Univerzita Palackého v Olomouci

Přírodovědecká fakulta

Katedra biotechnologií

Analýza transkriptomu v průběhu vývoje embrya *Sorghum purpureosericeum*

Transcriptome analysis of developing embryo of *Sorghum purpureosericeum*

DIPLOMOVÁ PRÁCE

Bc. Tereza Bojdová

Studijní program: Biotechnologie a genové inženýrství

Forma studia: Prezenční

Olomouc 2023

Vedoucí práce: Mgr. Jan Bartoš, Ph.D.

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně pod vedením školitele Mgr. Jana Bartoše, Ph.D., a uvedla v ní všechny použité zdroje a literaturu.

V Olomouci dne ……

Poděkování

Ráda bych poděkovala vedoucímu mé práce, Mgr. Janu Bartošovi, Ph.D., za odborné vedení. Jeho vstřícný a trpělivý přístup mi umožnil zdokonalit se v bioinformatice, kterou jsem si zamilovala a chtěla bych se jí dále věnovat. Neuvěřitelná míra jeho vědomostí je pro mě velkým přínosem, motivují mě se stále zlepšovat a učit se novým věcem v odvětví, které mě tak pohltilo.

Významné poděkování bych chtěla věnovat také Mgr. Miroslavě Karafiátové, Ph.D., která mi poskytuje nekonečné nadšení a inspiraci pro práci ve vědě. Je mi ctí, že mám příležitost se učit od tak talentované a zkušené vědkyně. Velmi si vážím času a důvěry, kterou mi v této práci věnovala, a která přispěla i k mému osobnímu rozvoji.

Dále děkuji všem zaměstnancům Centra strukturní a funkční genomiky rostlin Ústavu experimentální botaniky AV ČR v Olomouci, kteří vždy ochotně pomohli a poradili.

Tato práce byla vypracována za podpory projektu číslo 22-02108S financovaného Grantovou agenturou ČR. Výpočetní zdroje byly poskytnuty projektem e-INFRA CZ (ID:90140) podporovaným Ministerstvem školství, mládeže a tělovýchovy České republiky.

Bibliografická identifikace

Jméno a příjmení autora:	Bc. Tereza Bojdová
Název:	Analýza transkriptomu v průběhu vývoje embrya *Sorghum purpureosericeum*
Typ práce:	diplomová
Pracoviště:	Ústav experimentální botaniky AV ČR, v.v.i.
Vedoucí práce:	Mgr. Jan Bartoš, Ph.D.
Rok obhajoby práce:	2023
Klíčová slova:	*Sorghum purpureosericeum*, B chromozómy, eliminace B chromozómů, transkriptomika
Počet stran:	75
Jazyk:	čeština

Souhrn

Divoký druh čiroku *Sorghum purpureosericeum* může nad rámec své chromozomální sady obsahovat také nadbytečné B chromozómy, což jsou postradatelné elementy, které svým hostitelským organismům zpravidla neposkytují žádnou výhodu. B chromozómy *S. purpureosericeum* nejsou stabilní ve všech pletivech a během brzkého embryonálního vývoje dochází k jejich eliminaci ze všech pletiv kromě nediferenciovaných meristémů nadzemních orgánů.

Cílem této práce bylo vytvoření referenční genomové sekvence *S. purpureosericeum* a transkriptomická analýza RNA-Seq dat embryí v různém stádiu vývoje za účelem selekce kandidátních genů, které jsou odlišně exprimovány a mohly by být zodpovědné za eliminaci B chromozómů.

Reference byla získána pomocí nástroje SMARTdenovo a následně přečištěna programy Medaka a NextPolish. Výsledná sekvence má délku 2,82 Gb a kompletnost činila 98,7 % na základě BUSCO skóre. Pro transkriptomický experiment byla izolována B+ i B0 embrya a endospermy čtyř vývojových kategorií. Pro ověření B-statusu embryí byla použita DNA

izolovaná z endospermů a z vybraných embryí byla následně izolována RNA. Po NGS sekvenování RNA embryí byla provedena bioinformatická analýza RNA-seq dat.

Výsledky ukazují na možnou roli B chromozomů v regulaci genové exprese během vývoje zejména v nejmladších vzorcích embryí. Diferenciální analýza těchto mladých embryí odhalila 424 genů se zvýšenou expresí. Některé z těchto genů jsou pravděpodobně umístěny na B chromozómu a mohou představovat kandidáty pro soubor genů odpovědných za eliminaci B chromozómu z pletiv *S. purpureosericeum*.

Bibliographical identification

Author´s first name and surname:	Bc. Tereza Bojdová
Title:	Transcriptome analysis of developing embryo of *Sorghum purpureosericeum*
Type of thesis:	diploma
Department:	Institute of Experimental Botany of the Czech Academy of Sciences, v.v.i.
Supervisor:	Mgr. Jan Bartoš, Ph.D.
The Year of Presentation:	2023
Keywords:	*Sorghum purpureosericeum*, B chromosomes, B chromosome elimination, transkriptomics
Number of pages:	75
Language:	Czech

Summary

The wild species *Sorghum purpureosericeum* can contain supernumerary B chromosomes in addition to its normal chromosomal set. B chromosomes are dispensable elements that typically provide no advantage to their host organisms. B chromosomes of *S. purpureosericeum* are not stable and are eliminated from all tissues except apical meristem during early embryonic development.

The aim of this study was to create a reference genome sequence of *S. purpureosericeum* and perform transcriptomic analysis of RNA-Seq data from embryos at different stages of development in order to identify candidate genes that are differentially expressed and could be responsible for B chromosome elimination.

The reference sequence was obtained using the SMARTdenovo tool and subsequently polished using Medaka and NextPolish tools. The resulting sequence has a length of 2.82 Gb and completeness of 98.7 % BUSCO score. Both B+ and B0 embryos and endosperms of four developmental categories were isolated for the transcriptomic experiment. DNA isolated

from endosperms was used to verify the B-status of embryos, and RNA was subsequently isolated from selected embryos. After NGS sequencing, bioinformatics analysis of the RNA-seq data was performed.

The results suggest a possible role of the B chromosomes in the regulation of gene expression during development, particularly in the youngest embryonic category. The differential analysis of these young embryos revealed 424 genes with increased expression. Some of these genes are most likely located on the B chromosome and may represent candidates for the set of genes responsible for the elimination of the B chromosome from *S. purpureosericeum* tissues.

