3‘ konec transkriptů obsažených v genomu. To znamená např., že pokud bude tento parametr nastaven na hodnotu 20 a bude nalezeno místo s pokrytím 21x splňující parametr `threshold_coverage_min`, bude toto místo uvažováno jako potenciální 5‘ nebo 3‘ konec transkriptu genomu. Jestliže by bylo pokrytí menší než zvolený parametr, jako 5’nebo 3‘ konec by toto místo uvažováno nebylo. Doporučený rozsah tohoto parametru je od 2 do 5, pro experimentální účely i vyšší.

Parametr `threshold_gap_transcripts` (datový typ `int`) udává minimální rozestup dvou detekovaných transkriptů [pb], aby byly uvažovány jako samostatné transkripty. Doporučený rozsah tohoto parametru je od 20 pb, kdy by vyšší hodnoty měly být nastaveny pro BAM soubory s nižším průměrným pokrytím. Minimální hodnota 20 pb je určena kvůli tomu, že při detekci dochází ke spojení dvou transkriptů vždy, pokud je jejich vzdálenost menší než uvedená hodnota 20 pb.

Poslední dva parametry upravují až výsledný seznam sRNA obdržený po ukončení detekce. Parametrem `threshold_coverage_sRNA` (datový typ `int`) lze určit, jaké minimální průměrné pokrytí musí mít výsledná detekovaná sRNA. Defaultně je tento parametr nastaven na hodnotu 0, a pokud nedojde k jeho nastavení, jsou do výstupního seznamu uloženy všechny detekované sRNA bez ohledu na jejich průměrné pokrytí. Parametr `min_length_of_sRNA` (datový typ `int`) určuje, jakou minimální délku [pb] musí výsledná sRNA mít, aby byla uložena do výstupního seznamu detekovaných sRNA. Defaultně je tento parametr nastaven na 40 pb vzhledem k očekávaným minimálním délkám sRNA (viz kapitola 1).

4.3 Princip detekce a implementace

Jak už bylo zmíněno, nástroj byl implementován v programovacím prostředí R a byly vytvořeny dva skripty. Pro úspěšné spuštění nástroje je vyžadováno pomocí funkce `source()` načíst skript `functions`, který obsahuje funkce použité při detekci sRNA nástrojem SEARCHsRNA.

Skript `example` slouží jako ukázka pro spuštění nástroje SEARCHsRNA. Tento ukázkový skript `example` a jednotlivé funkce ze skriptu `functions` jsou detailněji popsány v následujících podkapitolách 4.3.1-4.3.8.

4.3.1 Skript `example`

Skript `example` je určený jako ukázka pro úspěšné spuštění detekce sRNA pomocí nástroje SEARCHsRNA. Nachází se zde sekce pro načtení potřebných knihoven (kapitola 4) pomocí funkce `library()` a skriptu `functions`, který obsahuje funkce nástroje SEARCHsRNA. Ten je načten pomocí funkce `source()`. Na závěr tohoto skriptu je zobrazen příklad spuštění funkce `search_sRNA()` (kapitola 4.3.2) s nastavením všech jejích parametrů (viz kapitola 4.2) manuálně.

4.3.2 Funkce `search_sRNA()`

Jediným povinným vstupním parametrem je parametr `path_of_files` (datový typ `str`) obsahující cestu ke složce, ve které se nachází všechna potřebná vstupní data: referenční sekvence ve formátu FASTA, anotace ve formátu GFF3 a BAM soubor/y. Dalšími, již volitelnými parametry, jsou parametry detailněji popsané v kapitole 4.2.

Ve funkci nejprve dochází na základě parametru `path_of_files` k načtení referenční sekvence do proměnné `fasta` (datový typ `list`), anotace do proměnné `gff` (datový typ `data.frame`) a uložení všech názvů dostupných souborů datového formátu BAM do proměnné `bamfiles` (datový typ `character`). Na základě proměnné `fasta` je poté určena proměnná `length_of_genome` obsahující informaci o délce genomu.

Následně dochází ke spuštění funkce `get_annotation()` (kapitola 4.3.4), díky níž jsou obdrženy pozice anotovaných genů v genomu rozdělené podle vlákna, na kterém se geny vyskytují.

V rámci funkce následuje for cyklus, jehož počet opakování odpovídá délce proměnné `bamfiles` a tedy odpovídá počtu BAM souborů, na kterých je detekce sRNA provedena. V rámci tohoto for cyklu je nejprve volána funkce `preparing_signals_from_reads()` (kapitola 4.3.5), na jejímž výstupu je obdržena datová struktura `list`, obsahující signál pokrytí pro pozitivní i negativní vlákno (datový typ `num`). Signálem pokrytí se rozumí vektor o délce genomu, který na jednotlivých pozicích obsahuje informaci o počtu namapovaných čtení na danou pozici. Z těchto signálů je poté vypočítáno průměrné pokrytí `coverage_signal` pro daný BAM soubor vzhledem k oběma vláknům. Průměrné pokrytí `coverage_signal` bylo vypočteno podle vzorce:

$$coverage\_signal = \frac{počet\ čtení * průměrná\ délka\ čtení}{2 * délka\ genomu}. \tag{0.1}$$

Informace o průměrném pokrytí je posléze použita při automatickém nastavení parametrů `threshold_coverage_steepness`, `threshold_coverage_min` a `threshold_gap_transcripts` v případě, kdy nejsou tyto parametry nastaveny uživatelem při spuštění funkce `search_sRNA()`. K nastavení těchto parametrů jsou využity po částech lineární funkce, jejichž výstup závisí právě na proměnné `coverage_signal`. Tyto po částech lineární funkce byly vytvořeny na základě detekce sRNA na různých souborech BAM s odlišnou hodnotou `coverage_signal`. V rovnici (4.2) je ukázána lineárně lomená funkce pro automatické nastavení parametru `threshold_gap_transcripts`.

$$\text{threshold\_gap\_transcripts}(coverage\_signal) = \begin{cases} 50 & pro\ \text{coverage\_signal} < 10; \\ 20 & pro\ \text{coverage\_signal} > 100; \\ \frac{2 * \text{coverage\_signal}}{45} + \frac{5}{9} & jinak. \end{cases} \tag{0.2}$$

Poté je dvakrát spuštěna funkce `search_transcripts()` (kapitola 4.3.6), ve které jsou nacházeny potenciální transkripty sRNA, nejprve pro pozitivní a posléze i pro negativní vlákno.

Seznamy detekovaných transkriptů sRNA pro pozitivní a negativní vlákno jsou poté využity funkcí `exporting_signals_TXT()` (kapitola 4.3.7) a funkcí `exporting_CSV()` (kapitola 4.3.8), které poskytují výstupní soubory nástroje SEARCHsRNA – tedy tabulku CSV obsahující seznam detekovaných sRNA a dva textové soubory obsahující informace o umístění těchto sRNA na pozitivním a negativním vlákně.

4.3.3 Funkce `readBAM()`

Funkce `readBAM()` je funkce sloužící pro uložení souborů BAM do datového formátu `DFrame`, se kterým je možné rychle a jednoduše pracovat (indexace, vyhledávání, …). Vstupní proměnnou je pouze název BAM souboru, který je uložen v proměnné `bamfiles`. Funkce je dostupná na: https://gist.github.com/SamBuckberry/9914246.

4.3.4 Funkce `get_annotation()`

Funkce `get_annotation()` slouží k získání pozic všech anotovaných genů vyskytujících se v organizmu. Vstupní proměnnou je proměnná `gff` z funkce `search_sRNA()`. Geny jsou v rámci funkce `get_annotation()` roztříděny na pozitivní a negativní vlákno. Na výstupu této funkce je tedy obdržen `list` se čtyřmi vektory obsahující pozice 5‘ a 3‘ konců všech anotovaných genů pro jednotlivé vlákna.

4.3.5 Funkce `preparing_signals_from_reads()`

Vstupními proměnnými této funkce jsou proměnné `length_of_genome`, `bamfiles` a parametr `type_of_data` (viz kapitola 4.3.2). Výstupem funkce je `list` obsahující dva vektory signálu pokrytí.

Ve funkci `preparing_signals_from_reads()` je pomocí funkce `readBAM()` (kapitola 4.3.3) načten BAM soubor, který je následně pročištěn o méně kvalitní čtení a o čtení, která nemají vhodnou délku. Nejprve jsou tedy odstraněna všechna čtení, která mají kvalitu nižší než Phred skóre rovno 255. K tomuto odstranění dochází z důvodu eliminování špatně namapovaných čtení. Následně jsou ze seznamu odstraněna ta čtení, která jsou kratší jak 20 pb a ta čtení, která jsou delší než zvolená délka čtení pro dané sekvenování.

Takto pročištěný seznam všech vyhovujících čtení je posléze rozdělen do dvou seznamů podle vlákna, ze kterého pochází. V této sekci je využit parametr `type_of_data`, který informuje o typu sekvenačních dat uložených v BAM souboru – zda se jedná o stranded nebo reversly-stranded data, což určuje, jak jsou čtení v BAM souborech orientovaná.

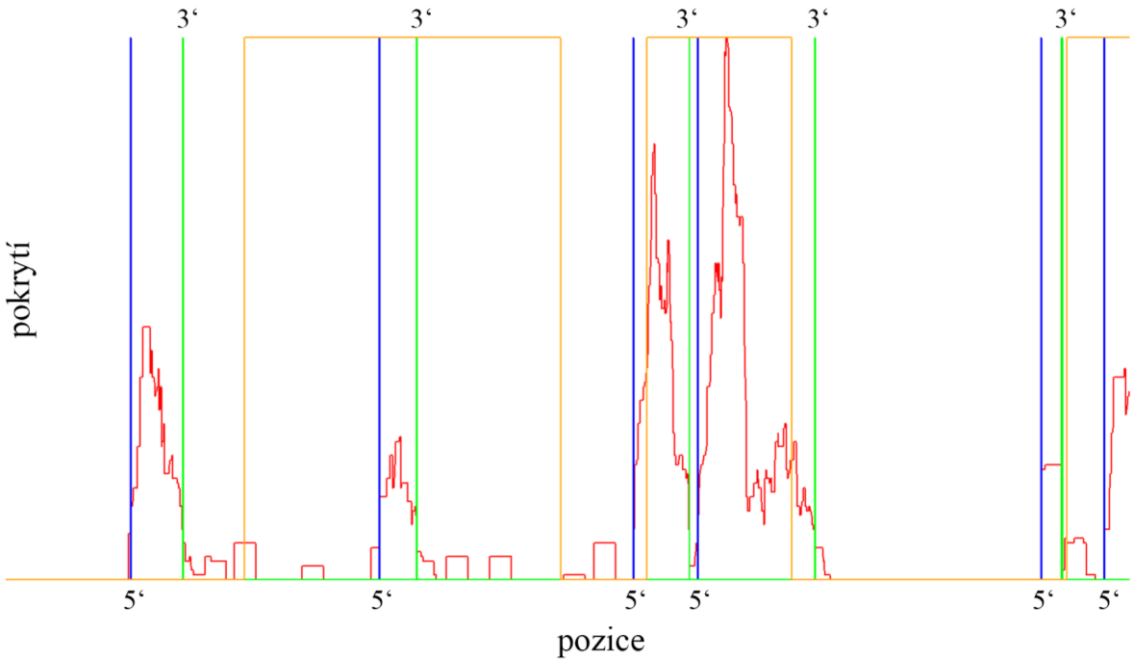
Z těchto seznamů jsou vytvořeny dva vektory o délce odpovídající délce genomu. Do každého z vektorů jsou zaznamenávána čtení z jednotlivých seznamů tak, že na pozice, kam bylo čtení namapováno, se do vektoru přičítá hodnota 1. Tím jsou získány tzv. signály pokrytí pro pozitivní i negativní vlákno, které jsou základem pro následnou predikci potenciálních transkriptů.

4.3.6 Funkce `search_transcripts()`

Funkce `search_transcripts()` je nejdůležitější částí celého nástroje, která ze signálu pokrytí (kapitola 4.3.5) detekuje potenciální sRNA. Vstupními proměnnými je již zmíněný signál pokrytí pro jedno vlákno, délka genomu, vektory obdržené z funkce `get_annotation()` pro stejné vlákno (kapitola 4.3.4), průměrné pokrytí

`coverage_signal` a parametry ovlivňující detekci (`threshold_coverage_min`, `threshold_coverage_steepness`, `threshold_gap_transcripts`, `min_length_of_sRNA`, `threshold_coverage_sRNA`).

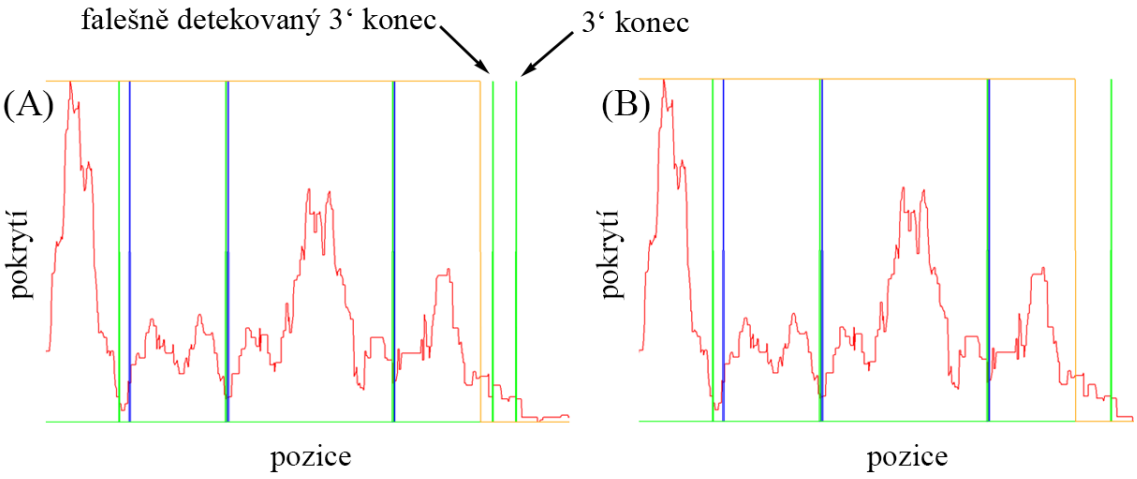
Nejprve jsou hledány potenciální 5‘ a 3‘ konce všech transkriptů nacházející se v signálu pokrytí. Konce jsou hledány v signálu pokrytí tak, že jsou nacházeny všechny oblasti, které mají minimální požadované pokrytí podle parametru `threshold_coverage_min` a změna pokrytí (tj. strmost v rozsahu 30 pb) minimálně dosahuje hodnoty parametru `threshold_coverage_steepness`. Ukázku takto nalezených 5‘ a 3‘ konců lze vidět na Obrázku 4.2, kde modrou barvou jsou značené pozice detekovaných 5‘ konců, zeleně jsou znázorněny pozice 3‘ konců, červeně je zobrazený signál pokrytí a žlutě je vyznačena oblast, kde se nachází některý z anotovaných genů daného vlákna. Tento a zbylé obrázky v této podkapitole byly získány během spuštění nástroje SEARCHsRNA na datech pro bakterii *Vibrio atlanticus* LGP32, přesněji pro chromozom NC_011753.2.



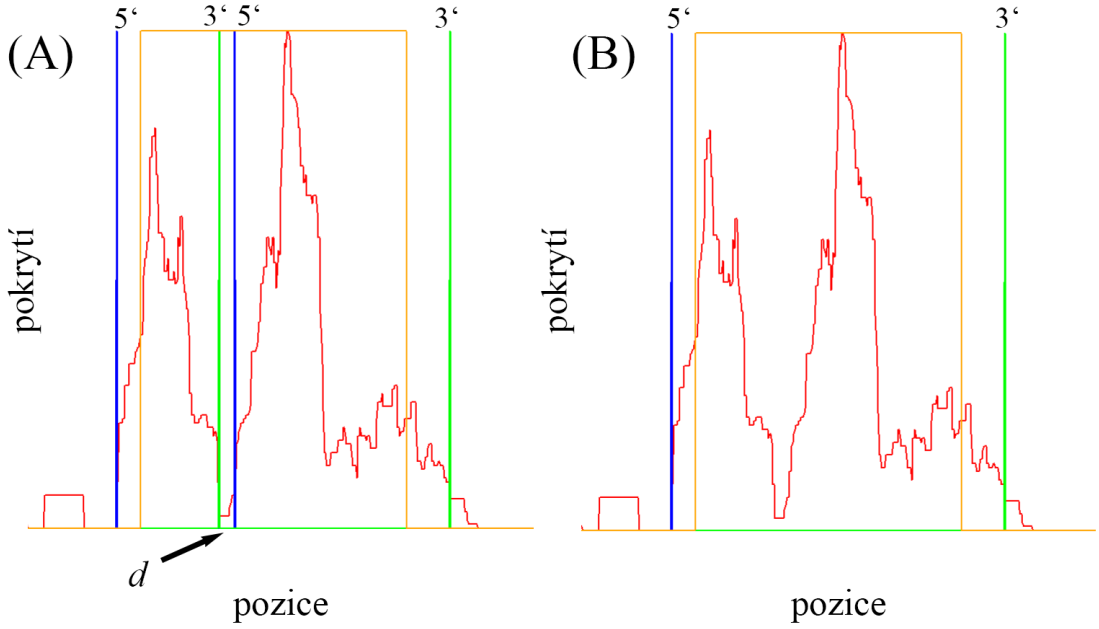
Obrázek 4.2 Detekce sRNA nástrojem SEARCHsRNA pro bakterii *Vibrio atlanticus* LGP32 (chromozom NC_011753.2): detekce 5‘ (modrá) a 3‘ (zelená) konců potenciálních transkriptů ze signálu pokrytí (červená).

Při této detekci 5‘ a 3‘ konců dochází také k detekování falešných konců, jak je vidět na Obrázku 4.3 v sekci (A). Kvůli těmto situacím jsou detekované konce kontrolovány a falešné konce jsou eliminovány. Eliminace probíhá na základě poklesu pokrytí mezi 5‘ a 3‘ koncem. Tím jsou eliminovány falešné konce, viz Obrázek 4.3 sekce (B). Eliminací falešných konců jsou získány potenciální pozice 3‘ a 5‘ konců všech transkriptů v datech. Pozice konců těchto transkriptů jsou posléze rozšiřovány

oběma směry, dokud nedojde k poklesu pokrytí v signálu pokrytí pod stanovený parametr `threshold_coverage_min`.



Obrázek 4.3 Detekce sRNA nástrojem SEARCHsRNA pro bakterii *Vibrio atlanticus* LGP32 (chromozom NC_011753.2): eliminace falešně detekovaných 5‘ (modrá) a 3‘ (zelená) konců transkriptů ze signálu pokrytí (červená): (A) v signálu se nachází falešně detekovaný 3‘ konec; (B) po eliminaci falešně detekovaných konců.

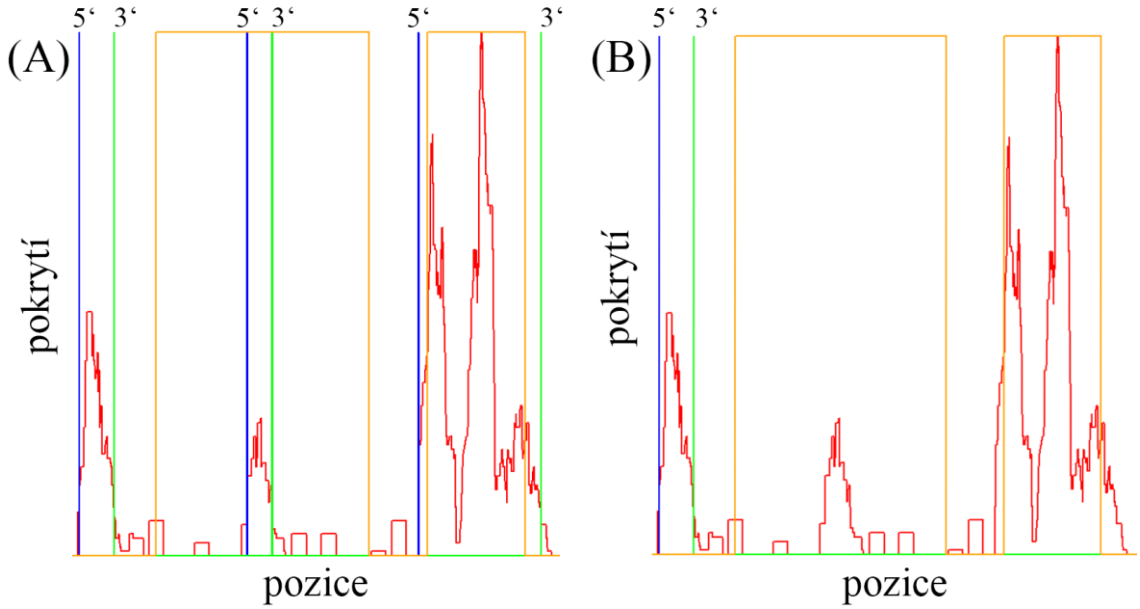


Obrázek 4.4 Detekce sRNA nástrojem SEARCHsRNA pro bakterii *Vibrio atlanticus* LGP32 (chromozom NC_011753.2): spojování blízkých transkriptů: (A) dva detekované transkripty se vzdáleností *d*; (B) po spojení transkriptů.