Obsah

1 Úvod

Rostoucí globální populace spolu se změnou klimatu vyvíjí na světovou potravinovou bezpečnost čím dál větší tlak. Do roku 2050 dosáhne světová populace 9,7 miliardy a bude potřeba produkovat o 70 % více potravin než nyní (Tripathi et al., 2019). Jako alternativa kukuřice se nabízí sladký čirok, další C4 plodina, která je oproti kukuřici schopna růst i v suchých oblastech, a proto je čirok vnímán jako potravina budoucnosti. Ačkoliv je nejvíce pěstovaný druh čiroku kultivovaný *Sorghum bicolor*, věnuje se nyní pozornost i divokým druhům čiroku, které obsahují geny rezistence na fytopatogeny či geny abiotické rezistence a mohou tak být využity ke křížení s kulturně pěstovaným druhem (Harlan, 1992; Kamala et al., 2002; Kamala et al., 2009).

Rod *Sorghum* může nad rámec své chromozomální sady obsahovat také nadpočetné B chromozómy. Tyto B chromozómy, které se vyskytují napříč říšemi rostlin, hub i živočichů, nepodléhají pravidlům Mendelovy dědičnosti a neposkytují svým hostitelům většinou žádnou výhodu. Výskyt B chromozómu u druhu *Sorghum purpureosericeum* je v rostlině nestabilní ve všech pletivech kromě květenství a k jeho selektivní eliminaci pravděpodobně dochází již v raném vývoji embrya.

Programovaná eliminace B chromozómů je velmi zajímavý biologický fenomén, který byl v rostlinách již dříve zaznamenán. U některých druhů, jako je *Haploppapus gracilis* (Ostergren & Frost, 1962), *Poa timoleontis* (Nygren, 1957) a *Aegilops speltoides* (Mochizuki, 1957; Mendelson & Zohary, 1972) je B chromozóm eliminován z kořenů.

Studium eliminace B chromozómů je sledováno také za účelem pochopení evoluce a adaptace B chromozómů v hostitelských organismech. Aplikace tohoto výzkumu by mohla inovovat studium lidských chromozomálních aberací, které by eventuelně mohlo vést k novým přístupům léčby. Výzkum eliminace B chromozómů má také potenciál k využití v agrikultuře a rostlinné biotechnologii za účelem šlechtění.

2 Cíle práce

Cílem předkládané práce byla analýza transkriptomu embryí čiroku *Sorghum purpureosericeum* v průběhu vývoje. Teoretická část se zabývá vypracováním literární rešerše na dané téma. Experimentální část práce zahrnuje:

- Sestavení kontigů genomové sekvence.
- RNA-Seq analýza v průběhu vývoje embryí.
- Identifikace kandidátních genů zodpovědné za eliminaci B chromozómu.

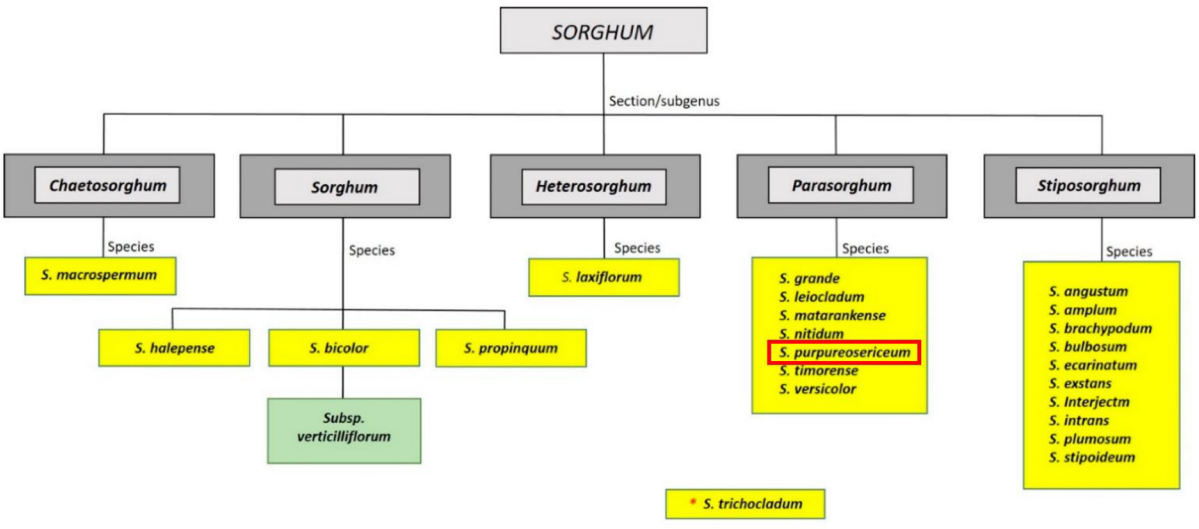
3 Literární přehled

3.1 Čirok *(Sorghum)*

Čirok, zejména pak čirok dvoubarevný (*Sorghum bicolor L.*), je základní potravinou pro miliony lidí žijících v polosuchých oblastech Afriky a Asie. Jde o krátkodenní C4 rostlinu s vysokou odolností růstu v teplých a suchých podmínkách. Vzhledem ke globální výzvě stanovené Světovým summitem o potravinové bezpečnosti navýšit produkci potravin o 70 % (FAO, 2009) je čirok za současných změn klimatu jednou z nadějných potravin budoucnosti. Kromě zdroje obživy je čirok pěstován i za účelem krmiva a aditiva do biopaliv, a to zejména ve Spojených státech amerických a Austrálii (Reddy & Yang, 2005).

3.1.1 Taxonomie a geografické rozšíření čiroku

Rod *Sorghum* je klasifikován do čeledi lipnicovitých (*Poacae*), tribu *Andropogoneae*. Na základě morfologických znaků jako je barva zrn či plev a perzistence květní stopky, byl rod *Sorghum* původně dělen na dva podrody – *Eu-sorghum* a *Parasorghum* (Snowden, 1936). Dnes se k těmto dvěma podrodům řadí navíc podrody *Chaetosorghum*, *Heterosorghum* a *Stiposorghum* (Lazarides et al., 1991) (Obr. 1). Divoký druh čiroku *Sorghum purpureosericeum* studovaný v této práci je řazen do podrodu *Parasorhum* (Obr. 1).



Obrázek 1: Taxonomická klasifikace rodu *Sorghum*. Studovaný druh *S. purpureosericeum* je označen červeně, hvězdičkou je označen ještě nezařazený druh. Převzato z Bednářová et al., 2021.

Domestikace čiroku se datuje 4500–8 000 let zpět v oblastech Afriky a Indie (Mann et al., 1983; Damania, 2002) a v 17. století byl rozšířen do USA (Rajendra Kumar & Patil, 2015).

Obecně platí, že v rozvojových zemích je čirok pěstován tam, kde není možné pěstovat z důvodu klimatických podmínek jiné plodiny. Takovými oblastmi jsou zejména suché tropy, avšak divoké druhy čiroků vykazují značnou adaptabilitu na rozmanitější podmínky klimatu (Harlan, 1992; Bramel-Cox, 1988).

Druhy čiroku z podrodu *Eu-sorghum* jsou považovány za „pravý čirok“, tedy čirok hospodářsky využívaný. Zahrnuje druhy: *S. bicolor, S. propinquum, S. halepense* a hybridní druh s názvem *Sorghum ×almum* Parodi (*USDA*, 2022). Druhy čiroku z podrodu *Eu-sorghum* lze najít po celé Africe, jihovýchodní Asii a jižní Eurasii (Ananda et al., 2020).

Podrod *Parasorghum* zahrnuje 5 druhů, z nichž 3 se nachází výhradně v Austrálii. Do tohoto podrodu spadá i *Sorghum purpureosericeum*, který lze nalézt v Indii, východní a západní tropické Africe a na jihu Sahary (Ananda et al., 2020). Poslední zástupce, *S. versicolor,* je opět africký druh.

Čiroky podrodů *Chaetosroghum*, *Heterosorghum* a *Stiposorghum* obvykle rostou v Austrálii či Nové Guinei (Ananda et al., 2020).

3.1.2 Využití čiroku

Využití čiroku je soustředěné zejména na druh *Sorghum bicolor*. V současnosti jsou největšími světovými producenty čiroku USA, Nigérie, Indie, Mexiko a Sudán. K roku 2021 bylo celosvětově vyprodukováno 62 milionů tun čiroku na téměř 41 milionech hektarech (FAO, 2009).

Čirok má vysokou toleranci k různým abiotickým stresům a vysokou efektivitu využití vody. Z toho důvodu se za účelem potravy pěstuje v suchých oblastech a nahrazuje plodiny v místech, kde se zhoršuje dostupnost vody (Xie & Xu, 2019). V rozvojových státech Afriky a Asie je tak čirok základním zdrojem obživy. Chemické složení a nutriční hodnota čirokového semena jsou podobné rýži, kukuřici a pšenici. Obsahuje velký podíl polysacharidů, bílkovin, lipidů, minerálů a karotenoidů (de Morais Cardoso et al., 2017). V čirokovém endospermu se také nachází vitamíny – B komplex i vitamíny rozpustné v tucích (Slavin, 2004). *Sorghum* je mimo jiné zdrojem bioaktivních látek většinou fenolického charakteru. Čirok dále stejně jako kukuřice neobsahuje lepek, a je tedy vhodnou obilninou pro celiaky (Pontieri et al., 2013).

Čirok má dále výborné vlastnosti pícniny. Je dobře stravitelný a obsahuje důležité živiny pro krmení hospodářských zvířat. Biomasa (stonky, listy i semena) se pak také používá k výrobě chemikálií z lignocelulózy a vláken na biomateriály (Silva et al., 2022). Biomasa čiroku je také udržitelný zdroj pro výrobu biopaliv.

Sladké kultivary čiroku byly vyšlechtěny jako alternativní zdroj cukru v oblastech nevhodných pro produkci cukrové třtiny, k výrobě sirupu a alkoholové destilaci (Silva et al., 2022). Divoké druhy čiroků mohou být vhodným zdrojem genů pro vylepšení plodin, jelikož jsou více geneticky rozmanité než kulturní čiroky (Meilleur & Hodgkin, 2004) (Barnaud et al., 2009). Některé z divokých čiroků vykazují jistou míru adaptability na biotické či abiotické stresy. Například divoký čirok *Sorghum arundinaceum* je adaptovaný na vlhké podmínky (Harlan, 1992), což je vlastnost nevyskytující se u pěstovaných čiroků, a divoký *Sorghum virgatum* je schopen růstu v suchých podmínkách s vyššími teplotami (Bramel-Cox, 1988). Divoké čiroky vykazují odolnost vůči různým parazitům. Jedním z nich je moucha *Atherigona soccata*, která v Asii a Africe působí průměrně 5% ztráty úrody, nebo paraziti z taxonu Oomycetes (Kamala et al., 2002; Kamala et al., 2009).

3.2 B chromozómy

Genetická informace jedinců je organizována do konstantního počtu chromozómů specifického pro daný druh (A chromozómy). Některé organismy však mohou obsahovat nadbytečné chromozomy, které se nazývají B chromozómy a vyskytují se v karyotypu nad rámec běžné chromozomální sady. Tyto chromozómy obvykle nepřinášejí nositeli žádnou výhodu a jejich přítomnost může naopak snížit vitalitu a fertilitu hostitelského organismu. Jsou proto často považovány za "sobecké" prvky, které se i přes svou neužitečnost udržují v populaci. B chromozómy se také vyznačují tím, že se jejich přenos do potomstva neřídí Mendelovými zákony dědičnosti a jejich výskyt se liší jak mezi populacemi, tak i mezi jejími jedinci.

3.2.1 Výskyt

Výskyt B chromozómu byl poprvé zaznamenán roku 1907 v rodu *Metapodius* (Wilson, 1907). V rostlinách byly identifikovány až ve dvacátých letech, a to v žitu (Gotoh, 1924) a v kukuřici (Kuwada, 1925). Tyto chromozómy byly označeny za „nadbytečné“ (z anglického výrazu „supernumerary“) v roce 1927 (Longley, 1927), a termín „B chromozóm“, který je odlišil od běžných „A chromozómů“, byl zaveden až v roce 1928 (Randolph, 1928). Tyto dva termíny jsou dodnes běžně používány.

Od objevení B chromozómů před více než sto lety byly tyto nadpočetné elementy popsány ve všech eukaryotických říších. Databáze B-chrom uvádí 2 828 organismů s B chromozómy, z toho rostliny představují 2 087 druhů, živočichové 736 druhů a houby 14 druhů (D'Ambrosio et al., 2017). Dle databáze CCDB (The Chromosome Counts Database), která

poskytuje chromozómové počty pro 77 958 druhů rostlin (Rice et al., 2015), lze tedy odvodit, že B chromozómy se vyskytují u 2–3 % druhů rostlin. Je však velmi pravděpodobné, že frekvence výskytu B chromozómů je podstatně větší. Výzkum druhů totiž není rovnoměrně rozdělen mezi veškeré organismy a komplikuje ho také omezená přítomnost B chromozómů v tkáních či pletivech, které se běžně nevyužívají ke stanovení karyotypu.