Takto prodloužené transkripty, které od sebe leží v těsné blízkosti, jsou posléze spojeny. Ke spojení dvou transkriptů může dojít za dvou podmínek. První z možností, kdy dojde ke spojení dvou transkriptů, je ta, když vzdálenost, kterou mezi sebou dva transkripty mají, je menší nebo rovna 20 pb. Pokud je splněna tato první podmínka, ke

spojení dochází za všech okolností. U druhé z možností je využit parametr `threshold_gap_transcripts`. Ten udává maximální vzdálenost dvou transkriptů, které jsou uvažovány Zda dojde ke spojení v tomto případě, je ovlivněno pokrytím těchto transkriptů. Jestliže je jejich průměrné pokrytí podobné (pokrytí se liší o hodnotu, která je menší než parametr `coverage_signal`), ke spojení dochází. Pokud je rozdíl jejich pokrytí vyšší než parametr `coverage_signal`, ke spojení nedochází i když splňují podmínku, že jsou od sebe vzdáleny méně než `threshold_gap_transcripts`. Na Obrázku 4.4 v sekci (A) lze vidět dva transkripty se vzdáleností *d*, která je menší než parametr `threshold_gap_transcripts`, pokrytí těchto transkriptů je podobné (dosahuje rozdílu pokrytí menší jak parametr `coverage_signal`), tudíž dochází ke spojení, viz sekce (B). Na Obrázku 4.4 v sekci (B) je také patrné, vzhledem k anotaci (žlutý obdélník), že se na dané pozici nachází gen a detekovaný transkript je tedy správný.

Tímto procesem je získán seznam pozic nalezených transkriptů. Tento seznam obsahuje transkripty reprezentující jak geny, tak sRNA. Seznam je zkontrolován a transkripty menší než délka udaná parametrem `min_length_of_sRNA` jsou odstraněny (pokud uživatel nenastaví jinak, jsou odstraněny všechny transkripty menší než 40 pb). Posléze dochází k eliminaci všech transkriptů, které se nacházejí na stejných pozicích jako některý z anotovaných genů. Tím je obdržen seznam zbylých transkriptů, který by měl odpovídat detekovaným potenciálním sRNA. Na Obrázku 4.4 lze vidět, že transkripty odpovídající některému z referenčních genů jsou z výsledného seznamu odstraněny a zůstávají pouze transkripty, které odpovídají potenciálním sRNA.



Obrázek 4.5 Detekce sRNA nástrojem SEARCHsRNA pro bakterii *Vibrio atlanticus* LGP32 (chromozom NC_011753.2): odstranění transkriptů náležící anotovaným genům: (A) v seznamu detekovaných transkriptů jsou i anotované geny; (B) po odstranění transkriptů odpovídajícím anotovaným genům.

Na závěr dochází k eliminaci transkriptů sRNA, které mají pokrytí nižší než zvolený parametr `threshold_coverage_sRNA`. Tento krok závisí na nastavení uživatele. Jestliže k nastavení uživatelem nedojde, jsou na výstupu funkce `search_transcripts` ve výsledném seznamu uloženy všechny potenciální sRNA bez ohledu na jejich finální pokrytí.

4.3.7 Funkce `exporting_signals_TXT()`

Funkce `exporting_signals_TXT()` slouží k vytvoření souboru ve formátu TXT, který obsahuje informace o umístění sRNA v genomu. Vstupními proměnnými jsou název BAM souboru v proměnné `bamfiles`, textový řetězec obsahující informaci, zda se jedná o pozitivní či negativní vlákno, proměnná `length_of_genome` a seznam detekovaných sRNA pro dané vlákno.

V rámci funkce `exporting_signals_TXT()` jsou informace získané funkcí `search_transcripts()` (kapitola 4.3.6) o detekovaných potenciálních sRNA využity k vytvoření vektoru obsahujícímu informaci o umístění těchto sRNA. Ve vektoru jsou na pozicích výskytu sRNA po její celé délce uloženy hodnoty 1 a na pozicích, kde se sRNA nevyskytuje, jsou uloženy hodnoty 0. Takto vytvořený vektor je exportovaný v datovém formátu TXT, což je cílem této funkce.

4.3.8 Funkce `exporting_CSV()`

Funkce slouží k doplnění informací o detekovaných sRNA a následnému exportování těchto informací o detekovaných potenciálních sRNA v datovém formátu CSV. Vstupními proměnnými jsou název BAM souboru v proměnné `bamfiles`, proměnná `gff` obsahující uloženou anotaci a výstupní soubory z funkce `search_transcripts` (kapitola 4.3.6) pro pozitivní i negativní vlákno.

V rámci funkce je tvořena tabulka obsahující informace o potenciálních sRNA – umístění (start a stop pozice), délka, vlákno, na kterém se sRNA nachází (‘+‘/‘-‘), pokrytí pro sRNA, typ (*cis-/trans*-sRNA) a informaci, který gen s vysokou pravděpodobností ovlivňuje. Poslední z informací je uvedena pouze pro *cis*-sRNA, neboť určení cílů účinku pro *trans*-sRNA je problematické a není účelem tohoto nástroje.

Většina z ukládaných informací o sRNA je obdržena již ze vstupních proměnných, kromě informace o typu sRNA a genovém cíli pro *cis*-sRNA. Tyto informace jsou určeny až v rámci funkce `exporting_CSV()`. Typ sRNA je zjištěn na základě znalosti anotace, kdy sRNA, která jsou komplementární k některému z anotovaných genů, byla prohlášena za *cis*-sRNA. Small RNA, která se nacházela v mezigenovém prostoru a byla komplementární vůči jinému mezigenovému prostoru, byla prohlášena za *trans*-sRNA. Genový cíl *cis*-sRNA byl také určen na základě anotace a jednalo se o gen, který byl vůči detekované sRNA alespoň z určité části komplementární.

Všechny tyto informace jsou uloženy do tabulky, která je na závěru funkce exportována ve formátu CSV s názvem odpovídajícím názvu použitého BAM souboru.

4.4 Výstupní data

Počet výstupních souborů, které jsou vytvořeny pomocí zde navrženého nástroje SEARCHsRNA, je celkem tři – pro jeden vstupní soubor BAM. Prvním z nich je tabulka v datovém formátu CSV obsahující informace o detekovaných sRNA – začátek a konec sRNA vzhledem k referenci, délku, vlákno výskytu (pozitivní/negativní), pokrytí, typ sRNA (*cis-/trans*-sRNA) a pokud se jedná o *cis*-sRNA, je zde uveden potenciální cílový gen, který s vysokou pravděpodobností ovlivňují. Predikce genových cílů pro *trans*-sRNA je nad rámec tohoto nástroje. Tato tabulka je uložena pod názvem **_sRNA*, kde * udává název vstupního BAM souboru.

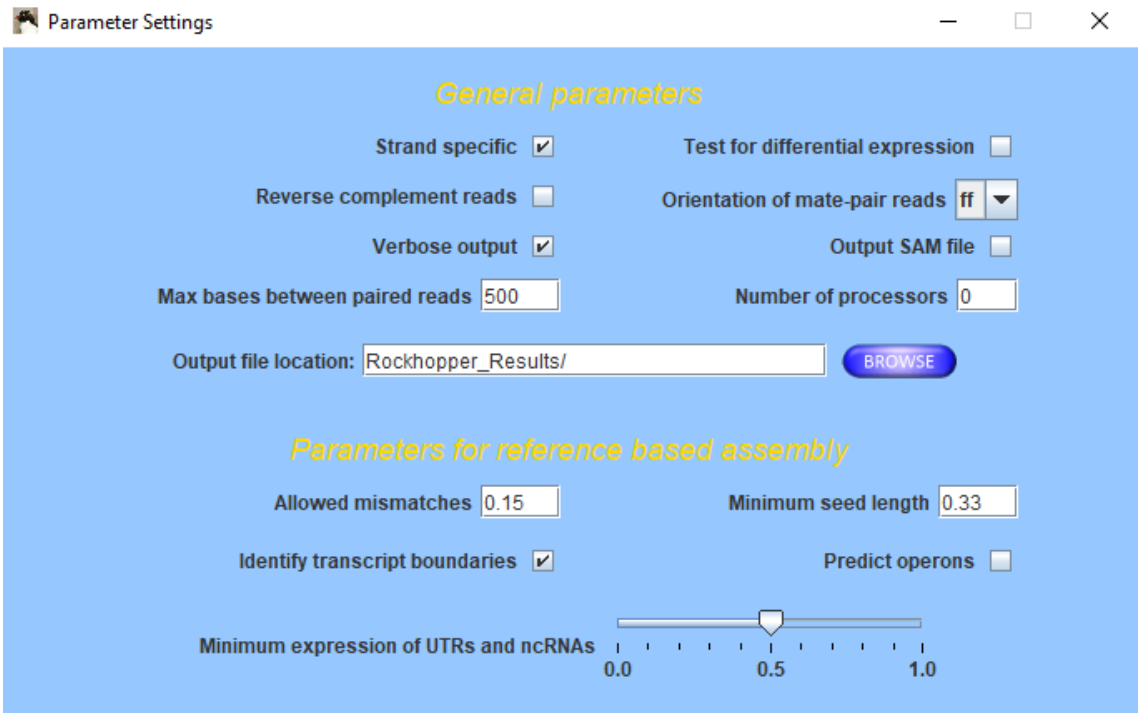
Dále jsou vytvořeny dva soubory v datovém formátu TXT, které obsahují řetězce nul a jedniček pro jednotlivá vlákna. Jedničky určují, že se na daném místě nachází potenciální sRNA, 0 značí, že se na dané pozici sRNA nenachází. Soubory se ukládají pod názvy **_positive_sRNA* a **_negative_sRNA* pro jednotlivá vlákna, kde * opět udává název použitého BAM souboru. Tyto textové soubory lze využít na rychlé zobrazení pozic výskytu potenciálních sRNA.

5. SROVNÁNÍ VÝSLEDKŮ PREDIKCE SRNA POMOCÍ DOSTUPNÝCH NÁSTROJŮ

5.1 Nastavení parametrů

5.1.1 Rockhopper

V programu Rockhopper bylo nastaveno, že se jedná o strand-specific single-end data. Zbylé parametry jako hodnota maximální odlišnosti zarovnaného čtení vůči referenci (*Allowed mismatches*), minimální délka semene (*Minimum seed length*) a hodnota minimální exprese nekódujících RNA (*Minimum expression of UTRs and ncRNAs*) byly ponechány na defaultním nastavení nástroje. Některé funkce nástroje Rockhopper byly vypnuty, neboť pro detekci sRNA nebyly potřebné. Například byla vypnuta funkce pro diferenciální genovou expresi či detekce operonů. Ukázka zvoleného nastavení je vyobrazen na Obrázku 5.1.



Obrázek 5.1 Nastavení programu Rockhopper pro data *Vibrio atlanticus* LGP32.

5.1.2 DETR'PROK

Pro nastavení u algoritmu DETR'PROK bylo zvoleno nastavení, které bylo využito při testování algoritmu samotným tvůrcem. Toto nastavení bylo zvoleno, neboťodpovídalo vhodnému nastavení pro danou hustotu genů u organizmu *Vibrio atlanticus* LGP32. Použité parametry a jejich hodnoty jsou uvedeny v Tabulce 5.1. Například minimální délka sRNA byla nastavena na 50 pb, což odpovídá minimální délce sRNA obecně.

Tabulka 5.1 Zvolené parametry pro algoritmus DETR'PROK.

Parametr	Hodnota	Parametr	Hodnota
op_gap	150	*asrna_min_reads*	22
clust_gap	20	*asrna_min_size*	50
rna_gap	25	*asrna_cov*	10
rna_merge	50	*utr5_min_reads*	10
srna_min_reads	12	*utr5_min_size*	50
srna_min_size	50	*utr5_cov*	0
srna_cov	5	*srna_inclusion*	0.00001

5.1.3 SEARCHsRNA

Parametry vytvořeného nástroje SEARCHsRNA byly nastaveny tak, aby byly co nejvíce srovnatelné s nastavenými parametry u nástroje Rockhopper a DETR'PROK. Jelikož připravený dataset obsahuje strand-specific single-end data, byl parametr *type_of_data* nastaven na hodnotu 'Stranded'. Stejně, jako u nástroje DETR'PROK, byl nastaven parametr určující minimální pokrytí výsledných sRNA *threshold_coverage_sRNA_user* na hodnotu 10, parametr určující minimální délku *min_length_of_sRNA_user* na hodnotu 50 a parametr určující maximální vzdálenost dvou transkriptů, které se mohou spojit *threshold_gap_transcripts_user* na hodnotu 25. Zbylé parametry byly nastaveny automaticky ('NULL'). Stručný přehled nastavených parametrů lze vidět v Tabulce 5.2.

Tabulka 5.2 Zvolené parametry pro nástroj SEARCHsRNA.

Parametr	Hodnota
type_of_data	'Stranded'
threshold_coverage_sRNA_user	10
min_length_of_sRNA_user	50
threshold_coverage_steepness_user	NULL
threshold_coverage_min_user	NULL
threshold_gap_transcripts_user	25

5.2 Obdržené výsledky

V této práci byly porovnány dva volně dostupné nástroje pro predikci sRNA (nástroj Rockhopper a DETR'PROK) a zde navržený nástroj SEARCHsRNA. Všechny tři nástroje byly vyzkoušeny na dostupných datech pro bakterii *Vibrio atlanticus* LGP32 pro oba chromozomy (viz kapitola 4.1.1). Získané výstupy z nástrojů byly následně zpracovány v programovém prostředí R, kde byly sRNA rozlišeny na *cis*-sRNA a *trans*-sRNA. Z obdržených dat byly také vyhotoveny grafy na porovnání vlastností těchto detekovaných sRNA. Následující podkapitoly se věnují obdrženým výsledkům. Výstupní soubory všech srovnávaných nástrojů jsou součástí elektronických příloh. Soupis těchto příloh se nachází v Příloze C.

5.2.1 Stručný přehled obdržených výsledků

Pomocí programu Rockhopper bylo pro chromozom NC_011753.2 predikováno celkově 101 sRNA. Z tohoto počtu se jednalo o 15 *cis*-sRNA a 86 *trans*-sRNA. Nástroj DETR'PROK detekoval celkově 169 sRNA. Z celkového počtu sRNA se jednalo o 48 *cis*-sRNA a 121 *trans*-sRNA. Nástrojem SEARCHsRNA bylo detekováno celkem 159 sRNA, z nichž se jednalo o 49 *cis*-sRNA a 110 *trans*-sRNA. Tyto počty detekovaných sRNA pro chromozom NC_011753.2 jsou uvedeny v Tabulce 5.3.

Tabulka 5.3 Přehled predikovaných sRNA pomocí nástroje Rockhopper, DETR'PROK a SEARCHsRNA pro chromozom NC_011753.2.

Vlákno	Typ sRNA	Rockhopper	DETR'PROK	SEARCHsRNA
Pozitivní	*cis*-	7	31	31
	trans-	48	60	49
Negativní	*cis*-	8	17	18
	trans-	38	61	61
celkem sRNA		**101**	**169**	**159**

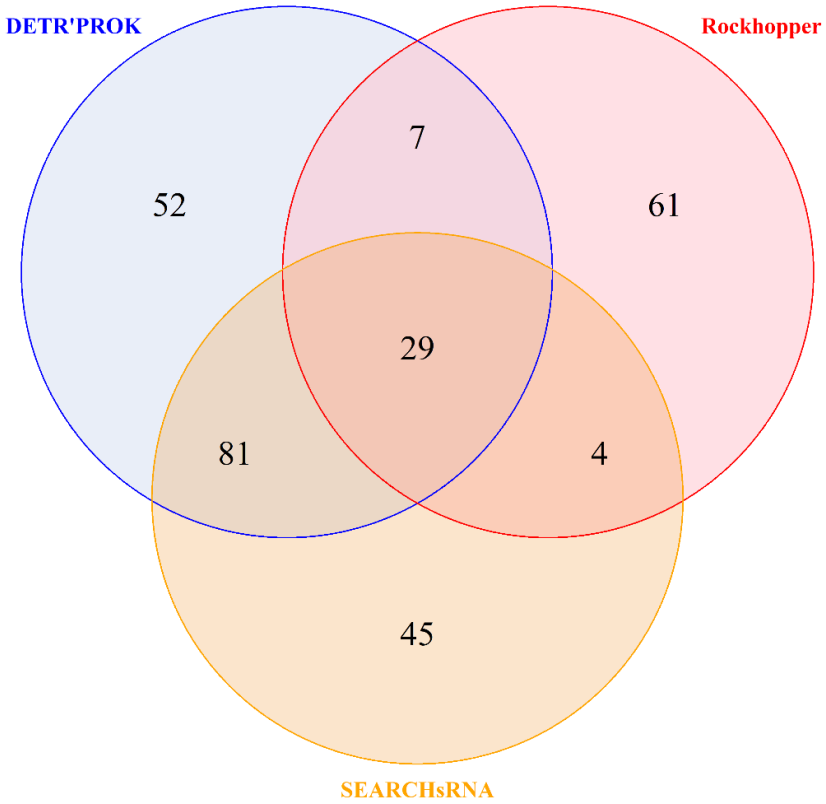
Pro druhý chromozom NC_011744.2 bylo pomocí programu Rockhopper detekováno celkově 451 sRNA. Z tohoto počtu detekovaných sRNA se jednalo o 151 *cis*-sRNA a 300 *trans*-sRNA. Nástroj DETR'PROK detekoval celkově 104sRNA. Z těchto 104 sRNA se jednalo o 22 *cis*-sRNA a 82 *trans*-sRNA. Nástrojem SEARCHsRNA bylo detekováno celkem 74 sRNA, z nichž se jednalo o 19 *cis*-sRNA a 55 *trans*-sRNA. Tyto počty detekovaných sRNA pro chromozom NC_011744.2 jsou uvedeny v Tabulce 5.4.

Seznamy všech predikovaných sRNA pomocí nástroje Rockhopper, DETR'PROK i SEARCHsRNA jsou zobrazeny v tabulkách, které jsou uvedeny v Příloze A.1 až A.6. Tabulky obsahují informace o pozici umístění, délce, vláknu a typu sRNA.

Tabulka 5.4 Přehled predikovaných sRNA pomocí nástroje Rockhopper a DETR'PROK a SEARCHsRNA pro chromozom NC_011744.2.

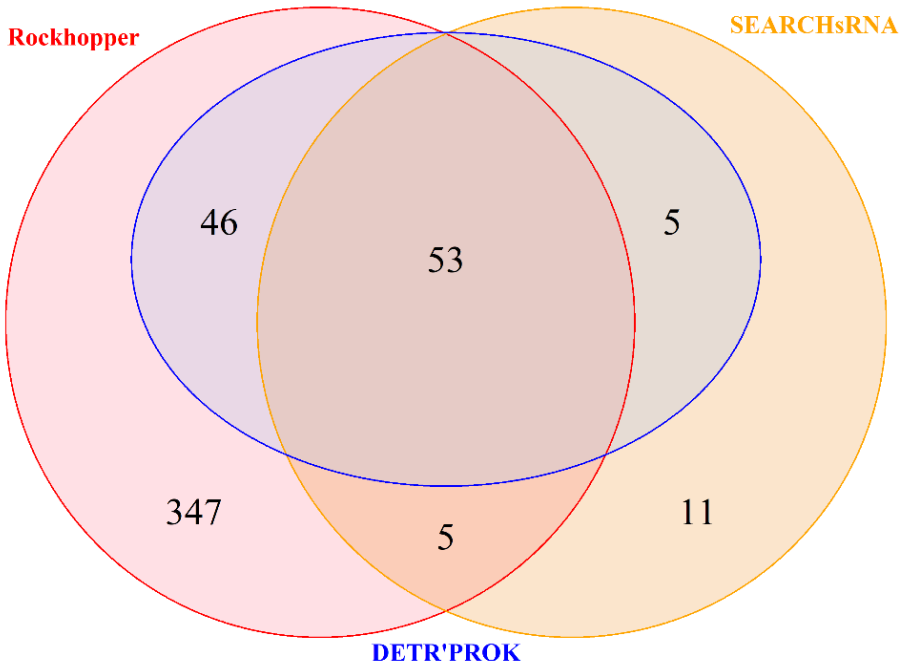
Vlákno	Typ sRNA	Rockhopper	DETR'PROK	SEARCHsRNA
Pozitivní	*cis*-	86	11	8
	trans-	160	41	21
Negativní	*cis*-	65	11	11
	trans-	140	41	34
celkem sRNA		**451**	**104**	**74**

Mezi těmito třemi nástroji došlo ke shodě u 29 detekovaných sRNA u chromozomu NC_011753.2. (pozice sRNA se alespoňčástečně shodovaly u všech tří nástrojů současně). Z celkových počtů predikovaných sRNA se jedná opravdu o malé číslo, které naznačuje, že některý či více z těchto nástrojů nepracují správně. Na Obrázku 5.2 je znázorněný Vennův diagram obsahující informace o tom, v kolika případech se jednotlivé nástroje shodly v detekci sRNA.



Obrázek 5.2 Vennův diagram detekovaných sRNA pro chromozom NC_011753.2: Rockhopper (červená), DETR'PROK (modrá) a SEARCHsRNA (oranžová).

U chromozomu NC_011744.2 došlo ke shodné detekci u 53 sRNA všemi použitými nástroji. Vennův diagram pro grafické znázornění shodujících se sRNA mezi jednotlivými nástroji pro tento chromozom je vyobrazen na Obrázku 5.3 Z diagramu je také patrné, že nástroj DETR'PROK se ve svých detekcích vždy shodl alespoň s jedním ze zbylých nástrojů. Podrobnější popis shodujících se sRNA mezi jednotlivými nástroji je uveden v kapitolách 5.2.2 až 5.2.4.



Obrázek 5.3 Vennův diagram detekovaných sRNA pro chromozom NC_011744.2: Rockhopper (červená), DETR'PROK (modrá) a SEARCHsRNA (oranžová).

Dalším parametrem, který může být u detekovaných sRNA obdržených jednotlivými nástroji porovnáván, je jejich délka. Na grafech v Příloze B.1 až B.4 jsou vyobrazeny box-ploty, které znázorňují variabilitu délek pro *cis*-sRNA a *trans*-sRNA z obdržených výsledků.

Detekované *cis*-sRNA pomocí nástroje Rockhopper pro chromozom NC_011753.2, dosahovaly hodnoty průměrné délky zhruba 44 pb a *trans*-sRNA hodnoty zhruba 50 pb. Nástroj DETR'PROK pak detekoval *cis*-sRNA s průměrnou délkou 140 pb a *trans*-sRNA s průměrnou délkou 136 pb. Pomocí nástroje SEARCHsRNA pak byly detekovány *cis*-sRNA o průměrné délce 108 pb a *trans*-sRNA o průměrné délce 147 pb.

Pro chromozom NC_011744.2 byla u nástroje Rockhopper stanovena průměrná délka *cis*-sRNA na hodnotu 45 pb a pro *trans*-sRNA na hodnotu 52 pb. Pro stejný chromozom u nástroje DETR'PROK dosahovala průměrná délka detekovaných *cis*-sRNA hodnotu 150 pb a *trans*-sRNA hodnoty 119 pb. Pro nástroj SEARCHsRNA dosahovaly *cis*-sRNA průměrné délky 112 pb a *trans*-sRNA 116 pb. Všechny výše zmíněné hodnoty průměrných délek jsou názorně uvedeny v Tabulka 5.5.

Tabulka 5.5 Shrnutí průměrných délek pro detekované sRNA nástroji Rockhopper, DETR'PROK a SEARCHsRNA.

Chromozom	Typ sRNA	Rockhopper	DETR'PROK	SEARCHsRNA
NC_011753.2	*cis*-	44 pb	140 pb	108 pb
	trans-	50 pb	136 pb	147 pb
NC_011744.2	*cis*-	45 pb	150 pb	112 pb
	trans-	52 pb	119 pb	116 pb

Z hodnot uvedených v Tabulce 5.5 lze pozorovat, že průměrné délky sRNA pro oba zkoumané chromozomy jsou u nástroje Rockhopper výrazně nižší než u nástrojů DETR'PROK a SEARCHsRNA. Tento fakt je způsobený tím, že u nástroje nelze definovat minimální délka predikovaných sRNA a spousta z detekovaných sRNA u nástroje Rockhopper dosahuje délky 38 pb, což odpovídá délce čtení použitého při sekvenování. Některé z detekovaných sRNA dosahují pouze délky 10-13 pb. Takovéto výsledky však nejsou přesné a mohou být nazvány za chybné.

U nástrojů DETR'PROK i SEARCHsRNA lze nastavit požadovaná minimální délka výsledných sRNA a chybné detekce tak mohou být odstraněny. Z grafů uvedených v Příloze B.1 až B.4 lze také pozorovat, že oba tyto nástroje dosahují podobných délek u detekovaných sRNA.

5.2.2 Srovnání nástroje Rockhopper s nástrojem DETR'PROK

Pro chromozom NC_011753.2 nástroje Rockhopper a DETR'PROK predikovaly celkově 270 sRNA (bez ohledu na duplicitu). Po analýze došlo ke shodě pouze u 36 sRNA (pozice detekovaných sRNA se z jednotlivých nástrojů alespoň částečně překrývaly). Oběma nástroji tedy bylo predikováno celkově 36 shodujících se sRNA, nástroj Rockhopper dále predikoval 65 odlišných sRNA oproti nástroji DETR'PROK, který jich predikoval odlišných 133. Celkem tedy bylo predikováno 234 jedinečných sRNA těmito nástroji. Veškeré tyto hodnoty jsou znázorněny pomocí Vennova diagramu na Obrázku 5.2.

Pro chromozom NC_011744.2 oba nástroje predikovaly celkově 555 sRNA (bez ohledu na duplicitu). Z tohoto počtu sRNA došlo ke shodě u 99 z nich. Oběma nástroji tedy bylo predikováno celkově 99 shodujících se sRNA. Nástroj Rockhopper dále predikoval 352 odlišných sRNA oproti nástroji DETR'PROK, který predikoval 5 odlišných sRNA. Celkem tedy bylo nalezeno 455 jedinečných sRNA. Tento poměr je vyobrazen na Obrázku 5.3.

Z Tabulky 5.3, Tabulky 5.4, Obrázku 5.2 a Obrázku 5.3 lze vidět, že celkové počty detekovaných sRNA pomocí nástrojů Rockhopper a DETR'PROK se výrazně liší.

Obzvláště je důležité si povšimnout, že nástroj Rockhopper detekuje více sRNA v chromozomu NC_011744.2. Naopak nástroj DETR'PROK detekuje více sRNA v prvním z chromozomů NC_011753.2.

5.2.3 Srovnání nástroje SEARCHsRNA s nástrojem Rockhopper

Nástroje SEARCHsRNA a Rockhopper detekovaly 260 sRNA (bez ohledu na duplicitu) u chromozomu NC_011753.2. Nástroje se shodly při detekci ve 33 sRNA. Nástroj SEARCHsRNA pak detekoval dalších 126 odlišných sRNA a nástroj Rockhopper dalších 68 odlišných sRNA. Tyto počty společných sRNA jsou graficky znázorněné na Obrázku 5.2.

Pro chromozom NC_011744.2 oba z nástrojů detekovaly celkově 525 sRNA (bez ohledu na duplicitu). Společně detekovaly 58 stejných sRNA. Nástrojem SEARCHsRNA bylo detekováno dalších 16 odlišných sRNA a nástrojem Rockhopper bylo detekováno dalších 393 odlišných sRNA. Poměr těchto detekovaných sRNA pro chromozom NC_011744.2 je vyobrazen na Obrázku 5.3.

I v tomto případě nástroj Rockhopper detekoval více sRNA u kratšího z chromozomů – NC_011744.2, a nástroj SEARCHsRNA detekoval více sRNA u delšího z chromozomů – NC_011753.2. Celkově se počty detekovaných sRNA mezi těmito nástroji výrazně lišily.

5.2.4 Srovnání nástroje SEARCHsRNA a nástroje DETR'PROK

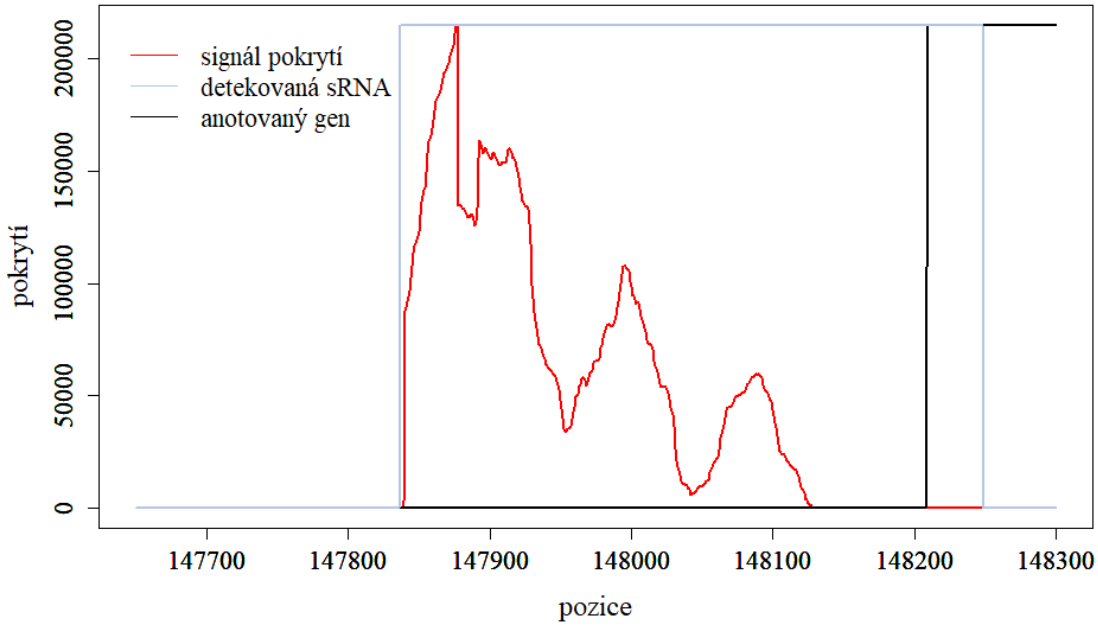
Na závěr byly srovnány nástroje SEARCHsRNA a DETR'PROK. Tyto nástroje detekovaly celkově 327 sRNA (bez ohledu na duplicitu) pro chromozom NC_011753.2. Ke shodě došlo u 110 sRNA, nástroj SEARCHsRNA detekoval dalších 49 odlišných sRNA a nástroj DETR'PROK detekoval dalších 59 odlišných sRNA. Celkem tedy identifikovaly 256 jedinečných sRNA pro chromozom NC_011753.2. Tento poměr predikovaných sRNA je vyobrazen na Obrázku 5.2.

Pro chromozom NC_011744.2 bylo oběma nástroji detekováno celkově 178 sRNA (bez ohledu na jejich duplicitu). Nástroje detekovaly celkově v 58 shodných sRNA, nástroj SEARCHsRNA pak nalezl dalších 16 sRNA a nástroj DETR'PROK nalezl dalších 46 jedinečných sRNA. Celkově tedy nástroje predikovaly 120 jedinečných sRNA. Tento poměr predikovaných sRNA je vyobrazen na Obrázku 5.3.

Celkově lze usoudit, že výsledky z nástrojů SEARCHsRNA a DETR'PROK jsou si navzájem podobnější, než s výsledky z nástroje Rockhopper. Oproti nástroji Rockhopper se shodly i v poměru detekovaných sRNA mezi jednotlivými chromozomy. Bylo tedy náhodně vybráno několik z detekovaných sRNA pomocí nástroje SEARCHsRNA nebo DETR'PROK, pro chromozom NC_011753.2, které byly ručně zkontrolovány za účelem zjištění, proč se v některých případech nástroje neshodly.

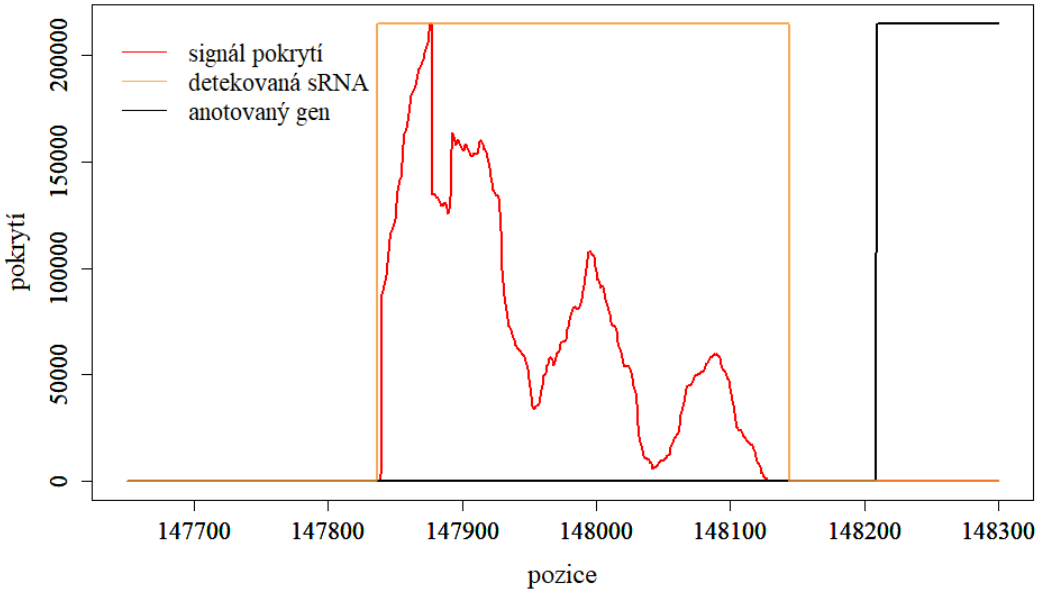
První sRNA, která byla ručně ověřena, byla nalezena na pozitivním vlákně jak pomocí nástroje SEARCHsRNA (Příloha A.3 řádek č. 8), tak i pomocí nástroje

DETR'PROK (Příloha A.2 řádek č. 13). Rozdílem však byla skutečnost, že nástroj SEARCHsRNA tuto sRNA zařadil jako *trans*-sRNA a nástroj DETR'PROK jako *cis*-sRNA. To bylo zapříčiněno tím, že nástroj DETR'PROK tuto sRNA detekoval o něco delší a tedy došlo k částečnému překryvu s anotovaným genem z opačného vlákna, viz Obrázek 5.4. Z obrázku si lze také povšimnout, že míra pokrytí napříč touto detekovanou sRNA není konzistentní a její koncová část má oproti zbylé oblasti výrazně nižší pokrytí. To však není úplně kýžený výsledek.



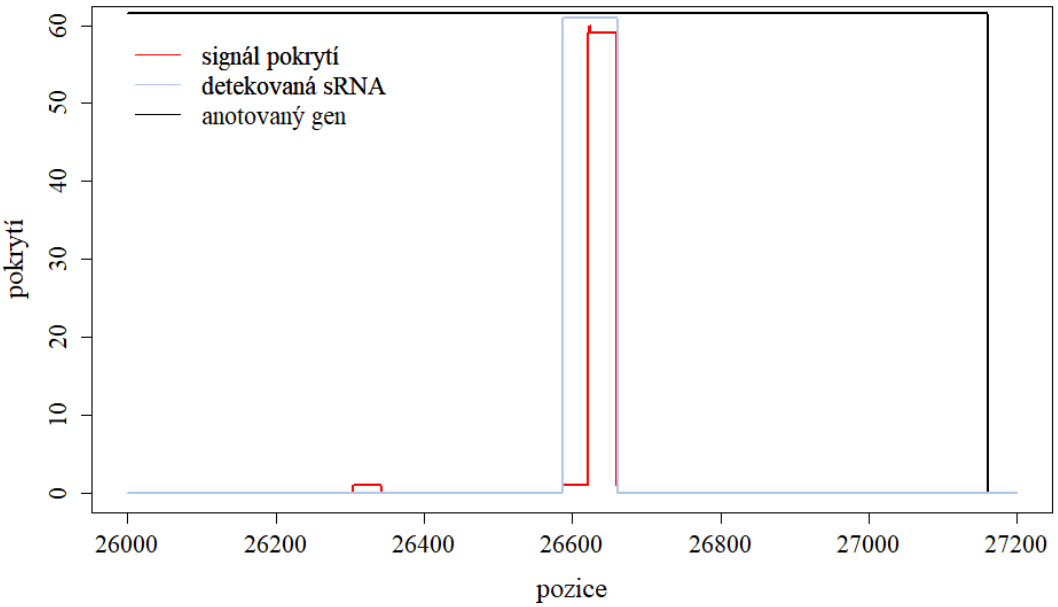
Obrázek 5.4 Ukázka detekované sRNA (Příloha A.2 řádek č. 13) pomocí nástroje DETR'PROK: signál pokrytí (červená), detekovaná sRNA (modrá), anotovaný gen z opačného vlákna (černá).

Naopak nástroj SEARCHsRNA tuto sRNA detekoval kratší a k překryvu s anotovaným genem tím pádem nedošlo. Detekovaná sRNA nástroje SEARCHsRNA je vyobrazena na Obrázku 5.5. Na tomto obrázku si lze povšimnout, že míra pokrytí pro takto detekovanou sRNA je konzistentnější než u sRNA detekované pomocí nástroje DETR'PROK, což je výsledek předpokládaný. Na základě Obrázku 5.4 a Obrázku 5.5 lze předpokládat, že v tomto případě lépe detekoval 3' konec potenciální sRNA nástroj SEARCHsRNA a tedy se s velkou pravděpodobností jedná spíše o sRNA typu *trans*-sRNA. Přesto oba z nástrojů tento úsek úspěšně detekovaly jako potenciální výskyt sRNA.



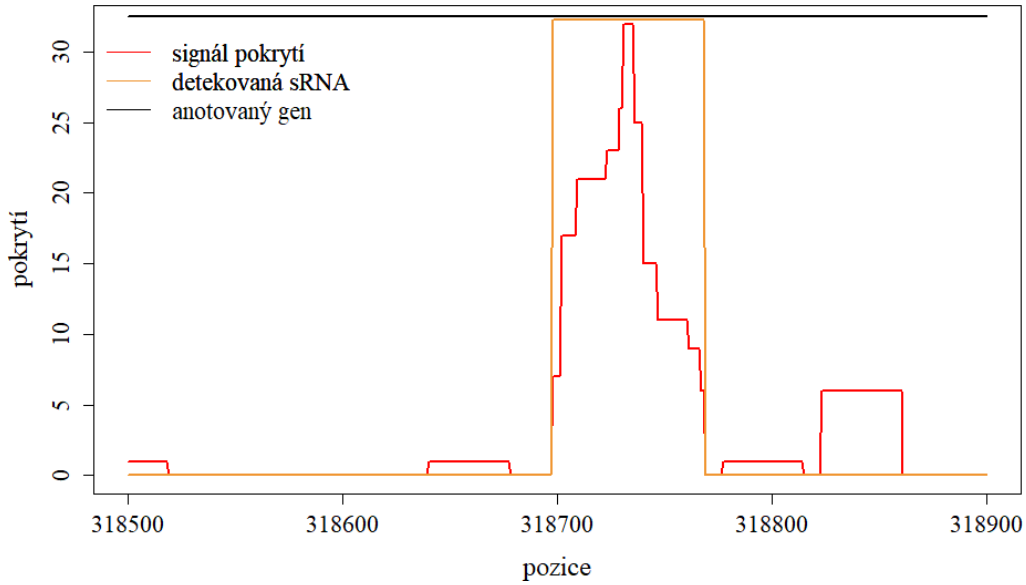
Obrázek 5.5 Ukázka detekované sRNA (Příloha A.3 řádek č. 8) pomocí nástroje SEARCHsRNA: signál pokrytí (červená), detekovaná sRNA (oranžová), anotovaný gen z opačného vlákna (černá).

Další kontrolovanou sRNA byla *cis*-sRNA, která byla detekována pouze nástrojem DETR'PROK (Příloha A.2 řádek č. 2). Tato potenciální sRNA byla nalezena na negativním vlákně chromozomu NC_011753.2 a dosahuje délky 74 pb. Na Obrázku 5.6 je patrné, že pokrytí u této sRNA je opět velmi nekonzistentní a dá se říci, že požadované pokrytí zadané na vstupu dosahuje pouze polovina z celé délky detekované sRNA. Lze tedy předpokládat, že nástroj DETR'PROK pochybil a zmíněný úsek predikován být neměl.