Frekvence výskytu B chromozómů je nezávislá na ploidii genomů (Palestis et al., 2004; Trivers et al., 2004). V živočišné říši je obecně výskyt B chromozómů častější u savců s akrocentrickými A chromozómy (Palestis et al., 2004). U rostlin jsou B chromozómy zastoupeny rostlinnými druhy s většími genomy (Trivers et al., 2004), které lépe tolerují nadbytečný genetický materiál (Levin et al., 2005). Literatura také popisuje korelaci mezi zvýšeným výskytem B chromozómů a znečištěním v životním prostředí či stresovými klimatickými podmínkami (Douglas & Birchler, 2017).

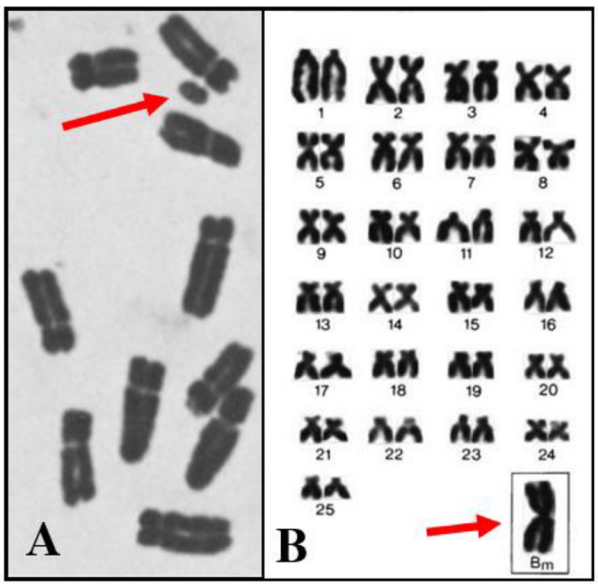
3.2.2 Diverzita, morfologie a stabilita

Vzhledem ke svému rozmanitému výskytu napříč eukaryotickými organismy se mezi B chromozómy vyskytuje velká míra diverzity. Karyotypová diverzita B chromozómů zahrnuje rozdíly v počtu kopií, velikosti či poloze centromery.

B chromozómy se obvykle v genomu nachází v nízkém počtu kopií, ale některé organismy jeví překvapivě velkou toleranci k jejich přítomnosti. Například u kukuřice seté bylo pozorováno až 34 kopií B chromozómu (Jones, & Rees, 1982). Sukulent *Pachyphytum fittkaui* je schopen ve svém genomu nést až 50 kopií a jedná se tak o doposud nejvyšší zaznamenaný počet kopií B chromozómu v rostlinách (Uhl. & Moran, 1973). Tyto vysoké počty jsou však spíše výjimkou a vypovídají o vysoké frekvenci přenosu B chromozómu či toleranci organismu k nesení nadbytečné genetické informace (Camacho, 2005).

Variabilita se týká i velikosti B chromozómů napříč druhy. U většiny druhů rostlin, pro které jsou dostupné informace, mají B chromozómy délku jedné až tří čtvrtin průměrné velikosti A chromozómů. Existuje pouze pár druhů, u kterých je v mitóze B chromozóm stejné velikosti jako A chromozóm. Takovým případem je *Clarkia elegans* (Lewis, 1951) nebo *Sorghum nitidum* (Raman & Krishnaswami, 1960). U některých druhů jsou B chromozómy podstatně menší než nejmenší A chromozómy. Takový příklad „mikro“ B chromozómu je u rostlin znám například u *Hypochoeris maculata* (Parker et al., 1982) (Obr. 2A). Zatím není známo mnoho rostlinných druhů, u kterého by B chromozómy přesahovaly velikost největšího z A chromozomů tak, jako je tomu například u kaprovité ryby *Alburnus alburnus* (Ziegler et al., 2003) (Obr. 2B). Diverzitu ve velikosti B chromozómů lze

běžně nalézt i v rámci jednoho druhu, kdy může existovat více typů B chromozómů i u jednoho jedince. (Henriques-Gil et al., 1984; López-León et al., 1993)



Obrázek 2: Diverzita velikosti B chromozómů. A – B chromozóm menší než A chromozómy *Hypochoeris maculata*, B – B chromozóm větší než A chromozómy ryby *Alburnus alburnus*. Modifikováno z Parker et al., 1982 a Schmid et al., 2006.

B chromozómy jsou nejčastěji metacentrické nebo akrocentrické, a to se opět může lišit i v rámci jednoho druhu. Je však zajímavé, že morfologie B chromozómů odráží morfologii A chromozómů (Henriques-Gil et al., 1984; López-León et al., 1993).

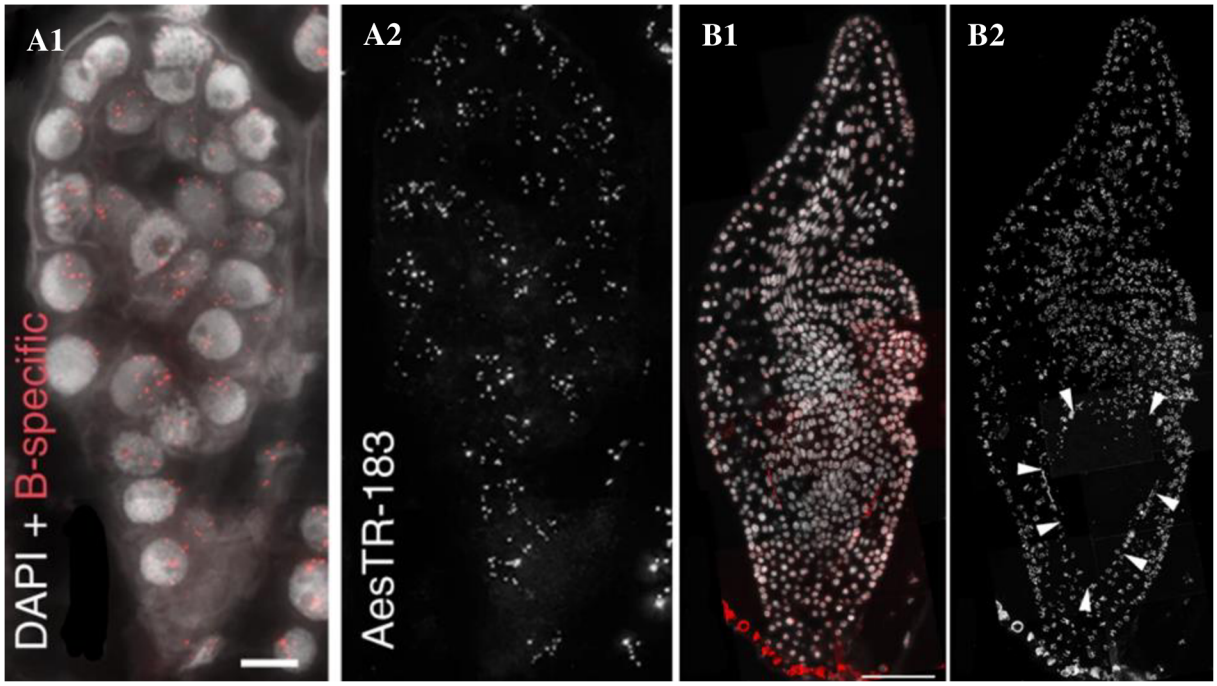
B chromozómy se vyznačují vysokým obsahem heterochromatinu s vysokou úrovní kondenzace po většinu buněčného cyklu. Tento znak vyplývá z vysokého obsahu repetitivní DNA, zejména satelitní (satDNA), ribozomální (rDNA) a DNA mobilních elementů (mobDNA) (Camacho, 2005). Organelová DNA, mobilní elementy a jiné repetitivní sekvence se zde mohou hromadit, protože B chromozómy nejsou pod selekčním tlakem (Martis et al., 2012; Lamb et al., 2007; Kour et al., 2013).