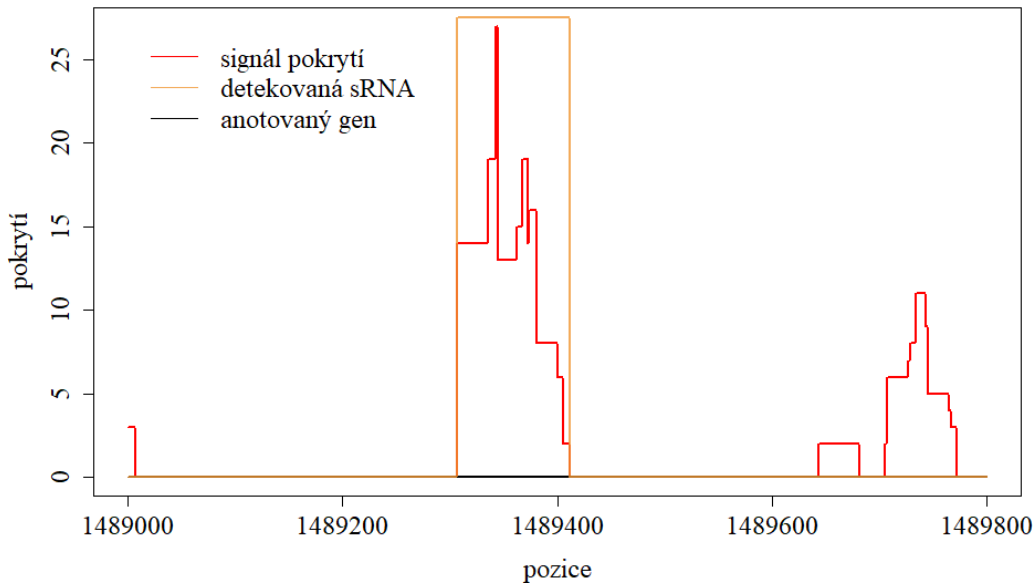


Obrázek 5.6 Ukázka detekované sRNA (Příloha A.2 řádek č. 2) pomocí nástroje DETR'PROK: signál pokrytí (červená), detekovaná sRNA (modrá), anotovaný gen z opačného vlákna (černá).

Dalšími kontrolovanými sRNA byla *cis*-sRNA vyskytující se na negativním vlákně (Příloha A.3 řádek č. 17) a *trans*-sRNA vyskytující se na pozitivním vlákně (Příloha A.3 řádek č. 77) detekovaná nástrojem SEARCHsRNA. První z těchto sRNA dosahuje délky 71 pb, druhá dosahuje délky 105 pb. Na Obrázku 5.7 a Obrázku 5.8 lze pro tyto sRNA pozorovat relativně konzistentní pokrytí bez skokových rozdílů, což je od detekce požadováno. Je tedy velkým předpokladem, že obě tyto detekce byly správné a opravdu se jedná o potenciální sRNA i přes fakt, že celkové průměrné pokrytí těchto sRNA nedosahuje tak vysokých hodnot.

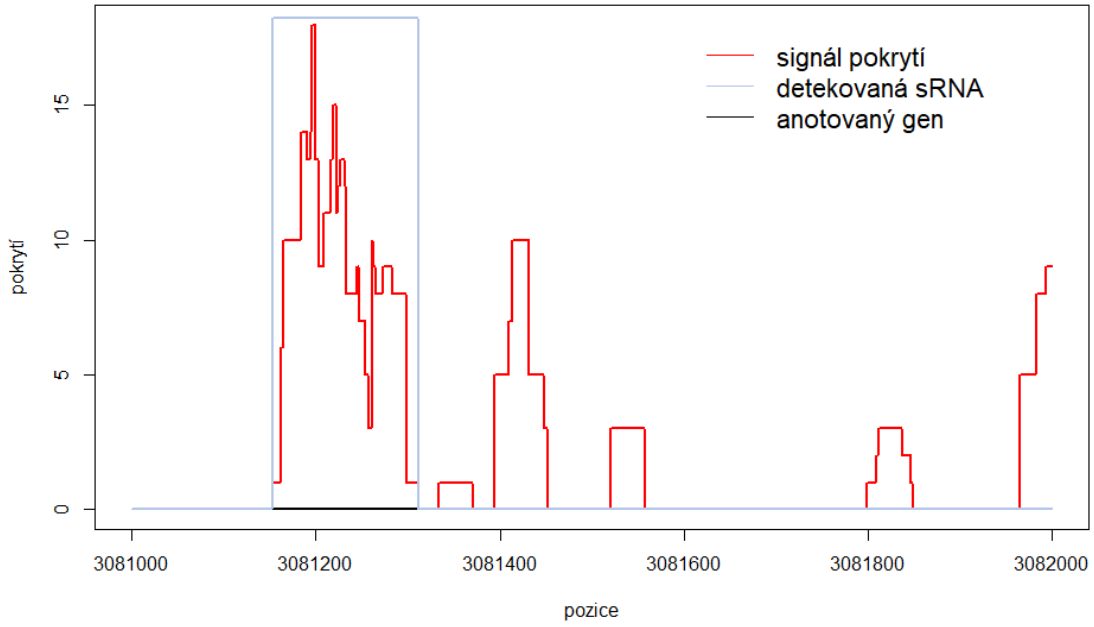


Obrázek 5.7 Ukázka detekované sRNA (Příloha A.3 řádek č. 17) pomocí nástroje SEARCHsRNA: signál pokrytí (červená), detekovaná sRNA (oranžová), anotovaný gen z opačného vlákna (černá).



Obrázek 5.8 Ukázka detekované sRNA (Příloha A.3 řádek č. 77) pomocí nástroje SEARCHsRNA: signál pokrytí (červená), detekovaná sRNA (oranžová), anotovaný gen z opačného vlákna (černá).

Na závěr byla ověřena *trans*-sRNA, která byla detekována pouze pomocí nástroje DETR'PROK (Příloha A.2 řádek č. 156). Opět se jedná o sRNA vyskytující se na pozitivním vlákně, grafické znázornění této detekce je na Obrázku 5.9, kde lze vidět, že i zde by se mohlo jednat o potenciální sRNA. Špatná detekce nástroje SEARCHsRNA zde mohla být způsobena automatickým nastavením parametrů určujících pozice 5' a 3' konců transkriptů, čemuž by manuální upravení těchto parametrů mohlo zabránit.



Obrázek 5.9 Ukázka detekované sRNA (Příloha A.2 řádek č. 156) pomocí nástroje DETR'PROK: signál pokrytí (červená), detekovaná sRNA (modrá), anotovaný gen z opačného vlákna (černá).

5.2.5 Celkové zhodnocení výsledků

Z výše uvedené analýzy predikovaných sRNA nástroji Rockhopper, DETR'PROK a zde navrženým nástrojem SEARCHsRNA lze usoudit, že nástroj Rockhopper predikuje nejhůře z vyzkoušených nástrojů. K tomuto názoru spěje fakt, že do výsledného seznamu detekovaných sRNA uvádí i úseky o délce 11-13 pb, což není požadovaný výsledek. Také dochází k predikci mnohonásobně vyššího počtu sRNA u kratšího z chromozomů – chromozomu NC_011744.2, kde lze předpokládat, že by počet vyskytujících se sRNA měl být nižší, než počet vyskytujících se sRNA u delšího z chromozomů – chromozomu NC_011753.2. Což tento předpoklad nesplňuje.

Jelikož byly nástroje testovány na souborech BAM, ve kterých je relativně nízké průměrné pokrytí, špatná detekce nástroje Rockhopper mohla být způsobena právě tímto. Je možné, že se nástroj Rockhopper neumí plně přizpůsobit vstupním datům a proto došlo k nepřesné detekci.

Přestože se některé sRNA detekované nástrojem Rockhopper shodovaly s výsledky obdrženými z nástrojů DETR'PROK i SEARCHsRNA a je tedy možné, že se v těchto případech jednalo o správnou detekci, v poměru se špatně detekovanými úseky je tento

počet zanedbatelný. Výsledky obdržené z nástroje Rockhopper lze tedy považovat za nekvalitní a nepříliš vhodné pro další analýzu.

Nástroj DETR'PROK společně s nástrojem SEARCHsRNA naopak dosahovaly podobných výsledků a téměř v polovině predikovaných potenciálních sRNA se shodly. Délky obdržených sRNA jsou u obou z nástrojů v normě a počty detekovaných sRNA u jednotlivých chromozomů odpovídají zmíněnému předpokladu výskytu menšího počtu sRNA u kratšího z chromozomů.

Z podrobnější analýzy náhodně vybraných detekovaných sRNA nástroji DETR'PROK a SEARCHsRNA lze říci, že ani jeden z nástrojů – DETR'PROK ani SEARCHsRNA – není při detekci stoprocentní a může docházet k chybám. Tyto chyby jsou však s porovnáním u nástroje Rockhopper v mnohem menší míře a mohou být pomocí manuálního nastavení parametrů opraveny. Oba z nástrojů tedy přináší kvalitní výsledky, které mohou být dále laboratorně ověřeny.

6. ZÁVĚR

Diplomová práce se v teoretické části zaměřuje na způsob získání RNA-Seq dat pomocí sekvenačních platforem NGS a TGS a na volně dostupné nástroje pro predikci sRNA u bakterií. Pro jednotlivé nástroje – Rockhopper, DETR'PROK, ANNOgesic, APERO a Baerhunter byly uvedeny základní informace a byl zde popsán jejich pracovní postup.

V rámci praktické části byl navržen a implementován v programovacím prostředí R nástroj SEARCHsRNA, který predikuje potenciální umístění sRNA. Jedním z hlavních důvodů pro navržení tohoto nástroje byla uživatelská přívětivost, neboť spuštění většiny z dostupných nástrojů pro detekci sRNA je velmi komplikované. Spuštění nástroje SEARCHsRNA nepožaduje žádné speciální knihovny funkcí, které by bylo obtížné instalovat, a pro spuštění automatické detekce stačí nastavit pouze dva parametry.

Následně byly dva z volně dostupných nástrojů – Rockhopper a DETR'PROK, společně s navrženým nástrojem SEARCHsRNA, vyzkoušeny na připraveném datasetu RNA-Seq dat pro bakterii *Vibrio atlanticus* LGP32 obsahující dva chromozomy. RNA-Seq data byla stažena z volně dostupných doplňkových dat pro predikční nástroj DETR'PROK a referenční sekvence s anotací byly staženy z databáze GenBank. Zbylé nástroje nebyly vyzkoušeny, neboť nepodporovaly detekci sRNA u tohoto typu RNA-Seq dat, nebo se jejich spuštění nezdařilo.

Z obdržených výsledků lze usoudit, že nástroj Rockhopper neprokazuje na zvoleném datasetu dobrých kvalit. Detekovaná sRNA jsou příliš krátká a jejich počty neodpovídají předpokladům pro zvolený dataset.

Nástroj DETR'PROK dosahoval mnohem kvalitnějších výsledků. Průměrné délky i počty detekovaných potenciálních sRNA byly přijatelné a v očekávaném rozsahu. U nástroje DETR'PROK však docházelo ke zbytečnému prodlužování predikovaných transkriptů bez ohledu na jejich pokrytí. Tím docházelo u některých sRNA k falešně pozitivní predikci či k nesprávnému určení výsledného typu sRNA.

Výsledky obdržené z nástroje SEARCHsRNA se z velké části podobaly výsledkům obdržených nástrojem DETR'PROK. Nástroj SEARCHsRNA v porovnání s nástrojem DETR'PROK úspěšně eliminoval z výsledného seznamu potenciálních sRNA ty sRNA, které neobsahovaly konstantní pokrytí na určité požadované minimální délce a nedovolil nežádoucí prodlužování transkriptů bez ohledu na jejich pokrytí. Nástroj SEARCHsRNA byl také oproti nástroji DETR'PROK citlivější na nalezení sRNA obsahující celkově nižší průměrné pokrytí, což umožňuje nalezení sRNA s nižší mírou exprese v dané chvíli.

Celkově lze označit nástroj SEARCHsRNA za jednoduše spustitelný a úspěšný nástroj pro predikci výskytů potenciálních sRNA, které mohou být využity pro detekci těchto sRNA laboratorními metodami. Přesto, jako u většiny podobných nástrojů, je i zde prostor na zdokonalení. Toho by bylo možné dosáhnout optimalizováním některých parametrů na základě dalších datasetů obsahující bakteriální RNA-Seq data.

LITERATURA

[1] MIZUNO, T., M. CHOU a M. INOUYE. A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). *Proceedings of the National Academy of Sciences*. 1984, **81**(7), 1966-1970. ISSN 0027-8424. Dostupné z: doi:10.1073/pnas.81.7.1966

[2] COWAN, S., R. GARAVITO, J. JANSONIUS et al. The structure of OmpF porin in a tetragonal crystal form. *Structure*. 1995, **3**(10), 1041-1050. ISSN 09692126. Dostupné z: doi:10.1016/S0969-2126(01)00240-4

[3] BERVOETS, I. a D. CHARLIER. Diversity, versatility and complexity of bacterial gene regulation mechanisms: opportunities and drawbacks for applications in synthetic biology. *FEMS Microbiology Reviews*. 2019, **43**(3), 304-339. ISSN 1574-6976. Dostupné z: doi:10.1093/femsre/fuz001

[4] CROUCHER, N. a N. THOMSON. Studying bacterial transcriptomes using RNA-seq. *Current Opinion in Microbiology*. 2010, **13**(5), 619-624. ISSN 13695274. Dostupné z: doi:10.1016/j.mib.2010.09.009

[5] What is RNA-Seq?. In: *ZYMO RESEARCH: The Beauty of Science is to Make Things Simple*. Dostupné také z: https://www.zymoresearch.com/pages/what-is-rna-seq

[6] KOLÍSKO, M. *Moderní metody sekvenování DNA*. 2017, . Dostupné také z: https://ziva.avcr.cz/files/ziva/pdf/moderni-metody-sekvenovani-dna.pdf

[7] A targeted method for both small RNA profiling and discovery applications: Sequence the complete range of small RNA and miRNA species. In: *Illumina*. 2015. Dostupné také z: https://www.illumina.com/techniques/sequencing/rna-sequencing/small-rna-seq.html

[8] TJADEN, Brian. De novo assembly of bacterial transcriptomes from RNA-seq data. *Genome Biology*. 2015, **16**(1). ISSN 1474-760X. Dostupné z: doi:10.1186/s13059-014-0572-2

[9] MCCLURE, R., D. BALASUBRAMANIAN, Y. SUN, M. BOBROVSKYY, P. SUMBY, C. GENCO, C. VANDERPOOL a B. TJADEN. Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Research*. 2013, **41**(14), 140-140. ISSN 0305-1048. Dostupné z: doi:10.1093/nar/gkt444

[10] TOFFANO-NIOCHE, C., Y. LUO, C. KUCHLY, C. WALLON, D. STEINBACH, M. ZYTNICKI, A. JACQ a D. GAUTHERET. Detection of non-coding RNA in bacteria and archaea using the DETR'PROK Galaxy pipeline. *Methods*. 2013, **63**(1), 60-65. ISSN 10462023. Dostupné z: doi:10.1016/j.ymeth.2013.06.003

[11] YU, S., J. VOGEL a K. FÖRSTNER. ANNOgesic: a Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes. *GigaScience*. 2018, **7**(9). ISSN 2047-217X. Dostupné z: doi:10.1093/gigascience/giy096

[12] LEONARD, S., S. MEYER, S. LACOUR, W. NASSER, F. HOMMAIS a S. REVERCHON. APERO: a genome-wide approach for identifying bacterial small RNAs from RNA-Seq data. *Nucleic Acids Research*. 2019, **47**(15), 88-88. ISSN 0305-1048. Dostupné z: doi:10.1093/nar/gkz485

[13] OZUNA, A., D. LIBERTO, R. JOYCE, K. ARNVIG, I. NOBELI a J. GORODKIN. Baerhunter: an R package for the discovery and analysis of expressed non-coding regions in bacterial RNA-seq data. *Bioinformatics*. 2019. ISSN 1367-4803. Dostupné z: doi:10.1093/bioinformatics/btz643

[14] VOGEL, J. a E. WAGNER. Target identification of small noncoding RNAs in bacteria. *Current Opinion in Microbiology*. 2007, **10**(3), 262-270. ISSN 13695274. Dostupné z: doi:10.1016/j.mib.2007.06.001

[15] GOTTESMAN, Susan. Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends in Genetics*. 2005, **21**(7), 399-404. ISSN 01689525. Dostupné z: doi:10.1016/j.tig.2005.05.008

[16] THOMASON, M. a G. STORZ. Bacterial Antisense RNAs: How Many Are There, and What Are They Doing?. *Annual Review of Genetics*. 2010, **44**(1), 167-188. ISSN 0066-4197. Dostupné z: doi:10.1146/annurev-genet-102209-163523

[17] WAGNER, E. a P. ROMBY. *Small RNAs in Bacteria and Archaea*. In: . Elsevier, 2015, s. 133-208. Advances in Genetics. ISBN 9780128036945. Dostupné z: doi:10.1016/bs.adgen.2015.05.001

[18] EDDY, Sean R. Non–coding RNA genes and the modern RNA world. *Nature Reviews Genetics*. 2001, **2**(12), 919-929. ISSN 1471-0056. Dostupné z: doi:10.1038/35103511

[19] UDEKWU, K., F. DARFEUILLE, J. VOGEL, J. REIMEGåRD, E. HOLMQVIST a E. WAGNER. Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA. *Genes & Development*. 2005, **19**(19), 2355-2366. ISSN 0890-9369. Dostupné z: doi:10.1101/gad.354405

[20] SELVARAJ, S., P. PERIANDYTHEVAR a N. PRASADARAO. Outer membrane protein A of Escherichia coli K1 selectively enhances the expression of intercellular adhesion molecule-1 in brain microvascular endothelial cells. *Microbes and Infection*. 2007, **9**(5), 547-557. ISSN 12864579. Dostupné z: doi:10.1016/j.micinf.2007.01.020

[21] ZHANG, A., S. ALTUVIA, A. TIWARI, L. ARGAMAN, R. HENGGE-ARONIS a G. STORZ. The OxyS regulatory RNA represses rpoS translation and binds the Hfq (HF-I) protein. *The EMBO Journal*. **17**(20), 6061-6068. ISSN 14602075. Dostupné z: doi:10.1093/emboj/17.20.6061

[22] WANG, J., T. LIU, B. ZHAO, Q. LU, Z. WANG, Y. CAO a W. LI. SRNATarBase 3.0: an updated database for sRNA-target interactions in bacteria. *Nucleic Acids Research*. 2016, **44**(1), 248-253. ISSN 0305-1048. Dostupné z: doi:10.1093/nar/gkv1127

[23] LEE, E. a E. GROISMAN. An antisense RNA that governs the expression kinetics of a multifunctional virulence gene. *Molecular Microbiology*. 2010, **76**(4), 1020-1033. ISSN 0950382X. Dostupné z: doi:10.1111/j.1365-2958.2010.07161.x

[24] BRANTL, S. a P. MÜLLER. Cis- and Trans-Encoded Small Regulatory RNAs in Bacillus subtilis. *Microorganisms*. 2021, **9**(9). ISSN 2076-2607. Dostupné z: doi:10.3390/microorganisms9091865

[25] LÓPEZ-GOMOLLÓN, Sara. Detecting sRNAs by Northern Blotting. In: DALMAY, Tamas, ed., Tamas DALMAY. *MicroRNAs in Development*. Totowa, NJ: Humana Press, 2011, s. 25-38. Methods in Molecular Biology. ISBN 978-1-61779-082-9. Dostupné z: doi:10.1007/978-1-61779-083-6_3

[26] HE, S. a R. GREEN. Northern Blotting. In: *Laboratory Methods in Enzymology: RNA*. Elsevier, 2013, s. 75-87. Methods in Enzymology. ISBN 9780124200371. Dostupné z: doi:10.1016/B978-0-12-420037-1.00003-8

[27] VELCULESCU, V., L. ZHANG, B. VOGELSTEIN a K. KINZLER. Serial Analysis of Gene Expression. *Science*. 1995, **270**(5235), 484-487. ISSN 0036-8075. Dostupné z: doi:10.1126/science.270.5235.484

[28] YE, S., T. LAVOIE, D. USHER a Li. ZHANG. Microarray, SAGE and their applications to cardiovascular diseases. *Cell Research*. 2002, **12**(2), 105-115. ISSN 1001-0602. Dostupné z: doi:10.1038/sj.cr.7290116

[29] TAYLOR, S., M. WAKEM, G. DIJKMAN, M. ALSARRAJ a M. NGUYEN. A practical approach to RT-qPCR—Publishing data that conform to the MIQE guidelines. *Methods*. 2010, **50**(4), 1-5. ISSN 10462023. Dostupné z: doi:10.1016/j.ymeth.2010.01.005

[30] TAYLOR, S., G. LAPERRIERE a H. GERMAIN. Droplet Digital PCR versus qPCR for gene expression analysis with low abundant targets: from variable nonsense to publication quality data. *Scientific Reports*. 2017, **7**(1). ISSN 2045-2322. Dostupné z: doi:10.1038/s41598-017-02217-x

[31] Analýza genové exprese: RNA-Seq nebo kvantitativní PCR?. In: *Baria*. 2021. Dostupné také z: https://www.baria.cz/blog/analyza-genove-exprese-rna-seq-nebo-kvantitativni-pcr/

[32] The Dynamic Range of RNA-Seq. Subio = sub_stream + bio. Dostupné z: https://www.subioplatform.com/products/subioplatform/the-dynamic-range-of-rna-seq

[33] ATHANASOPOULOU, K., M. BOTI, P. ADAMOPOULOS, P. SKOUROU a A. SCORILAS. Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. *Life*. 2022, **12**(1). ISSN 2075-1729. Dostupné z: doi:10.3390/life12010030

[34] STARK, R., M. GRZELAK a J. HADFIELD. RNA sequencing: the teenage years. *Nature Reviews Genetics*. 2019, **20**(11), 631-656. ISSN 1471-0056. Dostupné z: doi:10.1038/s41576-019-0150-2

[35] MORLAN, J., K. QU, D. SINICROPI a S. DADRAS. Selective Depletion of rRNA Enables Whole Transcriptome Profiling of Archival Fixed Tissue. *PLoS ONE*. 2012, **7**(8). ISSN 1932-6203. Dostupné z: doi:10.1371/journal.pone.0042882

[36] BETIN, V., C. PENARANDA, N. BANDYOPADHYAY et al. Hybridization-based capture of pathogen mRNA enables paired host-pathogen transcriptional analysis. *Scientific Reports*. 2019, **9**(1). ISSN 2045-2322. Dostupné z: doi:10.1038/s41598-019-55633-6

[37] *Oxford Dictionary of Biochemistry and Molecular Biology*. Revised Edit. New York: Oxford University Press, 2006. ISBN 978-0-19-852917-0.

[38] BIRD, I. M. Extraction of RNA From Cells and Tissue. In: FENNELL, Jérôme a Andrew BAKER. *Hypertension*. New Jersey: Humana Press, 2004, s. 139-148. ISBN 1-59259-850-1. Dostupné z: doi:10.1385/1-59259-850-1:139

[39] CORTÉS-MALDONADO, L., J. MARCIAL-QUINO, S. GÓMEZ-MANZO, F. FIERRO a A. TOMASINI. A method for the extraction of high quality fungal RNA suitable for RNA-seq. *Journal of Microbiological Methods*. 2020, **170**. ISSN 01677012. Dostupné z: doi:10.1016/j.mimet.2020.105855

[40] HRDLICKOVA, R., M. TOLOUE a B. TIAN. RNA -Seq methods for transcriptome analysis. *WIREs RNA*. 2017, **8**(1). ISSN 1757-7004. Dostupné z: doi:10.1002/wrna.1364

[41] Read Length. In: *Illumina*. Dostupné také z: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/read-length.html

[42] GUPTA, A. a U. GUPTA. Next Generation Sequencing and Its Applications. In: *Animal Biotechnology*. Elsevier, 2014, s. 345-367. ISBN 9780124160026. Dostupné z: doi:10.1016/B978-0-12-416002-6.00019-5

[43] RHOADS, A. a K. AU. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*. Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China, 2015, **13**(5), 278-289. Dostupné z: doi:10.1016/j.gpb.2015.08.002

[44] LU, H., F. GIORDANO a Z. NING. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*. 2016, **14**(5), 265-279. ISSN 16720229. Dostupné z: doi:10.1016/j.gpb.2016.05.004

[45] HAFNER, M., P. LANDGRAF, J. LUDWIG et al. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*. 2008, **44**(1), 3-12. ISSN 10462023. Dostupné z: doi:10.1016/j.ymeth.2007.09.009

[46] Small RNA Library Preparation. In: *New England BioLabs*. Dostupné také z: https://international.neb.com/applications/ngs-sample-prep-and-target-enrichment/rna-library-preparation/small-rna-library-preparation

[47] SEGERMAN, B. The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases. *Frontiers in Cellular and Infection Microbiology*. 2020, **10**. ISSN 2235-2988. Dostupné z: doi:10.3389/fcimb.2020.527102

[48] CHAITANKAR, V., G. KARAKÜLAH, R. RATNAPRIYA, F. GIUSTE, M. BROOKS a A. SWAROOP. Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. *Progress in Retinal and Eye Research*. 2016, **55**, 1-31. ISSN 13509462. Dostupné z: doi:10.1016/j.preteyeres.2016.06.001

[49] HEATHER, J. a B. CHAIN. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 2016, **107**(1), 1-8. ISSN 08887543. Dostupné z: doi:10.1016/j.ygeno.2015.11.003

[50] ROTHBERG, J., W. HINZ, T. REARICK et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 2011, **475**(7356), 348-352. ISSN 0028-0836. Dostupné z: doi:10.1038/nature10242

[51] WILLIAMS, R., S. PEISAJOVICH, O. MILLER, S. MAGDASSI, D. TAWFIK a A. GRIFFITHS. Amplification of complex gene libraries by emulsion PCR. *Nature Methods*. 2006, **3**(7), 545-550. ISSN 1548-7091. Dostupné z: doi:10.1038/nmeth896

[52] 454-roche. In: *LAB Guide: Průvodce laboratoří*. Dostupné také z: https://labguide.cz/metody/sekvenovani-dna/sekvenovani-nove-generace/454-roche/

[53] OHTA, Jun. *Smart CMOS image sensors and applications*. 2. Boca Raton: CRC Press, 2020. ISBN 9781420019155.

[54] FICHOT, E. a R. NORMAN. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome*. 2013, **1**(1). ISSN 2049-2618. Dostupné z: doi:10.1186/2049-2618-1-10

[55] NGUYEN, J. Oxford Nanopore sequncing. In: *APOLLO INSTITUTE*. 2021. Dostupné také z: https://apollo-institute.org/oxford-nanopore-sequencing/

[56] LEVIN, J., M. YASSOUR, X. ADICONIS, Ch. NUSBAUM, D. THOMPSON, N. FRIEDMAN, A. GNIRKE a A. REGEV. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods*. 2010, **7**(9), 709-715. ISSN 1548-7091. Dostupné z: doi:10.1038/nmeth.1491

[57] TSAI, K., B. CHANG, Ch. PAN, W. LIN, T. CHEN a S. LI. Evaluation and Application of the Strand-Specific Protocol for Next-Generation Sequencing. *BioMed Research International*. 2015, **2015**, 1-8. ISSN 2314-6133. Dostupné z: doi:10.1155/2015/182389

[58] How do strand specific sequencing protocols work?. In: *ECSEQ BIOINFORMATICS*. ecSeq Bioinformatics, 2018. Dostupné také z: https://www.ecseq.com/support/ngs/how-do-strand-specific-sequencing-protocols-work

[59] Directional/stranded RNA-seq data -which parameters to choose?. In: *Chipster: Open source platform for data analysis*. 2011-2023. Dostupné také z: https://chipster.rahtiapp.fi/manual/library-type-summary.html

[60] LEE-LIU, D., L. ALMONACID, F. FAUNES, F. MELO a J. LARRAIN. Transcriptomics Using Next Generation Sequencing Technologies. In: HOPPLER, STEFAN a Peter D VIZE, ed., S. HOPPLER, P. VIZE. *Xenopus Protocols*. Totowa, NJ: Humana Press, 2012, s. 293-317. Methods in Molecular Biology. ISBN 978-1-61779-991-4. Dostupné z: doi:10.1007/978-1-61779-992-1_18

[61] TJADEN, Brian. Rockhopper: User Guide. In: *Rockhopper*. Wellesley College: Department of Computer Science, 2020. Dostupné také z: http://cs.wellesley.edu/~btjaden/Rockhopper/user_guide.html

[62] LANGMEAD, B. a S. SALZBERG. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, **9**(4), 357-359. ISSN 1548-7091. Dostupné z: doi:10.1038/nmeth.1923

[63] LANGMEAD, B. Introduction. In: *Bowtie 2: Fast and sensitive read alignment*. Johns Hopkins University. Dostupné také z: https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#introduction

[64] AFGAN, E., D. BAKER, M. VAN DEN BEEK et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*. 2016, **44**(1), 3-10. ISSN 0305-1048. Dostupné z: doi:10.1093/nar/gkw343

[65] LI, L., D. HUANG, M. CHEUNG, W. NONG, Q. HUANG a H. KWAN. BSRD: a repository for bacterial small regulatory RNA. *Nucleic Acids Research*. 2013, **41**(1), 233-238. ISSN 0305-1048. Dostupné z: doi:10.1093/nar/gks1264

[66] LAWRENCE, M., W. HUBER, H. PAGèS et al. Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*. 2013, **9**(8). ISSN 1553-7358. Dostupné z: doi:10.1371/journal.pcbi.1003118

[67] MORGAN, M., H. PAGES, V. OBENCHAIN a N. HAYDEN. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. R package version 2.8.0. In: *Bioconductor*. Dostupné také z: https://bioconductor.org/packages/release/bioc/html/Rsamtools.html

[68] CHARIF, D. a J. LOBRY, U. BASTOLLA a M. PORTO, ed., H. ROMAN, M. VENDRUSCOLO. *Structural approaches to sequence evolution: Molecules, networks, populations: Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis.* New York: Springer Verlag, 2007. ISBN 978-3-540-35306-5. Dostupné také z: https://link.springer.com/book/10.1007/978-3-540-35306-5

[69] WICKHAM, H., R. FRANçOIS, L. HENRY, K. MÜLLER a D. VAUGHAN. Dplyr: Overview. In: *Dplyr*. 2023. Dostupné také z: https://dplyr.tidyverse.org

[70] LE ROUX, F., J. BINESSE, D. SAULNIER a D. MAZEL. Construction of a Vibrio splendidus Mutant Lacking the Metalloprotease Gene vsm by Use of a Novel Counterselectable Suicide Vector. *Applied and Environmental Microbiology*. 2007, **73**(3), 777-784. ISSN 0099-2240. Dostupné z: doi:10.1128/AEM.02147-06

[71] MAZEL, D. a F. LE ROUX. Proteomes · Vibrio atlanticus (strain LGP32) (Vibrio splendidus (strain Mel32)). In: *UniProt*. Dostupné také z: https://www.uniprot.org/proteomes/UP000009100

[72] GAY, M., T. RENAULT, A. PONS a F. LE ROUX. Two Vibrio splendidus related strains collaborate to kill Crassostrea gigas: taxonomy and host alterations. *Diseases of Aquatic Organisms*. 2004, **62**, 65-74. ISSN 0177-5103. Dostupné z: doi:10.3354/dao062065

[73] PATAKOVA, P., B. BRANSKA, K. SEDLAR, M. VASYLKIVSKA, K. JURECKOVA, J. KOLEK, P. KOSCOVA a I. PROVAZNIK. Acidogenesis, solventogenesis, metabolic stress response and life cycle changes in Clostridium beijerinckii NRRL B-598 at the transcriptomic level. *Scientific Reports*. 2019, **9**(1). ISSN 2045-2322. Dostupné z: doi:10.1038/s41598-018-37679-0

SEZNAM SYMBOLŮ A ZKRATEK

Zkratky:

DNA	deoxyribonukleová kyselina
cDNA	komplementární DNA
dsDNA	dvouvláknová DNA
CDS	kódující sekvence
RNA	ribonukleová kyselina
asRNA	antisense RNA
mRNA	mediátorová RNA
tRNA	transferová RNA
rRNA	ribonukleová RNA
sRNA	small RNA, malá RNA
ncRNA	nekódující RNA
cis-sRNA	*cis*-encoded small RNA
trans-sRNA	*trans*-encoded small RNA
3'UTR	3' nepřekládaná oblast, 3' untranslated region
5'UTR	5' nepřekládaná oblast, 5' untranslated region
RBP	RNA-vázající se protein, RNA-binding protein
RNáza	RNA ribonukleázy
DNáza	DNA ribonukleázy
RT	reverzní transkriptáza
PCR	polymerázová řetězová reakce
RT-qPCR	reversně transkripční kvantitativní PCR
ddPCR	dropletová digitální PCR
RNA-Seq	RNA sekvenování, RNA-Sequencing
SAGE	sériová analýza genové exprese
NGS	sekvenování nové generace
TGS	sekvenování třetí generace
RPKM	počet čtení na milion kilobází
FPKM	počet fragmentů na milion kilobází
TPM	počet transkriptů na milion kilobází
GITC	guanidin isothiokyanát
SMRT	jediná molekula v reálném čase, single molecule real time
PacBio	Pacific Biosciences
CMOS	komplementární polovodič oxidu kovu
ISFET	iontově senzitivní tranzistor
dNTP	deoxynukleosidtrifosfát
pH	potenciál vodíku
ZMW	zero-mode waveguide

NCBI	Národní centrum pro biotechnologické informace
SRA	Sequence Read Archive
pb	páry bází

Symboly:

H^+	vodíkový kation

SEZNAM PŘÍLOH

A. Tabulky

A.1 Tabulka detekovaných sRNA pro chromozom NC_011753.2 nástrojem Rockhopper

Č.	Start	Stop	Délka [pb]	Vlákno	Typ	Č.	Start	Stop	Délka [pb]	Vlákno	Typ	Č.	Start	Stop	Délka [pb]	Vlákno	Typ
1	6383	6399	17	-	*trans-*	35	1083418	1083590	173	+	*trans-*	69	2296300	2296325	26	-	*trans-*
2	26622	26659	38	-	*cis-*	36	1126240	1126337	98	+	*trans-*	70	2316099	2316205	107	-	*trans-*
3	37408	37445	38	-	*trans-*	37	1182999	1183012	14	+	*trans-*	71	2357045	2357277	233	-	*trans-*
4	58182	58198	17	+	*trans-*	38	1198040	1198078	39	-	*trans-*	72	2392147	2392223	77	-	*trans-*
5	68222	68387	166	+	*trans-*	39	1223152	1223228	77	-	*trans-*	73	2436114	2436172	59	-	*trans-*
6	104586	104766	181	+	*trans-*	40	1238171	1238198	28	-	*trans-*	74	2440018	2440045	28	-	*trans-*
7	120269	120311	43	+	*trans-*	41	1242857	1242902	46	+	*trans-*	75	2460118	2460171	54	-	*trans-*
8	143299	143341	43	+	*trans-*	42	1243651	1243688	38	+	*cis-*	76	2469929	2469967	39	+	*trans-*
9	180204	180249	46	+	*trans-*	43	1359903	1359925	23	+	*trans-*	77	2556709	2556734	26	+	*trans-*
10	180603	180640	38	-	*cis-*	44	1363490	1363515	26	+	*trans-*	78	2581355	2581368	14	-	*trans-*
11	205858	205924	67	+	*trans-*	45	1369432	1369504	73	+	*trans-*	79	2597678	2597687	10	+	*trans-*
12	214031	214042	12	+	*trans-*	46	1371945	1371973	29	+	*trans-*	80	2599163	2599200	38	-	*trans-*
13	214414	214536	123	+	*trans-*	47	1376754	1376778	25	+	*trans-*	81	2625987	2626012	26	+	*trans-*
14	219705	219741	37	+	*trans-*	48	1378297	1378324	28	+	*trans-*	82	2707127	2707218	92	-	*trans-*
15	236738	236761	24	+	*trans-*	49	1402627	1402664	38	+	*trans-*	83	2817437	2817459	23	+	*cis-*
16	286587	286692	106	-	*trans-*	50	1413522	1413559	38	-	*cis-*	84	2844214	2844299	86	-	*trans-*
17	316429	316454	26	+	*trans-*	51	1442567	1442621	55	-	*trans-*	85	2867756	2867798	43	+	*trans-*
18	389023	389048	26	+	*trans-*	52	1461450	1461485	36	-	*trans-*	86	2884580	2884632	53	-	*trans-*
19	449770	449815	46	-	*trans-*	53	1473515	1473571	57	-	*trans-*	87	2908632	2908669	38	+	*trans-*
20	463780	463849	70	+	*trans-*	54	1592376	1592388	13	-	*trans-*	88	2942529	2942565	37	+	*trans-*
21	516525	516559	35	+	*trans-*	55	1641961	1641990	30	+	*trans-*	89	2942892	2942901	10	-	*trans-*

22	537476	537593	118	+	*trans-*	56	1648352	1648365	14	-	*trans-*	90	2968051	2968088	38	+	*cis-*
23	570758	570789	32	+	*trans-*	57	1653400	1653413	14	-	*trans-*	91	2977722	2977829	108	+	*cis-*
24	583264	583381	118	+	*trans-*	58	1657921	1657952	32	+	*trans-*	92	3042839	3042866	28	+	*trans-*
25	599830	599845	16	-	*trans-*	59	1855525	1855608	84	-	*trans-*	93	3088293	3088330	38	-	*trans-*
26	618064	618095	32	-	*trans-*	60	1882083	1882169	87	+	*trans-*	94	3096202	3096225	24	-	*trans-*
27	712090	712099	10	-	*trans-*	61	1910209	1910238	30	-	*trans-*	95	3118741	3118828	88	+	*trans-*
28	771886	771909	24	-	*cis-*	62	1941297	1941334	38	+	*trans-*	96	3157159	3157175	17	-	*trans-*
29	904943	904958	16	+	*trans-*	63	2051318	2051327	10	-	*trans-*	97	3158614	3158660	47	-	*cis-*
30	922306	922369	64	+	*cis-*	64	2100673	2100710	38	-	*trans-*	98	3173828	3173891	64	+	*trans-*
31	977292	977329	38	+	*cis-*	65	2104977	2105014	38	-	*cis-*	99	3227651	3227710	60	-	*cis-*
32	1032121	1032158	38	+	*cis-*	66	2197148	2197158	11	-	*trans-*	100	3277153	3277190	38	+	*trans-*
33	1059241	1059278	38	-	*trans-*	67	2198239	2198275	37	-	*cis-*	101	3277365	3277418	54	+	*trans-*
34	1081000	1081019	20	+	*trans-*	68	2261285	2261389	105	+	*trans-*						

A.2 Tabulka detekovaných sRNA pro chromozom NC_011753.2 nástrojem DETR'PROK

Č.	Start	Stop	Délka [pb]	Vlákno	Typ	Č.	Start	Stop	Délka [pb]	Vlákno	Typ	Č.	Start	Stop	Délka [pb]	Vlákno	Typ
1	13	87	75	+	*trans-*	58	960538	960659	122	+	*cis-*	114	2022071	2022380	310	-	*trans-*
2	26587	26660	74	-	*cis-*	59	975697	975774	78	+	*trans-*	115	2088318	2088450	133	+	*trans-*
3	34845	35137	293	+	*trans-*	60	977225	977333	109	+	*cis-*	116	2095945	2096265	321	-	*trans-*
4	41018	41087	70	-	*trans-*	61	980122	980268	147	-	*cis-*	117	2149280	2149354	75	-	*trans-*
5	43815	44196	382	+	*trans-*	62	995937	996060	124	+	*cis-*	118	2175145	2175237	93	-	*trans-*
6	62317	62452	136	+	*trans-*	63	1031399	1031455	57	+	*trans-*	119	2197201	2197259	59	-	*trans-*
7	75699	75793	95	+	*trans-*	64	1031975	1032214	240	+	*cis-*	120	2198173	2198300	128	-	*cis-*
8	98212	98392	181	+	*trans-*	65	1043744	1043914	171	-	*trans-*	121	2231216	2231516	301	-	*trans-*
9	101779	101869	91	+	*trans-*	66	1074048	1074166	119	-	*trans-*	122	2306494	2306612	119	+	*cis-*
10	102209	102309	101	+	*cis-*	67	1077437	1077553	117	+	*trans-*	123	2315053	2315199	147	-	*trans-*