Vznikem B chromozómů se zabývá více hypotéz, z nichž nejvíce pravděpodobný je jejich původ z A chromozómového komplementu, a to buď intraspecificky (ze stejného druhu současného hostitele) či interspecificky (mezidruhově). Intraspecificky vznikl B chromozóm kukuřice, žita a také některých druhů ryb (Lamb et al., 2005; Martis et al., 2012; Serrano et al., 2017). Naopak na interspecifický původ B chromozómu ukazuje případ vosy *Nasonia vitripennis*, kde molekulární fylogeneze retrotranspozonu NATE odhalila jeho přítomnost u cizího druhu vosy rodu *Trichomalopsis* (McAllister & Werren, 1997) a interspecifickým křížením byl potvrzen vznik B chromozómů *de novo* (Perfectti & Werren, 2001). Další hypotézou původu B chromozómů je vznik z pohlavních chromozómů, což bylo potvrzeno

například u saranče*te Eyprepocnemis plorans* (López-León et al., 1994) a u ryb rodu *Characidium* (Pansonato-Alves et al., 2014).

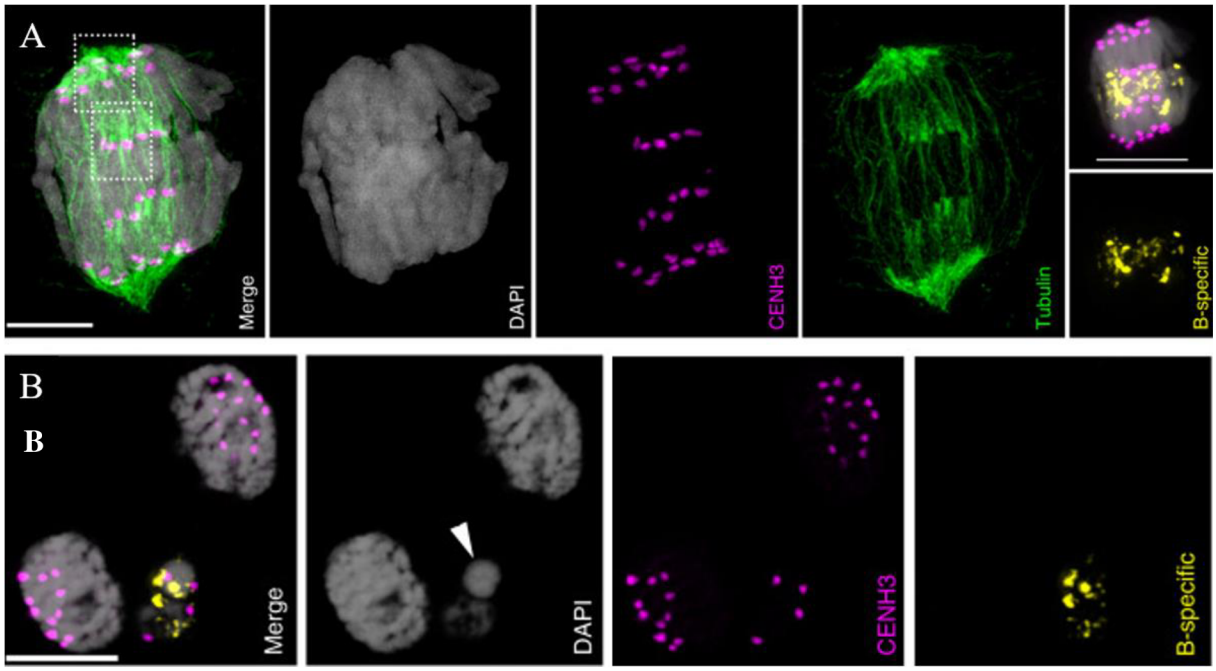
B chromozómy se vyskytují stabilně ve všech rostlinných pletivech během vývoje rostliny a jejího růstu, nicméně následkem abnormálního chování během buněčného dělení existují výjimky z tohoto pravidla. Somatická variace ve frekvenci počtu kopií se vyskytuje například v *Crepis capillaris,* kde se počet B chromozómů v nadzemních pletivech liší od počtu kopií B chromozómů v kořenech (Rutishauser & Röthlisberger, 1966). Úplná absence B chromozómů z kořenů byla popsána v *Haploppapus gracilis* (Ostergren & Frost, 1962), *Poa timoleontis* (Nygren, 1957) či *Aegilops speltoides* (Mochizuki, 1957; Mendelson & Zohary, 1972).

Mechanismus této pletivově specifické distribuce B chromozómů byl nedávno objasněn právě u druhu *Aegilops speltoides*. Již dříve bylo známo, že tento druh obsahuje B chromozómy ve všech pletivech kromě kořenů dospělých rostlin (Jones & Rees, 1982). Nicméně až studie z roku 2020 odhalila, že k eliminaci B chromozómů dochází již pár dní po opylení v brzkém stádiu vývoje embrya (Ruban et al., 2020). V pětidenním embryu (5 DAP, z anglického „days after pollination) *A. speltoides* lze pozorovat B chromozóm ve všech buňkách embrya (Obr. 3A), zatímco v patnáctidenním embryu (15 DAP) je zřetelná oblast embryonálního kořene, z něhož je B chromozóm zcela eliminován (Obr. 3B).