11	104585	104834	250	+	*trans-*	68	1108674	1108776	103	-	*cis-*	124	2327493	2327673	181	-	*cis-*
12	128532	128608	77	-	*cis-*	69	1114697	1114808	112	+	*trans-*	125	2349740	2350005	266	+	*cis-*
13	147837	148248	412	+	*cis-*	70	1139920	1139974	55	-	*trans-*	126	2355599	2355752	154	+	*cis-*
14	148657	148744	88	+	*cis-*	71	1184705	1184755	51	+	*trans-*	127	2366200	2366348	149	+	*trans-*
15	173080	173145	66	-	*trans-*	72	1197999	1198137	139	-	*trans-*	128	2440008	2440060	53	-	*trans-*
16	180566	180640	75	-	*cis-*	73	1223977	1224104	128	-	*cis-*	129	2449487	2449663	177	+	*trans-*
17	193587	193694	108	+	*cis-*	74	1238150	1238254	105	-	*trans-*	130	2449603	2449703	101	-	*trans-*
18	196450	196604	155	+	*trans-*	75	1239693	1239776	84	+	*cis-*	131	2453030	2453175	146	+	*trans-*
19	219213	219337	125	+	*trans-*	76	1243546	1243688	143	+	*cis-*	132	2459412	2459497	86	-	*cis-*
20	245183	245269	87	+	*trans-*	77	1275644	1275779	136	+	*cis-*	133	2460063	2460269	207	-	*cis-*
21	255742	255848	107	+	*trans-*	78	1335101	1335150	50	-	*trans-*	134	2460238	2460292	55	+	*trans-*
22	286552	286747	196	-	*trans-*	79	1359881	1360017	137	+	*trans-*	135	2530814	2530870	57	+	*cis-*
23	316001	316104	104	+	*cis-*	80	1389767	1390100	334	-	*trans-*	136	2569878	2570082	205	+	*trans-*
24	362303	362362	60	+	*cis-*	81	1390169	1390298	130	+	*trans-*	137	2590681	2590767	87	-	*trans-*
25	386853	386981	129	+	*trans-*	82	1402627	1402722	96	+	*trans-*	138	2599074	2599200	127	-	*trans-*
26	450867	451351	485	+	*trans-*	83	1428651	1428744	94	+	*cis-*	139	2625945	2626091	147	-	*cis-*
27	491191	491244	54	+	*trans-*	84	1428867	1429009	143	+	*cis-*	140	2634972	2635062	91	-	*trans-*
28	505021	505118	98	-	*trans-*	85	1461432	1461526	95	-	*cis-*	141	2651334	2651399	66	-	*trans-*
29	505021	505231	211	+	*trans-*	86	1470217	1470305	89	+	*trans-*	142	2685748	2685812	65	-	*trans-*
30	506402	506592	191	+	*trans-*	87	1472827	1473000	174	+	*cis-*	143	2685769	2685829	61	+	*trans-*
31	515259	515365	107	+	*trans-*	88	1473494	1473594	101	-	*trans-*	144	2817353	2817534	182	+	*cis-*
32	516427	516585	159	+	*trans-*	89	1473718	1473840	123	+	*trans-*	145	2844143	2844332	190	-	*trans-*
33	549430	549521	92	-	*trans-*	90	1519682	1519741	60	-	*trans-*	146	2884535	2884731	197	-	*trans-*
34	573741	573924	184	+	*cis-*	91	1529342	1529418	77	+	*cis-*	147	2885931	2886015	85	+	*trans-*
35	575367	575437	71	-	*trans-*	92	1534750	1534921	172	+	*trans-*	148	2886083	2886252	170	+	*trans-*
36	582579	582669	91	+	*trans-*	93	1601739	1601850	112	+	*cis-*	149	2908632	2908703	72	+	*trans-*
37	582584	582652	69	-	*trans-*	94	1626964	1627066	103	-	*trans-*	150	2931684	2931818	135	-	*trans-*

38	583264	583427	164	+	*trans-*	95	1641108	1641407	300	+	*cis-*	151	2942479	2942589	111	+	*trans-*
39	592865	592940	76	+	*trans-*	96	1641923	1642049	127	+	*trans-*	152	2942808	2942982	175	-	*trans-*
40	597581	597680	100	+	*trans-*	97	1646879	1646960	82	+	*cis-*	153	2956282	2956407	126	-	*trans-*
41	652597	652759	163	+	*trans-*	98	1648283	1648389	107	-	*trans-*	154	2977718	2977836	119	+	*cis-*
42	668773	669203	431	-	*trans-*	99	1653371	1653437	67	-	*trans-*	155	3048074	3048198	125	+	*cis-*
43	700722	700787	66	+	*trans-*	100	1689161	1689289	129	-	*cis-*	156	3081153	3081310	158	+	*trans-*
44	705571	705676	106	-	*cis-*	101	1704457	1704517	61	-	*trans-*	157	3118740	3118976	237	+	*cis-*
45	712276	712493	218	+	*cis-*	102	1757648	1757984	337	+	*trans-*	158	3120450	3120611	162	+	*trans-*
46	771769	771923	155	-	*cis-*	103	1794117	1794166	50	-	*trans-*	159	3146027	3146362	336	-	*trans-*
47	777524	777586	63	+	*trans-*	104	1804486	1804606	121	-	*trans-*	160	3158530	3158688	159	-	*cis-*
48	781460	781522	63	+	*trans-*	105	1847239	1847311	73	-	*trans-*	161	3173269	3173359	91	-	*trans-*
49	782469	782532	64	-	*trans-*	106	1855437	1855608	172	-	*trans-*	162	3173629	3173717	89	-	*trans-*
50	810079	810146	68	-	*trans-*	107	1877514	1877596	83	-	*trans-*	163	3173828	3173950	123	+	*trans-*
51	816192	816262	71	+	*trans-*	108	1882036	1882212	177	+	*trans-*	164	3204603	3204711	109	+	*trans-*
52	827913	828017	105	-	*trans-*	109	1896965	1897180	216	-	*trans-*	165	3204770	3204883	114	-	*trans-*
53	832962	833117	156	-	*trans-*	110	1936117	1936380	264	-	*trans-*	166	3210128	3210488	361	-	*trans-*
54	839679	839736	58	-	*trans-*	111	1985110	1985217	108	+	*trans-*	167	3227635	3227758	124	-	*cis-*
55	884131	884214	84	+	*trans-*	112	2015252	2015414	163	-	*trans-*	168	3250454	3250522	69	-	*trans-*
56	888201	888294	94	+	*trans-*	113	2021864	2021964	101	-	*trans-*	169	3252934	3253120	187	-	*trans-*
57	957362	957492	131	+	*cis-*												

A.3 Tabulka detekovaných sRNA pro chromozom NC_011753.2 nástrojem SEARCHsRNA

Č.	Start	Stop	Délka [pb]	Vlákno	Typ	Č.	Start	Stop	Délka [pb]	Vlákno	Typ	Č.	Start	Stop	Délka [pb]	Vlákno	Typ
1	43815	44195	381	+	*trans-*	54	1043744	1043799	56	-	*trans-*	107	2099584	2099646	63	-	*trans-*
2	62317	62452	136	+	*trans-*	55	1074048	1074162	115	-	*trans-*	108	2149303	2149354	52	-	*trans-*
3	98212	98351	140	+	*trans-*	56	1108684	1108776	93	-	*cis-*	109	2175171	2175234	64	-	*trans-*

4	101779	101869	91	+	*trans-*	57	1108804	1108874	71	-	*cis-*	110	2197201	2197257	57	-	*trans-*
5	102209	102309	101	+	*cis-*	58	1184705	1184755	51	+	*trans-*	111	2198173	2198300	128	-	*cis-*
6	104586	104825	240	+	*trans-*	59	1197999	1198133	135	-	*trans-*	112	2231240	2231413	174	-	*trans-*
7	128536	128608	73	-	*cis-*	60	1223977	1224104	128	-	*cis-*	113	2231444	2231516	73	-	*trans-*
8	147837	148143	307	+	*trans-*	61	1238150	1238248	99	-	*trans-*	114	2306498	2306602	105	+	*cis-*
9	148657	148744	88	+	*cis-*	62	1239693	1239776	84	+	*cis-*	115	2327496	2327673	178	-	*trans-*
10	173085	173141	57	-	*trans-*	63	1243546	1243688	143	+	*cis-*	116	2349740	2349949	210	+	*cis-*
11	180566	180640	75	-	*cis-*	64	1275644	1275771	128	+	*cis-*	117	2355599	2355752	154	+	*cis-*
12	193587	193694	108	+	*cis-*	65	1387305	1387368	64	-	*trans-*	118	2366200	2366342	143	+	*trans-*
13	214005	214549	545	+	*trans-*	66	1389768	1390100	333	-	*trans-*	119	2423284	2423394	111	+	*trans-*
14	245183	245266	84	+	*trans-*	67	1402627	1402705	79	+	*trans-*	120	2449487	2449651	165	+	*trans-*
15	286563	286747	185	-	*trans-*	68	1428651	1428737	87	+	*cis-*	121	2453883	2453940	58	+	*cis-*
16	316001	316104	104	+	*cis-*	69	1428867	1429001	135	+	*cis-*	122	2459412	2459497	86	-	*cis-*
17	318698	318768	71	-	*cis-*	70	1444616	1444704	89	-	*trans-*	123	2460082	2460234	153	-	*cis-*
18	362311	362362	52	+	*trans-*	71	1447888	1447971	84	+	*trans-*	124	2530814	2530870	57	+	*trans-*
19	463731	464341	611	+	*trans-*	72	1461432	1461526	95	-	*cis-*	125	2569878	2570082	205	+	*trans-*
20	486565	486904	340	-	*trans-*	73	1466029	1466081	53	+	*trans-*	126	2599074	2599200	127	-	*trans-*
21	488746	488906	161	-	*trans-*	74	1470217	1470305	89	+	*trans-*	127	2651334	2651398	65	-	*trans-*
22	489768	489876	109	-	*trans-*	75	1472827	1473000	174	+	*cis-*	128	2817353	2817534	182	+	*cis-*
23	490056	490136	81	-	*trans-*	76	1473494	1473594	101	-	*trans-*	129	2844210	2844332	123	-	*trans-*
24	490438	490568	131	-	*trans-*	77	1489307	1489411	105	+	*trans-*	130	2886127	2886252	126	+	*trans-*
25	490883	491010	128	-	*trans-*	78	1508056	1508106	51	-	*cis-*	131	2891856	2892173	318	-	*trans-*
26	491191	491244	54	+	*trans-*	79	1519198	1519285	88	+	*cis-*	132	2897368	2897949	582	+	*trans-*
27	505153	505205	53	+	*trans-*	80	1519682	1519739	58	-	*trans-*	133	2898144	2898462	319	+	*trans-*
28	506455	506511	57	+	*trans-*	81	1529342	1529418	77	+	*cis-*	134	2898504	2898649	146	+	*trans-*
29	516427	516585	159	+	*trans-*	82	1530073	1530174	102	-	*trans-*	135	2898678	2898782	105	+	*trans-*
30	518231	518284	54	+	*cis-*	83	1574754	1574808	55	-	*cis-*	136	2898848	2899033	186	+	*trans-*

31	573741	573858	118	+	*cis-*	84	1583706	1583762	57	+	*trans-*	137	2908632	2908703	72	+	*trans-*
32	582579	582669	91	+	*trans-*	85	1600503	1600563	61	-	*trans-*	138	2942479	2942580	102	+	*trans-*
33	582584	582642	59	-	*trans-*	86	1601743	1601831	89	+	*cis-*	139	2942830	2942982	153	-	*trans-*
34	583264	583420	157	+	*trans-*	87	1623914	1623977	64	+	*trans-*	140	3008041	3008108	68	-	*cis-*
35	592865	592921	57	+	*trans-*	88	1625189	1625680	492	-	*trans-*	141	3048074	3048197	124	+	*cis-*
36	597581	597663	83	+	*trans-*	89	1625947	1626102	156	-	*trans-*	142	3118740	3118874	135	+	*trans-*
37	705604	705676	73	-	*trans-*	90	1626964	1627066	103	-	*trans-*	143	3118897	3118976	80	+	*cis-*
38	712357	712493	137	+	*cis-*	91	1641108	1641341	234	+	*cis-*	144	3146030	3146160	131	-	*trans-*
39	771769	771923	155	-	*cis-*	92	1646879	1647021	143	+	*cis-*	145	3146190	3146362	173	-	*trans-*
40	783032	783240	209	-	*trans-*	93	1648283	1648389	107	-	*trans-*	146	3158558	3158688	131	-	*cis-*
41	783366	783418	53	-	*trans-*	94	1653371	1653437	67	-	*trans-*	147	3173269	3173359	91	-	*trans-*
42	816201	816262	62	+	*trans-*	95	1689094	1689289	196	-	*trans-*	148	3173828	3173940	113	+	*trans-*
43	817510	817606	97	+	*cis-*	96	1722315	1722366	52	-	*cis-*	149	3204624	3204696	73	+	*trans-*
44	827924	828017	94	-	*trans-*	97	1757648	1757980	333	+	*trans-*	150	3210128	3210272	145	-	*trans-*
45	832964	833117	154	-	*trans-*	98	1855519	1855608	90	-	*trans-*	151	3210300	3210488	189	-	*trans-*
46	839679	839736	58	-	*trans-*	99	1916668	1916863	196	+	*trans-*	152	3227337	3227853	517	+	*trans-*
47	957362	957483	122	+	*trans-*	100	1976301	1976351	51	+	*cis-*	153	3227446	3227521	76	-	*trans-*
48	960540	960657	118	+	*cis-*	101	1985110	1985217	108	+	*trans-*	154	3227635	3227710	76	-	*trans-*
49	977249	977333	85	+	*cis-*	102	2015272	2015414	143	-	*trans-*	155	3244443	3244614	172	-	*trans-*
50	979778	979832	55	-	*cis-*	103	2022071	2022376	306	-	*trans-*	156	3250457	3250522	66	-	*trans-*
51	980182	980268	87	-	*cis-*	104	2095945	2096265	321	-	*trans-*	157	3253028	3253120	93	-	*trans-*
52	995937	996045	109	+	*cis-*	105	2098638	2098837	200	-	*trans-*	158	3271008	3271098	91	-	*trans-*
53	1031975	1032194	220	+	*cis-*	106	2098985	2099101	117	-	*trans-*	159	3277153	3277308	156	+	*trans-*

A.4 Tabulka detekovaných sRNA pro chromozom NC_011744.2 nástrojem Rockhopper

Č.	Start	Stop	Délka [pb]	Vlákno	Typ	Č.	Start	Stop	Délka [pb]	Vlákno	Typ	Č.	Start	Stop	Délka [pb]	Vlákno	Typ
1	31	102	72	-	*trans-*	152	539458	539495	38	+	*trans-*	302	1230415	1230449	35	+	*cis-*
2	12869	12886	18	+	*cis-*	153	541302	541313	12	-	*trans-*	303	1230728	1230741	14	+	*cis-*
3	15899	15930	32	+	*cis-*	154	546455	546475	21	-	*trans-*	304	1236799	1236848	50	+	*cis-*
4	18612	18728	117	-	*trans-*	155	547158	547195	38	-	*cis-*	305	1236999	1237095	97	-	*trans-*
5	18823	18848	26	+	*trans-*	156	550261	550298	38	+	*cis-*	306	1239607	1239821	215	-	*trans-*
6	27429	27503	75	+	*trans-*	157	550909	550950	42	-	*trans-*	307	1242954	1242966	13	-	*trans-*
7	27924	27961	38	-	*cis-*	158	552431	552468	38	-	*trans-*	308	1244724	1244927	204	+	*cis-*
8	29092	29185	94	+	*trans-*	159	555762	555799	38	+	*trans-*	309	1244928	1244935	8	-	*trans-*
9	33650	33831	182	-	*trans-*	160	556397	556434	38	-	*cis-*	310	1247166	1247203	38	-	*trans-*
10	36824	36860	37	+	*cis-*	161	562256	562293	38	+	*trans-*	311	1249350	1249367	18	-	*trans-*
11	38998	39035	38	+	*cis-*	162	565371	565408	38	-	*trans-*	312	1250794	1250831	38	-	*cis-*
12	39453	39563	111	+	*cis-*	163	579678	579806	129	+	*cis-*	313	1251544	1251564	21	-	*trans-*
13	40756	40815	60	+	*trans-*	164	580116	580242	127	+	*cis-*	314	1252641	1252652	12	+	*trans-*
14	42314	42351	38	-	*trans-*	165	580308	580334	27	+	*cis-*	315	1264692	1264724	33	+	*trans-*
15	55708	55745	38	+	*cis-*	166	581212	581237	26	+	*cis-*	316	1273249	1273409	161	+	*trans-*
16	56749	56777	29	+	*cis-*	167	598367	598404	38	+	*cis-*	317	1277673	1277739	67	-	*trans-*
17	57220	57305	86	+	*trans-*	168	607151	607167	17	+	*trans-*	318	1280743	1280781	39	-	*trans-*
18	60066	60077	12	+	*trans-*	169	611977	612076	100	+	*trans-*	319	1280999	1281023	25	+	*cis-*
19	60764	60806	43	-	*trans-*	170	615965	616002	38	+	*trans-*	320	1282759	1282781	23	+	*cis-*
20	60947	61039	93	+	*trans-*	171	618213	618243	31	-	*cis-*	321	1294758	1294777	20	-	*trans-*
21	77600	77637	38	+	*cis-*	172	631107	631139	33	-	*trans-*	322	1296701	1296738	38	+	*trans-*
22	80625	80637	13	+	*trans-*	173	636853	636872	20	-	*trans-*	323	1297756	1297794	39	+	*cis-*
23	82556	82575	20	+	*trans-*	174	638715	638922	208	+	*trans-*	324	1300080	1300109	30	-	*trans-*
24	84953	84990	38	-	*trans-*	175	645164	645220	57	+	*trans-*	325	1301258	1301295	38	+	*trans-*

25	85114	85141	28	-	*trans-*	176	651594	651622	29	-	*trans-*	326	1317197	1317274	78	+	*trans-*
26	86423	86460	38	-	*cis-*	177	652051	652088	38	-	*trans-*	327	1317796	1317943	148	+	*trans-*
27	86609	86697	89	-	*cis-*	178	652501	652525	25	+	*cis-*	328	1324236	1324277	42	+	*trans-*
28	93527	93564	38	+	*trans-*	179	661765	661783	19	+	*cis-*	329	1326600	1326667	68	-	*trans-*
29	94553	94660	108	+	*trans-*	180	669271	669338	68	+	*trans-*	330	1327538	1327575	38	+	*trans-*
30	97976	98013	38	+	*trans-*	181	674157	674194	38	-	*trans-*	331	1330216	1330292	77	+	*trans-*
31	103487	103549	63	+	*trans-*	182	676884	676924	41	-	*trans-*	332	1330888	1330949	62	-	*cis-*
32	106104	106126	23	+	*trans-*	183	679594	679604	11	-	*trans-*	333	1331472	1331494	23	-	*cis-*
33	106749	106825	77	+	*trans-*	184	679776	679804	29	+	*trans-*	334	1334628	1335182	555	-	*trans-*
34	109470	109507	38	+	*trans-*	185	683308	683345	38	-	*cis-*	335	1338297	1338334	38	+	*trans-*
35	112588	112607	20	+	*trans-*	186	687155	687192	38	+	*trans-*	336	1352210	1352239	30	+	*cis-*
36	113698	113720	23	-	*trans-*	187	687373	687410	38	-	*trans-*	337	1354406	1354443	38	-	*cis-*
37	113727	113758	32	+	*trans-*	188	687425	687496	72	+	*trans-*	338	1355397	1355494	98	+	*trans-*
38	126220	126280	61	-	*trans-*	189	690000	690038	39	+	*trans-*	339	1357340	1357355	16	+	*trans-*
39	140804	140841	38	+	*trans-*	190	692089	692126	38	+	*trans-*	340	1359445	1359481	37	+	*trans-*
40	142537	142574	38	-	*cis-*	191	692191	692228	38	+	*trans-*	341	1364952	1365034	83	-	*cis-*
41	144661	144721	61	-	*trans-*	192	695153	695222	70	-	*trans-*	342	1366612	1366649	38	+	*trans-*
42	146016	146129	114	+	*cis-*	193	695827	695845	19	+	*trans-*	343	1384180	1384193	14	+	*cis-*
43	154654	154691	38	+	*cis-*	194	695946	695983	38	+	*trans-*	344	1384566	1384651	86	+	*cis-*
44	161788	161825	38	+	*trans-*	195	697469	697506	38	+	*cis-*	345	1390669	1390717	49	-	*trans-*
45	161826	162090	265	-	*trans-*	196	699551	699669	119	+	*cis-*	346	1390795	1390879	85	-	*cis-*
46	162091	162128	38	+	*trans-*	197	699811	699858	48	+	*cis-*	347	1397987	1398146	160	+	*trans-*
47	162129	162167	39	-	*trans-*	198	700002	700039	38	-	*trans-*	348	1401778	1401815	38	-	*cis-*
48	173266	173303	38	-	*cis-*	199	711526	711563	38	+	*cis-*	349	1404619	1404656	38	+	*cis-*
49	175911	175979	69	+	*cis-*	200	735309	735346	38	+	*trans-*	350	1412787	1412804	18	+	*trans-*
50	177677	177714	38	+	*cis-*	201	735563	735582	20	+	*trans-*	351	1412907	1412944	38	+	*trans-*
51	177994	178031	38	+	*trans-*	202	741194	741207	14	-	*trans-*	352	1412968	1413005	38	-	*trans-*

52	179480	179517	38	+	*cis-*	203	741336	741363	28	+	*trans-*	353	1414049	1414071	23	+	*trans-*
53	181321	181354	34	+	*trans-*	204	753761	753777	17	-	*trans-*	354	1415073	1415103	31	+	*cis-*
54	187061	187098	38	-	*trans-*	205	763742	763757	16	-	*trans-*	355	1418809	1418835	27	-	*trans-*
55	193002	193024	23	+	*trans-*	206	765243	765274	32	-	*trans-*	356	1423698	1423816	119	-	*trans-*
56	200483	200520	38	-	*trans-*	207	765777	765795	19	-	*trans-*	357	1424036	1424112	77	+	*trans-*
57	200709	200792	84	-	*trans-*	208	767276	767349	74	+	*trans-*	358	1424420	1424431	12	-	*cis-*
58	201370	201407	38	-	*trans-*	209	769215	769231	17	+	*cis-*	359	1424507	1424523	17	-	*cis-*
59	202658	202673	16	-	*trans-*	210	778978	779043	66	+	*trans-*	360	1425056	1425239	184	-	*trans-*
60	209998	210035	38	+	*trans-*	211	779571	779608	38	-	*cis-*	361	1425713	1425764	52	-	*cis-*
61	215309	215346	38	+	*cis-*	212	784125	784158	34	-	*trans-*	362	1426688	1426764	77	-	*cis-*
62	215509	215546	38	+	*cis-*	213	784868	784891	24	+	*cis-*	363	1430361	1430398	38	-	*trans-*
63	223049	223086	38	+	*cis-*	214	791969	792061	93	+	*trans-*	364	1431442	1431479	38	-	*trans-*
64	225777	225814	38	+	*cis-*	215	792598	792635	38	+	*cis-*	365	1441486	1441523	38	+	*trans-*
65	226782	226801	20	+	*trans-*	216	794640	794677	38	-	*cis-*	366	1450147	1450171	25	+	*trans-*
66	227795	227832	38	-	*trans-*	217	796147	796184	38	-	*trans-*	367	1453236	1453278	43	+	*trans-*
67	239929	239944	16	+	*trans-*	218	796400	796524	125	+	*trans-*	368	1457622	1457640	19	+	*trans-*
68	249128	249149	22	-	*trans-*	219	797274	797311	38	+	*trans-*	369	1459075	1459112	38	-	*trans-*
69	253545	253664	120	-	*trans-*	220	798861	798898	38	-	*cis-*	370	1461053	1461068	16	+	*trans-*
70	255307	255431	125	-	*cis-*	221	800622	800659	38	+	*trans-*	371	1462926	1462959	34	+	*trans-*
71	258012	258166	155	-	*trans-*	222	805398	805435	38	-	*trans-*	372	1463775	1463803	29	-	*trans-*
72	258655	258664	10	-	*trans-*	223	807084	807121	38	+	*cis-*	373	1465728	1465765	38	+	*trans-*
73	259654	259748	95	-	*trans-*	224	808294	808416	123	-	*trans-*	374	1467055	1467092	38	-	*cis-*
74	259899	259936	38	+	*trans-*	225	811421	811505	85	-	*trans-*	375	1473331	1473415	85	-	*trans-*
75	263636	263673	38	+	*cis-*	226	812148	812185	38	-	*trans-*	376	1475318	1475334	17	-	*cis-*
76	268204	268213	10	-	*trans-*	227	816273	816310	38	-	*cis-*	377	1480568	1480605	38	+	*cis-*
77	271804	271841	38	+	*cis-*	228	818524	818537	14	-	*trans-*	378	1482537	1482574	38	+	*trans-*
78	285792	285829	38	-	*cis-*	229	820949	821033	85	+	*trans-*	379	1483205	1483242	38	+	*trans-*

79	289592	289611	20	-	*cis-*	230	825816	825853	38	+	*cis-*	380	1487591	1487681	91	+	*trans-*
80	289662	289699	38	-	*cis-*	231	827975	828012	38	+	*cis-*	381	1489562	1489594	33	+	*trans-*
81	289808	289886	79	-	*cis-*	232	828051	828113	63	-	*cis-*	382	1490449	1490464	16	+	*trans-*
82	300005	300139	135	+	*cis-*	233	838479	838506	28	-	*cis-*	383	1491772	1491809	38	+	*trans-*
83	300760	300781	22	+	*cis-*	234	838963	838973	11	-	*trans-*	384	1493434	1493490	57	-	*cis-*
84	302380	302417	38	+	*trans-*	235	842524	842633	110	+	*cis-*	385	1493869	1493900	32	-	*trans-*
85	305380	305417	38	+	*trans-*	236	843206	843227	22	-	*cis-*	386	1493939	1494103	165	+	*trans-*
86	317351	317388	38	+	*cis-*	237	848906	848943	38	-	*trans-*	387	1496497	1496525	29	-	*trans-*
87	317424	317448	25	-	*trans-*	238	863627	863664	38	+	*trans-*	388	1501617	1501647	31	+	*trans-*
88	317615	317626	12	+	*cis-*	239	871555	871592	38	-	*cis-*	389	1503267	1503304	38	-	*cis-*
89	319807	319840	34	+	*trans-*	240	877108	877130	23	-	*cis-*	390	1503855	1503892	38	-	*trans-*
90	320529	320566	38	+	*cis-*	241	880696	880733	38	+	*cis-*	391	1509227	1509350	124	+	*trans-*
91	325320	325357	38	+	*cis-*	242	881612	881668	57	-	*trans-*	392	1509820	1509907	88	-	*cis-*
92	333902	333931	30	+	*trans-*	243	881987	882060	74	-	*cis-*	393	1510098	1510121	24	+	*trans-*
93	335481	335518	38	+	*trans-*	244	889043	889080	38	+	*trans-*	394	1510171	1510255	85	-	*trans-*
94	336205	336235	31	+	*trans-*	245	893088	893118	31	-	*trans-*	395	1513072	1513126	55	-	*cis-*
95	337991	338028	38	+	*cis-*	246	900456	900533	78	-	*trans-*	396	1515055	1515113	59	+	*trans-*
96	341524	341701	178	-	*trans-*	247	905647	905684	38	+	*trans-*	397	1515309	1515346	38	-	*trans-*
97	344322	344359	38	-	*trans-*	248	906285	906347	63	-	*trans-*	398	1522219	1522233	15	+	*cis-*
98	344387	344407	21	+	*trans-*	249	913218	913255	38	+	*cis-*	399	1522883	1522920	38	+	*cis-*
99	358151	358204	54	-	*trans-*	250	923736	923848	113	-	*cis-*	400	1523446	1523483	38	-	*cis-*
100	362501	362538	38	+	*cis-*	251	925439	925476	38	-	*cis-*	401	1531009	1531026	18	-	*trans-*
101	362749	362814	66	+	*trans-*	252	929659	929695	37	-	*cis-*	402	1540621	1540702	82	+	*trans-*
102	366471	366572	102	-	*trans-*	253	930391	930430	40	+	*trans-*	403	1540860	1540897	38	+	*trans-*
103	368935	369022	88	+	*trans-*	254	964360	964397	38	-	*cis-*	404	1547373	1547425	53	-	*cis-*
104	369393	369430	38	+	*cis-*	255	970809	970822	14	-	*trans-*	405	1547945	1547982	38	-	*trans-*
105	374198	374253	56	+	*trans-*	256	970842	970879	38	+	*trans-*	406	1548099	1548136	38	+	*trans-*

106	374936	374973	38	+	*cis-*	257	973796	973810	15	-	*trans-*	407	1549351	1549404	54	+	*trans-*
107	381579	381611	33	-	*trans-*	258	992760	992778	19	-	*trans-*	408	1549546	1549607	62	+	*trans-*
108	381738	381821	84	+	*trans-*	259	996460	996494	35	-	*trans-*	409	1550011	1550042	32	+	*trans-*
109	385314	385351	38	-	*trans-*	260	996571	996608	38	+	*cis-*	410	1554184	1554221	38	-	*trans-*
110	390442	390494	53	-	*trans-*	261	998760	998770	11	+	*trans-*	411	1558383	1558420	38	-	*trans-*
111	394031	394068	38	+	*cis-*	262	1003152	1003189	38	-	*trans-*	412	1558421	1558470	50	+	*trans-*
112	399078	399132	55	-	*cis-*	263	1008502	1008540	39	-	*trans-*	413	1563002	1563024	23	+	*trans-*
113	401995	402032	38	-	*trans-*	264	1008815	1008870	56	+	*cis-*	414	1564285	1564311	27	-	*cis-*
114	405331	405368	38	-	*trans-*	265	1011464	1011501	38	+	*trans-*	415	1574838	1574875	38	-	*trans-*
115	409147	409184	38	+	*trans-*	266	1039021	1039079	59	-	*trans-*	416	1576476	1576502	27	+	*trans-*
116	412127	412137	11	+	*trans-*	267	1047653	1047690	38	+	*trans-*	417	1576557	1576594	38	-	*cis-*
117	418813	418824	12	+	*trans-*	268	1057698	1057741	44	-	*trans-*	418	1576852	1576886	35	-	*cis-*
118	420610	420647	38	+	*trans-*	269	1079555	1079651	97	-	*trans-*	419	1576944	1576981	38	-	*trans-*
119	420839	420876	38	+	*cis-*	270	1081928	1081951	24	-	*trans-*	420	1577078	1577110	33	-	*trans-*
120	421558	421587	30	+	*trans-*	271	1089373	1089410	38	+	*cis-*	421	1577145	1577182	38	+	*trans-*
121	422969	423006	38	+	*trans-*	272	1091644	1091680	37	+	*cis-*	422	1579361	1579398	38	-	*cis-*
122	424178	424195	18	+	*trans-*	273	1093090	1093127	38	-	*trans-*	423	1579544	1579610	67	-	*trans-*
123	424382	424419	38	+	*cis-*	274	1093373	1093392	20	-	*trans-*	424	1580579	1580592	14	-	*trans-*
124	424683	424720	38	-	*cis-*	275	1095176	1095197	22	-	*cis-*	425	1580658	1580775	118	-	*trans-*
125	425742	425759	18	+	*trans-*	276	1096061	1096098	38	-	*cis-*	426	1592137	1592198	62	+	*trans-*
126	434197	434234	38	+	*cis-*	277	1096217	1096303	87	-	*cis-*	427	1597904	1597941	38	+	*trans-*
127	435318	435350	33	+	*trans-*	278	1098795	1098827	33	-	*cis-*	428	1603211	1603295	85	-	*trans-*
128	436686	436781	96	+	*trans-*	279	1107952	1107989	38	+	*trans-*	429	1608638	1608675	38	+	*trans-*
129	440610	440647	38	+	*trans-*	280	1114700	1114736	37	-	*trans-*	430	1611263	1611300	38	-	*trans-*
130	443457	443499	43	+	*trans-*	281	1127210	1127247	38	-	*trans-*	431	1615544	1615589	46	-	*trans-*
131	458736	458853	118	+	*trans-*	282	1129109	1129118	10	+	*cis-*	432	1615689	1615724	36	-	*trans-*
132	460567	460610	44	-	*trans-*	283	1137685	1137735	51	-	*trans-*	433	1616341	1616378	38	+	*cis-*

133	461554	461635	82	-	*cis-*	284	1137905	1137999	95	+	*trans-*	434	1616860	1616879	20	+	*trans-*
134	461885	461922	38	+	*cis-*	285	1145937	1145974	38	+	*cis-*	435	1627579	1627616	38	-	*trans-*
135	466204	466241	38	+	*trans-*	286	1146125	1146181	57	+	*cis-*	436	1628045	1628082	38	+	*cis-*
136	484687	484724	38	-	*trans-*	287	1150169	1150202	34	-	*trans-*	437	1634042	1634077	36	-	*cis-*
137	484725	484775	51	+	*trans-*	288	1163646	1163658	13	+	*trans-*	438	1634557	1634582	26	-	*cis-*
138	488689	488726	38	-	*trans-*	289	1181985	1182021	37	-	*trans-*	439	1637368	1637381	14	-	*cis-*
139	491928	491965	38	+	*trans-*	290	1182168	1182205	38	-	*cis-*	440	1639812	1639843	32	+	*cis-*
140	493401	493438	38	+	*trans-*	291	1193518	1193604	87	+	*trans-*	441	1642765	1642794	30	+	*cis-*
141	493518	493581	64	+	*trans-*	292	1210086	1210117	32	-	*trans-*	442	1646422	1646461	40	-	*cis-*
142	495088	495102	15	-	*trans-*	293	1210118	1210155	38	+	*trans-*	443	1646557	1646600	44	+	*trans-*
143	499013	499050	38	+	*cis-*	294	1210192	1210217	26	-	*trans-*	444	1651605	1651660	56	-	*trans-*
144	501006	501043	38	+	*cis-*	295	1210218	1210609	392	+	*trans-*	445	1651845	1651940	96	-	*cis-*
145	504466	504503	38	+	*trans-*	296	1210610	1210629	20	-	*trans-*	446	1657903	1657937	35	+	*trans-*
146	510194	510231	38	+	*trans-*	297	1210673	1210710	38	+	*trans-*	447	1663781	1663819	39	-	*trans-*
147	515262	515371	110	+	*trans-*	298	1210820	1210857	38	+	*cis-*	448	1666932	1666969	38	+	*cis-*
148	517978	517997	20	-	*trans-*	299	1211069	1211106	38	+	*cis-*	449	1669240	1669253	14	+	*trans-*
149	524000	524037	38	+	*trans-*	300	1211445	1211634	190	-	*trans-*	450	1669776	1669886	111	-	*trans-*
150	524110	524147	38	+	*trans-*	301	1217581	1217610	30	+	*trans-*	451	1672446	1672488	43	-	*trans-*
151	528451	528480	30	-	*cis-*												

A.5 Tabulka detekovaných sRNA pro chromozom NC_011744.2 nástrojem DETR'PROK

Č.	**Start**	**Stop**	**Délka [pb]**	**Vlákno**	**Typ**	**Č.**	**Start**	**Stop**	**Délka [pb]**	**Vlákno**	**Typ**	**Č.**	**Start**	**Stop**	**Délka [pb]**	**Vlákno**	**Typ**
1	23	121	99	-	*trans-*	36	687404	687498	95	+	*trans-*	71	1273217	1273416	200	+	*trans-*
2	18605	18728	124	-	*trans-*	37	689979	690040	62	+	*trans-*	72	1277664	1277739	76	-	*trans-*
3	39450	39565	116	+	*cis-*	38	695144	695222	79	-	*trans-*	73	1317796	1317947	152	+	*trans-*

4	40732	40820	89	+	*trans-*	39	695787	696038	252	+	*trans-*	74	1324236	1324299	64	+	*trans-*
5	57217	57308	92	+	*trans-*	40	699782	699883	102	+	*cis-*	75	1326595	1326667	73	-	*trans-*
6	86604	86697	94	-	*cis-*	41	700002	700055	54	-	*trans-*	76	1330214	1330315	102	+	*trans-*
7	93517	93571	55	+	*trans-*	42	721839	721904	66	+	*cis-*	77	1330880	1330976	97	-	*cis-*
8	103487	103567	81	+	*trans-*	43	765237	765298	62	-	*trans-*	78	1334582	1335017	436	-	*trans-*
9	144661	144756	96	-	*trans-*	44	791949	792044	96	-	*trans-*	79	1335046	1335191	146	-	*cis-*
10	175911	175986	76	+	*cis-*	45	791955	792084	130	+	*trans-*	80	1359442	1359499	58	+	*trans-*
11	192987	193079	93	+	*trans-*	46	796380	796442	63	+	*trans-*	81	1366551	1366649	99	+	*trans-*
12	199054	199132	79	+	*cis-*	47	811397	811505	109	-	*trans-*	82	1378675	1378779	105	-	*trans-*
13	200678	200792	115	-	*trans-*	48	818395	818561	167	-	*trans-*	83	1412902	1413031	130	+	*trans-*
14	253522	253715	194	-	*trans-*	49	820949	821035	87	+	*trans-*	84	1425041	1425252	212	-	*trans-*
15	255236	255431	196	-	*cis-*	50	838896	839000	105	-	*trans-*	85	1425692	1425781	90	-	*cis-*
16	258000	258194	195	-	*trans-*	51	848906	848961	56	-	*trans-*	86	1426656	1426822	167	-	*cis-*
17	259614	259757	144	-	*trans-*	52	900455	900607	153	-	*trans-*	87	1482537	1482613	77	+	*trans-*
18	300005	300139	135	+	*cis-*	53	923705	923874	170	-	*cis-*	88	1487585	1487742	158	+	*trans-*
19	335481	335584	104	+	*trans-*	54	1038947	1039141	195	-	*trans-*	89	1493417	1493548	132	-	*cis-*
20	374172	374263	92	+	*trans-*	55	1096163	1096331	169	-	*cis-*	90	1509815	1509911	97	-	*cis-*
21	421527	421659	133	+	*trans-*	56	1137612	1137735	124	-	*trans-*	91	1513059	1513201	143	-	*cis-*
22	422898	423006	109	+	*trans-*	57	1137862	1138007	146	+	*trans-*	92	1515055	1515130	76	+	*trans-*
23	436667	436830	164	+	*trans-*	58	1145937	1145999	63	+	*cis-*	93	1540621	1540737	117	+	*trans-*
24	458736	458862	127	+	*trans-*	59	1146125	1146193	69	+	*cis-*	94	1547929	1547982	54	-	*trans-*
25	460549	460610	62	-	*trans-*	60	1150161	1150219	59	-	*trans-*	95	1549303	1549404	102	+	*trans-*
26	484636	484724	89	-	*trans-*	61	1210072	1210868	797	+	*cis-*	96	1558402	1558478	77	+	*trans-*
27	484721	484857	137	+	*trans-*	62	1210152	1210270	119	-	*trans-*	97	1568862	1569168	307	+	*trans-*
28	491907	491965	59	+	*trans-*	63	1210569	1210645	77	-	*trans-*	98	1577036	1577134	99	-	*trans-*
29	504655	504754	100	-	*trans-*	64	1211592	1211643	52	-	*trans-*	99	1579497	1579627	131	-	*trans-*
30	523995	524044	50	+	*trans-*	65	1217573	1217653	81	+	*trans-*	100	1580658	1580775	118	-	*trans-*

31	555744	555808	65	+	*trans-*	66	1228247	1228406	160	-	*trans-*	101	1669189	1669277	89	+	*trans-*
32	565323	565468	146	-	*trans-*	67	1236763	1236850	88	+	*cis-*	102	1669762	1669886	125	-	*trans-*
33	638715	638972	258	+	*trans-*	68	1236901	1237100	200	-	*trans-*	103	1672417	1672562	146	-	*trans-*
34	669226	669367	142	+	*trans-*	69	1239607	1239848	242	-	*trans-*	104	1675157	1675206	50	+	*trans-*
35	687364	687429	66	-	*trans-*	70	1244724	1244937	214	+	*cis-*						