Obrázek 3: Fluorescenční *in situ* lokalizace B specifické sondy na řezech embryí *A. speltoides*. B-specifická sonda AesTR-183 je na sloučených obrázcích (A1, B1) znázorněna červeně . Na obrázcích A2, B2 je zdůrazněna pouze vrstva s B-specifickým signálem v bílém zobrazení. A – 5 DAP embryo s B chromozómem ve všech buňkách. Měřítko 10 µm. B – 15 DAP embryo, šipky ukazují na region kořene bez přítomnosti B chromozómů. Měřítko 100 µm. Modifikováno z Ruban et al., 2020.

K eliminaci B chromozómů dochází v průběhu mitotického dělení jako následku segregačního selhání. V somatických kořenových buňkách lze pozorovat v anafázi zpožďování B chromozómů při tažení chromozómů k opačným pólům (Obr. 4A). Opožděné chromozómy se nezačlení do nových jader a zůstávají v cytoplazmě. V telofázi se poté formuje mikrojádro, které obsahuje B chromozómy, a které je následně degradováno (Obr. 4B).



Obrázek 4: Eliminace B chromozómů *A. speltoides* cestou tvorby mikrojader. Připojení chromozómů k dělícímu vřeténku je vizualizováno imunologicky (α-tubulin zeleně) a CENH3 protein růžově). B-chromozóm lokalizován pomocí FISH s B-specifickou repeticí (žlutě). A – Zpožďující se B chromozómy v anafázi, měřítko 5 µm, B – Formace mikrojádra (naznačeno šipkou) v telofázi, měřítko 10 µm. Modifikováno z Ruban et al., 2020.

3.2.3 Akumulační mechanismy

Frekvence přenosu B chromozómů jsou na rozdíl od A chromozómů vyšší než 0.5 a neřídí se tedy Mendelovským zákonem rovnoměrné segregace (Jones, 1991; Houben, 2017). Tato výhoda přenosu je označována jako „akumulační mechanismus" nebo jednoduše „drive“ a umožnila sobeckým B chromozómům se navzdory jejich postradatelnosti udržet v populaci. Dle fáze životního cyklu, ve kterém se drive projevuje, jsou známé tři typy: pre-meiotický, meiotický a post-meiotický drive (Jones, 1991; Houben, 2017).

Pre-meiotický drive způsobuje zvýšení počtu kopií B chromozómů v buňkách zárodečné linie tak, že se B chromozómy pohybují preferenčně k pólu raného embrya, ze kterého mají vznikat buňky zárodečné linie v následujících děleních (Obr. 5A). Poprvé byl tento jev pozorován u kobylky *Calliptamus palaestinensis* (Nur, 1969) a později byl objeven také v rostlinách, například v již zmíněné *Crepis capillaris* (Rutishauser & Röthlisberger, 1966).

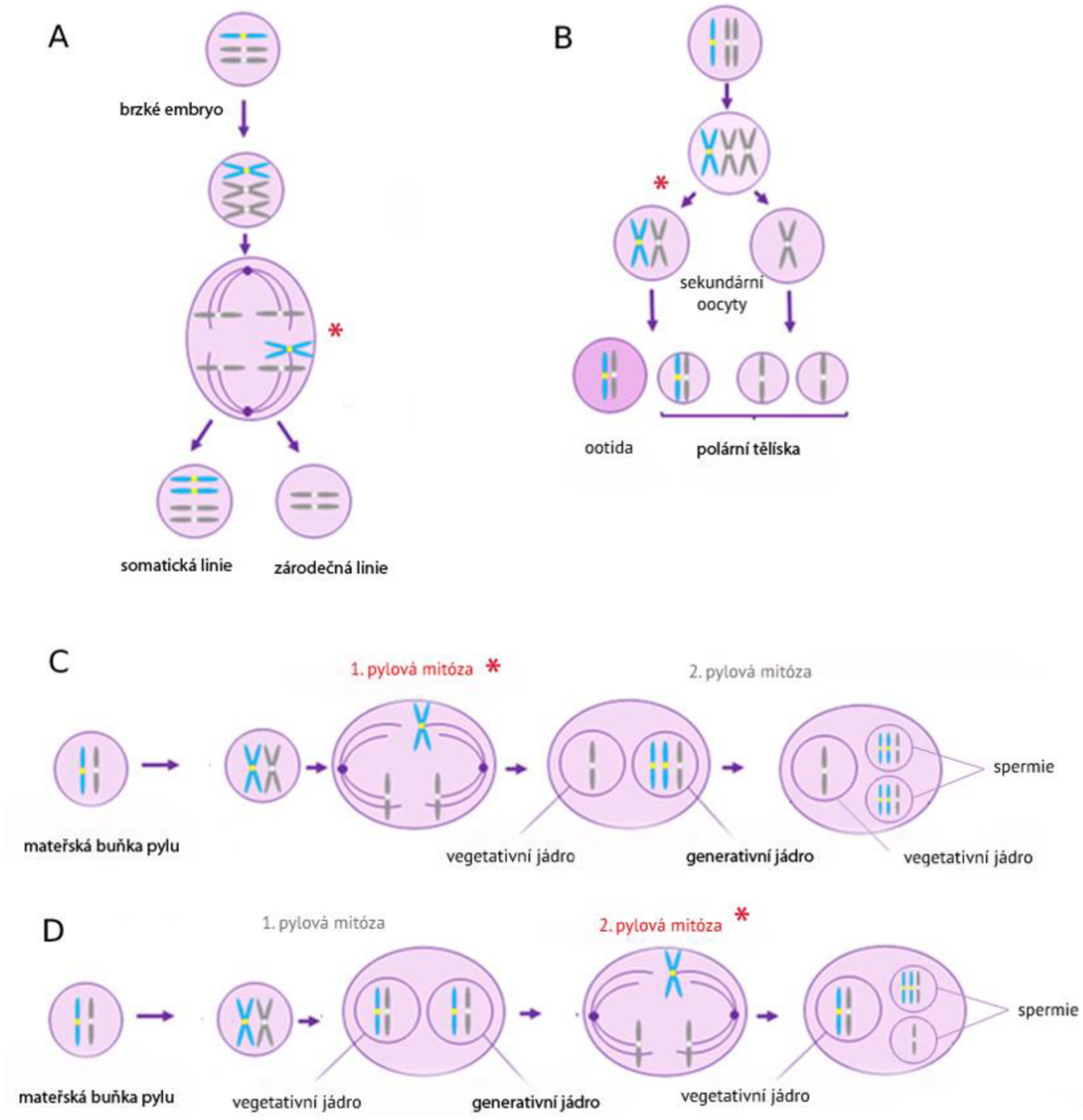
Meiotický drive je závislý na funkční asymetrii dělících vřetének. V běžném případě je v meióze jedna ze dvou buněk vzniklých u meiotických dělení přirozeně neživotaschopná (polární tělíska) a B chromozómy, které se dostanou do polárních tělísek tak přirozeně vymizí. Této 50% šanci na eliminaci jsou B chromozómy schopné uniknout preferenční