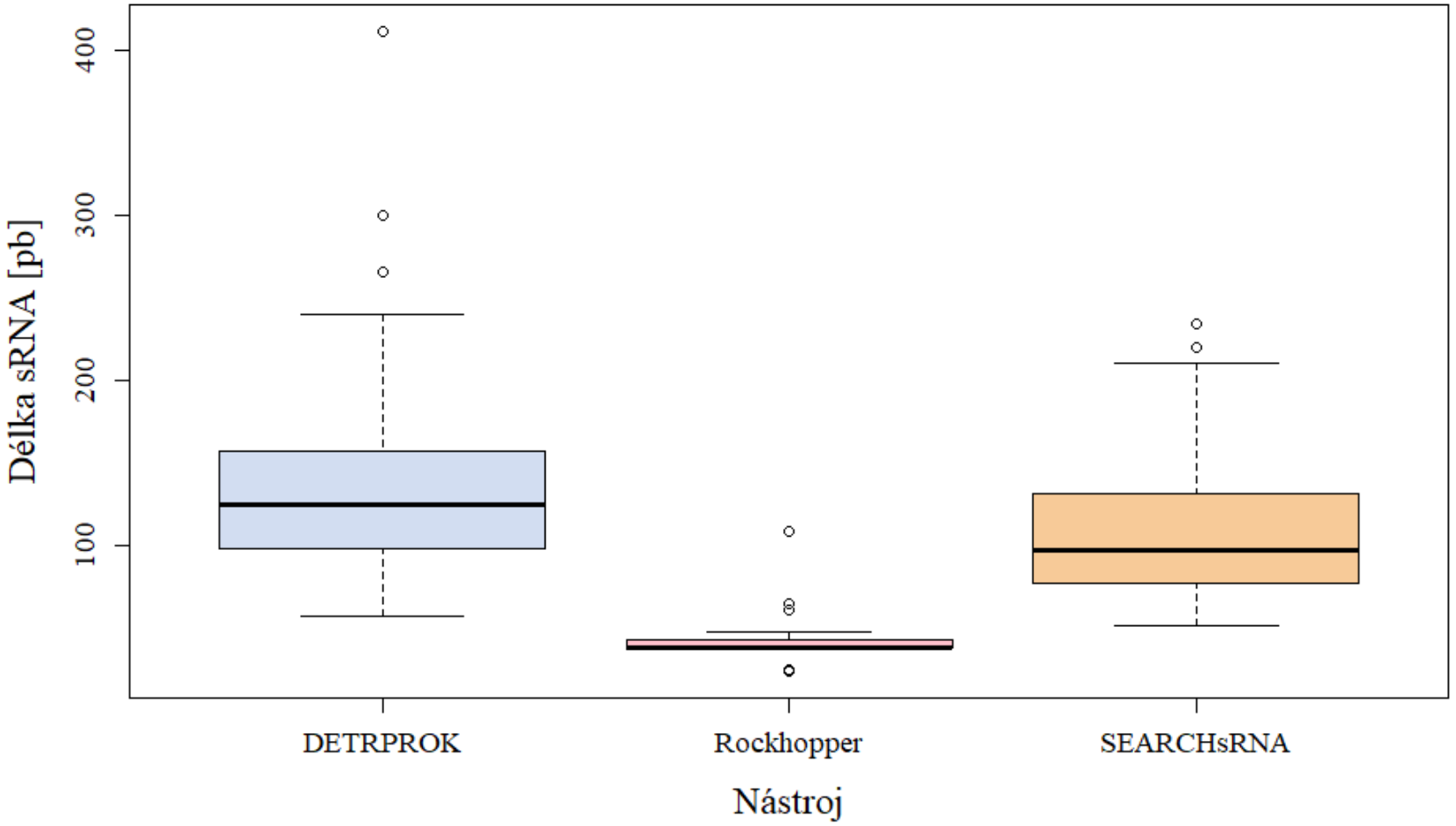
A.6 Tabulka detekovaných sRNA pro chromozom NC_011744.2 nástrojem SEARCHsRNA

Č.	Start	Stop	Délka [pb]	Vlákno	Typ	Č.	Start	Stop	Délka [pb]	Vlákno	Typ	Č.	Start	Stop	Délka [pb]	Vlákno	Typ
1	24	121	98	-	*trans-*	26	695144	695222	79	-	*trans-*	51	1280717	1280784	68	-	*trans-*
2	18605	18728	124	-	*trans-*	27	699786	699883	98	+	*cis-*	52	1326600	1326667	68	-	*trans-*
3	39452	39565	114	+	*cis-*	28	700002	700055	54	-	*trans-*	53	1330214	1330292	79	+	*trans-*
4	57217	57308	92	+	*trans-*	29	721839	721896	58	+	*cis-*	54	1330883	1330949	67	-	*cis-*
5	86604	86697	94	-	*cis-*	30	791949	792044	96	-	*trans-*	55	1334582	1335014	433	-	*trans-*
6	93517	93571	55	+	*trans-*	31	791960	792062	103	+	*trans-*	56	1335046	1335182	137	-	*trans-*
7	144661	144756	96	-	*trans-*	32	853981	854079	99	-	*trans-*	57	1382172	1382242	71	-	*trans-*
8	175911	175986	76	+	*cis-*	33	855379	855469	91	-	*trans-*	58	1390667	1390718	52	-	*trans-*
9	199054	199132	79	+	*trans-*	34	855809	855875	67	-	*trans-*	59	1425041	1425252	212	-	*trans-*
10	200678	200792	115	-	*trans-*	35	900456	900546	91	-	*trans-*	60	1425692	1425781	90	-	*cis-*
11	255236	255431	196	-	*cis-*	36	901627	901838	212	-	*trans-*	61	1426657	1426822	166	-	*cis-*
12	258000	258194	195	-	*trans-*	37	923734	923874	141	-	*cis-*	62	1493420	1493548	129	-	*cis-*
13	259623	259757	135	-	*trans-*	38	1096217	1096331	115	-	*cis-*	63	1506389	1506445	57	-	*trans-*
14	300005	300139	135	+	*cis-*	39	1128524	1128685	162	-	*trans-*	64	1509815	1509911	97	-	*cis-*
15	335481	335540	60	+	*trans-*	40	1137612	1137735	124	-	*trans-*	65	1513063	1513201	139	-	*cis-*
16	373615	373680	66	-	*trans-*	41	1137862	1138007	146	+	*trans-*	66	1547929	1547982	54	-	*trans-*
17	374198	374263	66	+	*trans-*	42	1145937	1145999	63	+	*cis-*	67	1549349	1549404	56	+	*trans-*

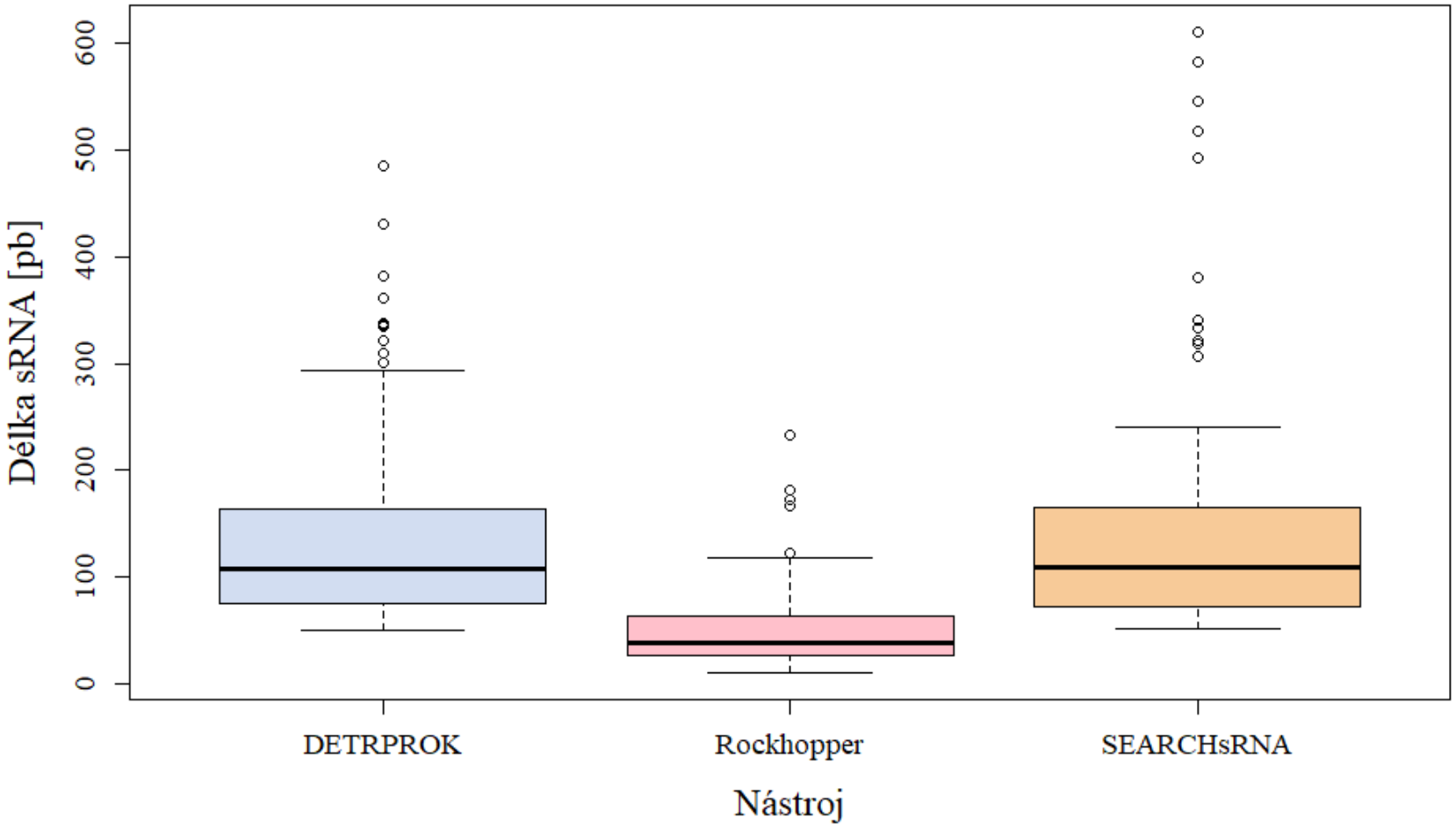
18	391318	391461	144	+	*trans-*	43	1146125	1146193	69	+	*cis-*	68	1568862	1569167	306	+	*trans-*
19	391895	392004	110	+	*trans-*	44	1210171	1210270	100	-	*trans-*	69	1576550	1576624	75	-	*cis-*
20	420822	420887	66	+	*trans-*	45	1228252	1228371	120	-	*trans-*	70	1580658	1580775	118	-	*trans-*
21	436667	436803	137	+	*trans-*	46	1236763	1236849	87	+	*trans-*	71	1608339	1608446	108	+	*trans-*
22	484721	484856	136	+	*trans-*	47	1236973	1237100	128	-	*trans-*	72	1615544	1615636	93	-	*trans-*
23	638715	638963	249	+	*trans-*	48	1244724	1244935	212	+	*cis-*	73	1669764	1669886	123	-	*trans-*
24	669226	669367	142	+	*trans-*	49	1273217	1273416	200	+	*trans-*	74	1672431	1672517	87	-	*trans-*
25	687404	687498	95	+	*trans-*	50	1277664	1277739	76	-	*trans-*						

B. Grafy

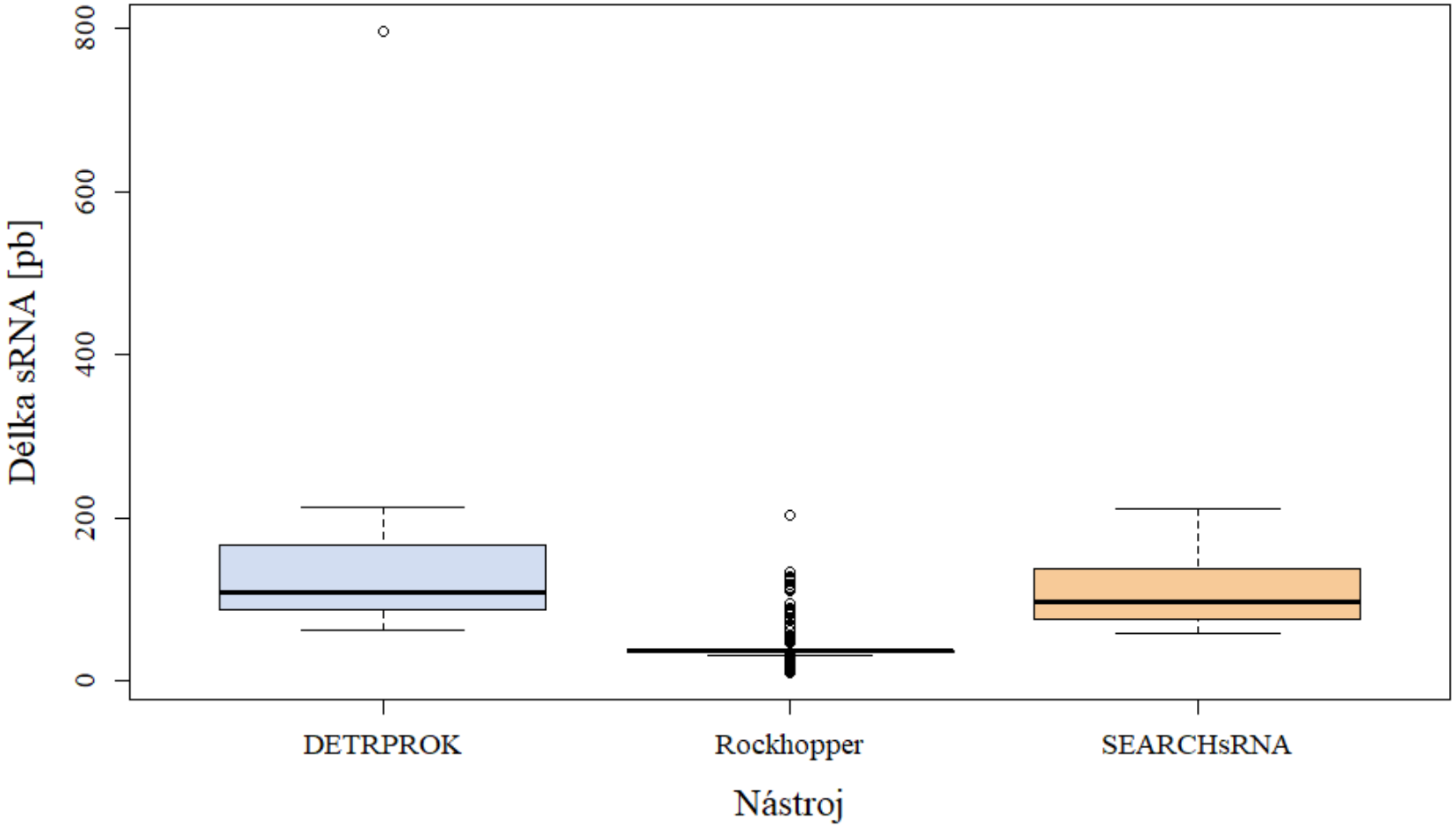
B.1 Box-plot pro délky *cis*-sRNA u chromozomu NC_011753.2



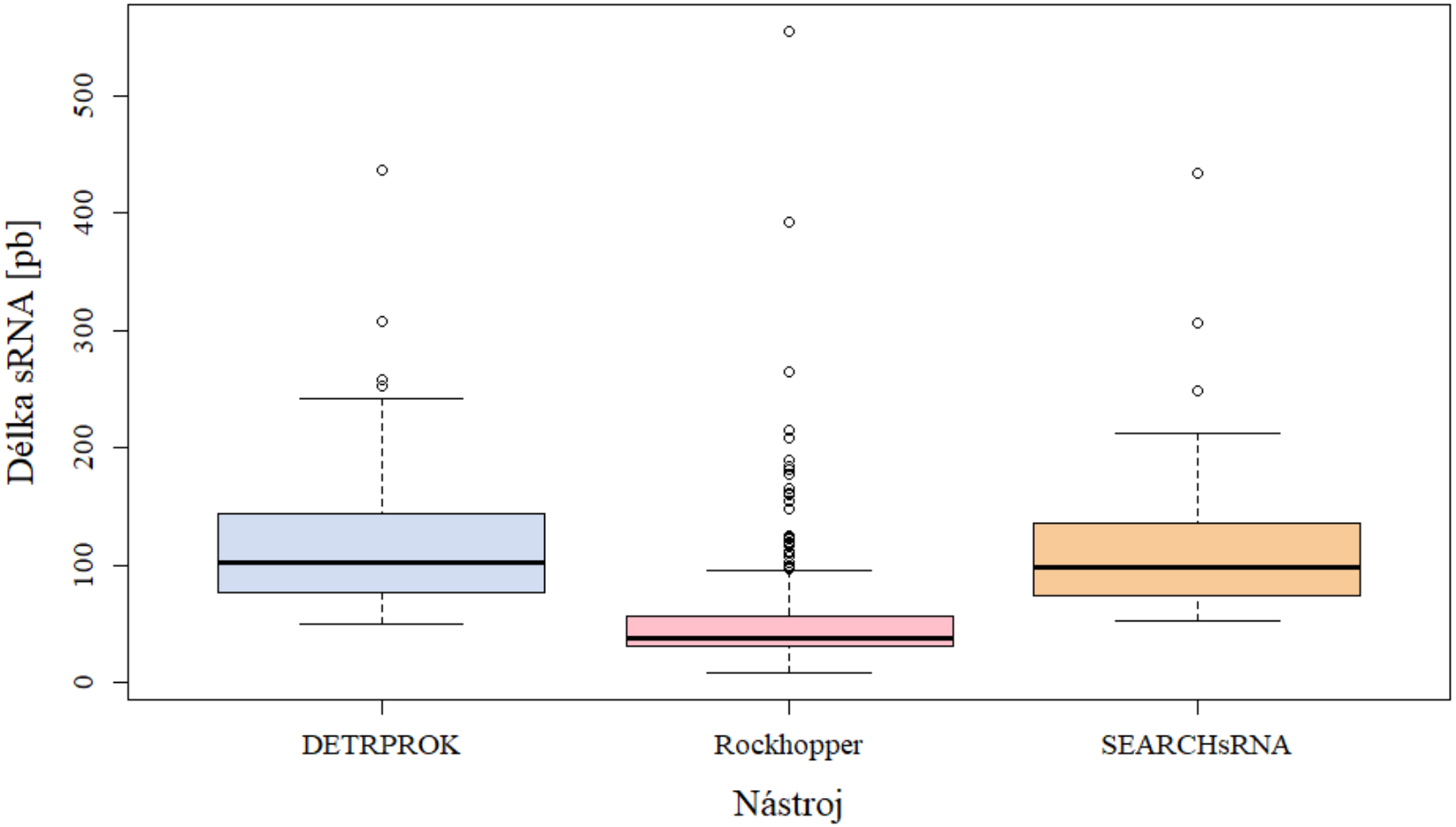
B.2 Box-plot pro délky *trans*-sRNA u chromozomu NC_011753.2



B.3 Box-plot pro délky *cis*-sRNA u chromozomu NC_011744.2



B.4 Box-plot pro délky *trans*-sRNA u chromozomu NC_011744.2



C. Soupis elektronických příloh

- *example.R* – skript obsahující příklad pro úspěšné spuštění nástroje SEARCHsRNA.
- *functions.R* – skript obsahující všechny funkce nástroje SEARCHsRNA.
- *installation_required.R* – skript obsahující instalační balíčky potřebné pro spuštění nástroje SEARCHsRNA.
- *NC_011753_sRNA.CSV* – výstupní CSV soubor nástroje SEARCHsRNA pro chromozom NC_011753.2 obsahující informace o detekovaných sRNA.
- *NC_011753_positive_sRNA.TXT* – výstupní TXT soubor nástroje SEARCHsRNA pro chromozom NC_011753.2 obsahující informace o pozicím sRNA na pozitivním vlákně.
- *NC_011753_negative_sRNA.TXT* – výstupní TXT soubor nástroje SEARCHsRNA pro chromozom NC_011753.2 obsahující informace o pozicím sRNA na negativním vlákně.
- *NC_011744_sRNA.CSV* – výstupní CSV soubor nástroje SEARCHsRNA pro chromozom NC_011744.2 obsahující informace o detekovaných sRNA.
- *NC_011744_positive_sRNA.TXT* – výstupní TXT soubor nástroje SEARCHsRNA pro chromozom NC_011744.2 obsahující informace o pozicím sRNA na pozitivním vlákně.
- *NC_011744_negative_sRNA.TXT* – výstupní TXT soubor nástroje SEARCHsRNA pro chromozom NC_011744.2 obsahující informace o pozicím sRNA na negativním vlákně.
- *NC_011753_sRNAs.GFF* – výstupní GFF soubor nástroje DETR'PROK pro chromozom NC_011753.2 obsahující nové anotace detekovaných *trans*-sRNA.
- *NC_011753_asRNAs.GFF* – výstupní GFF soubor nástroje DETR'PROK pro chromozom NC_011753.2 obsahující nové anotace detekovaných *cis*-sRNA.
- *NC_011744_sRNAs.GFF* – výstupní GFF soubor nástroje DETR'PROK pro chromozom NC_011744.2 obsahující nové anotace detekovaných *trans*-sRNA.
- *NC_011744_asRNAs.GFF* – výstupní GFF soubor nástroje DETR'PROK pro chromozom NC_011744.2 obsahující nové anotace detekovaných *cis*-sRNA.
- *NC_011753_ncRNAs* – výstupní WIG soubor nástroje Rockhopper pro chromozom NC_011753.2 obsahující pozice výskytů detekovaných sRNA pro pozitivní i negativní vlákno.
- *NC_011744_ncRNAs* – výstupní WIG soubor nástroje Rockhopper pro chromozom NC_011744.2 obsahující pozice výskytů detekovaných sRNA pro pozitivní i negativní vlákno.