migrací k životaschopnému pólu – oocytu II v prvním dělení či ootidě v druhém dělení (Obr. 5B). Tento drive je typický hlavně pro živočichy jak v samičí (Hewitt, 1973), tak v samčí (Nur, 1962) meióze, byl však pozorován i u rostlin (Kayano, 1956).

Post-meiotická nondisjunkce zajišťuje především u rostlin preferenční přenos obou B chromatid do jednoho z dceřiných jader. Zralý pyl vzniká dvěma post-meiotickými děleními a nondisjunkce B chromozómů může nastat jak v prvním, tak ve druhém pylovém dělení.

Při nondisjunkce v prvním mitotickém dělení pylu nedochází k rovnoměrné segregaci B chromozómů do vegetativního a generativního jádra. B chromozóm je obsažen pouze v generativním jádře, a ve druhé pylové mitóze z něj vznikají dvě stejné spermatické buňky, které obě obsahující B chromozóm. Při dvojitém oplození splyne jedna spermatická buňka (n) s vaječnou buňkou (n) a vzniká tak zygota (2n), která se dále vyvíjí v embryo obsahující B chromozómy. Druhá spermatická buňka (n) splývá s centrálním jádrem (2n) a vzniká tak endosperm (3n) obsahující B chromozómy. Mechanismus přenosu B chromozómu v první pylové mitóze byl popsán například u žita a tento typ přenosu se uplatňuje i u čiroku *S. purpureosericeum* studovaného v této práci (Hasegawa, 1934) (Obr. 5C).

Při nondisjunkci v druhé pylové mitóze, která je dobře popsána u B chromozómů kukuřice, se vytvoří z generativního jádra dvě spermie, z nichž pouze jedna nese B chromozómy, a právě ta je schopna přednostně oplodnit vajíčko (Carlson, 1969) (Obr. 5D). Tento typ post-meiotického dělení vyústí ve vznik embrya obsahujícího B chromozómy a endospermu, který B chromozómy neobsahuje.



Obrázek 5: Akumulační mechanismy B chromozómu. A – pre-meiotický drive, B – meiotický drive, C – post-meiotický drive v 1. pylové mitóze, který dává za vznik 2 spermiím s B chromozómy, D – post-meiotický drive v 2. pylové mitóze, kdy vznikají dvě spermie, z nichž pouze jedna obsahuje B chromozóm. B chromozóm je znázorněn modře a nondisjunkce je označena hvězdičkou. Modifikováno z Johnson Pokorná & Reifová, 2021.

3.2.4 Efekt přítomnosti B chromozómů v hostiteli

B chromozómy obecně hostitelskému organismu navzdory přítomnosti mnoha genů neposkytují žádnou výhodu, naopak jsou pro jeho růst a vývoj zcela postradatelné. V ojedinělých případech můžou B chromozómy ovlivňovat fenotyp hostitele, a to zejména negativně. Buňky obsahující vysoké počty B chromozómů mají více DNA, která se musí

replikovat, což vede k prodloužení buněčného cyklu (Evans et al., 1972). Buněčné dělení vyžaduje velké množství energie a metabolitů, a tudíž může přítomnost nadbytečných B chromozómů mít negativní vliv na fenotyp, jako je tomu například u embryí *Myrmeleotettix maculatus*, kde B chromozómy způsobují zpomalení vývoje (Hewitt, 1973) .

Nízké počty kopií B chromozómů obvykle v rostlinách nemají významný fenotypový efekt. Za přítomnosti vyššího počtu kopií B chromozómů je však jejich vliv na fenotyp viditelný – dochází ke zhoršení vitality rostliny a ke snížení fertility (Jones, 1995). Typickým příkladem takového efektu je zakrslý růst kukuřice s B chromozómy (Staub, 1987). B chromozómy kukuřice také negativně ovlivňují fertilitu rostliny, což má za následek produkci defektního pylu a semen (Randolph, 1941). Ve vzácných případech může B chromozóm obsahovat geny poskytující hostiteli výhodu. Například patogenní houba *Nectria haematococca* nese B chromozóm s geny zodpovědné za degradaci antimikrobiální látky produkované hrachem setým (Miao et al., 1991). B chromozóm tím umožňuje houbě úspěšně infikovat rostlinu, protože houba je schopná se vyhnout obranné reakci hrachu.

Ačkoliv se dříve považovaly B chromozómy za transkripčně neaktivní elementy, některé studie nedávno odhalily, že B chromozómy jsou v jisté míře transkripčně aktivní a obsahují protein-kódující sekvence. Většina genových sekvencí B chromozómů však není funkční v důsledku pseudogenizace (Banaei-Moghaddam et al., 2013). Některé geny lokalizované na B chromozómech jsou schopné ovlivnit samičí determinaci pohlaví (Yoshida et al., 2011), nebo kódují protein kinázy podílející se na kontrole buněčného cyklu např. lišky obecné a psíka mývalovitého (Makunin et al., 2018).

B chromozómy jsou také schopné ovlivnit expresi genů na A chromozómech, což bylo prokázáno například v kukuřici (Huang et al., 2016). B chromozómy také jistou mírou ovlivňují expresi na A chromozómech žita, a stejně jako u kukuřice jsou tyto transkripčně aktivní geny sekvenčně podobné svým homologům, což komplikuje výzkum diferenciální exprese (Banaei-Moghaddam et al., 2013).

3.2.5 Aplikace B chromozómů

Aplikovaný výzkum B chromozómů je nejvíce propracován u kukuřice, jejíž B chromozóm patří mezi nejlépe popsané. B chromozómy kukuřice byly využity pro mapování A genomu (Birchler, 1991), aspekty evoluce genomu či studium struktury centromery (Lamb et al., 2005).

Velká pozornost se také upíná na fenomén nerovnoměrné meiotické či mitotické segregace, kterou využívají B chromozómy ve svůj prospěch k přežití v populaci (viz kapitola

3.2.2). Poruchy v rozchodu chromozómů při buněčném dělení se vyskytují běžně i u lidí, kde mitotická nondisjunkce často souvisí s tumorogenezí a meiotická nondisjunkce způsobuje formaci aneuploidní zygoty, z níž vznikají embrya s vážnými vývojovými abnormalitami (Taylor, 1998). B chromozómy by tak mohly sloužit jako modelový systém pro plné pochopení příčin a mechanismů nondisjunkce lidských chromozómů, což by mohlo přinést nové způsoby léčby abnormalit u lidí.

Vnesení B chromozómu do cizích genomů má potenciál v sestrojení umělého chromozómu nesoucího transgeny (Jones et al., 2008). Umístění transgenů mimo hostitelský A genom je obecně výhodné z důvodu stability a vysoké klonovací kapacity. Většina plazmidů je schopna nést až 15 kb inzertu, kdežto umělé chromozómy 100–2000 kb inzertu (Bajpai, 2014).

3.2.6 B chromozómy rodu *Sorghum*

Ačkoliv jsou rostlinné B chromozómy biologicky zajímavým tématem, je jejich výzkum z velké části zaměřen na rostliny agronomického využití. Na rozdíl od široce pěstované kukuřice se rodu *Sorghum* nedostává velké pozornosti a o jeho B chromozómech není mnoho dostupných informací.

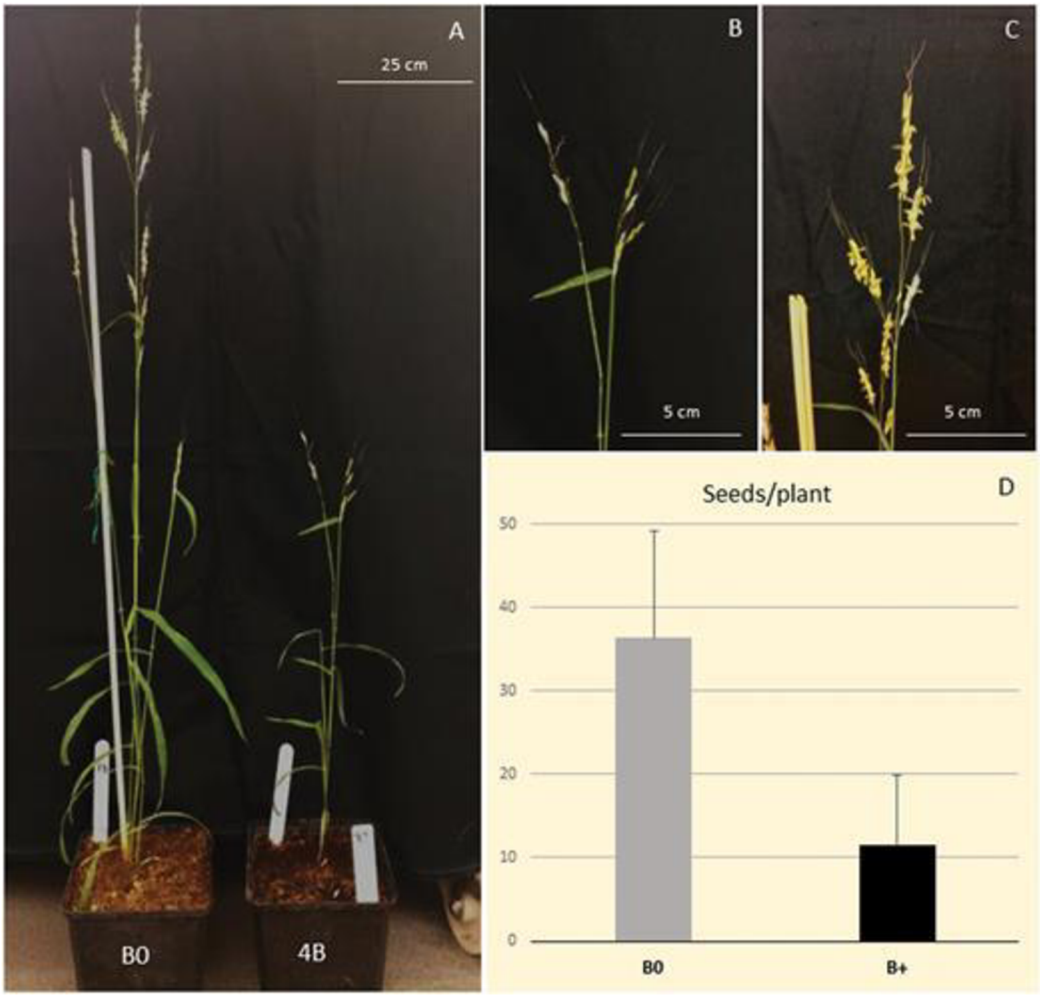
V rodu *Sorghum* byly B chromozómy popsány v pěti druzích divokých čiroků. Jako první byly zaznamenány v *S. verticilliflorum* (Huskins & Smith, 1934) a *S. purpureosericeum* (Janaki-Amma, 1939), později v *S. nitidum* (Raman & Krishnaswami, 1959) a *S. halepense* (Raman et al., 1964) a jako poslední v *S. stipoideum* (Wu, 1992).

3.2.6.1 B chromozómy *Sorghum purpureosericeum*

Čirok *Sorghum purpureosericeum* je jednoletá bylina, která dorůstá do výšky přibližně 1 metru. Čirok se řadí do skupiny často-cizosprašných rostlin. To znamená, že primárně je samosprašný, nicméně existuje významná míra cizosprašnosti. U divokých druhů čiroků může k cizosprašnosti docházet až s frekvencí 30 % (Rakshit & Bellundagi, 2019), což může komplikovat práci s potomky.

Genom tohoto divokého druhu čiroku je diploidní s 10 velkými chromozómy a nad rámec této sady A chromozómů může obsahovat B chromozómy. Již první studie u tohoto druhu čiroku popsaly přítomnost tří heterochromatických typů B chromozómu (Darlington et al., 1941). Nejvíce bylo zaznamenáno šest kopií B chromozómu v *S. purpureosericeum,* rostliny jeho přítomnost příliš netolerují (Janaki-Ammal, 1940). Vliv B chromozómů na fenotyp rostliny je lehce patrný už v 2B rostlinách, kdy jsou rostliny mnohdy menší a produkují až

o dvě třetiny méně semen než 0B rostliny (Obr. 6). V 3B a 4B rostlinách je tento efekt ještě výraznější (Karafiátová et al., 2021).

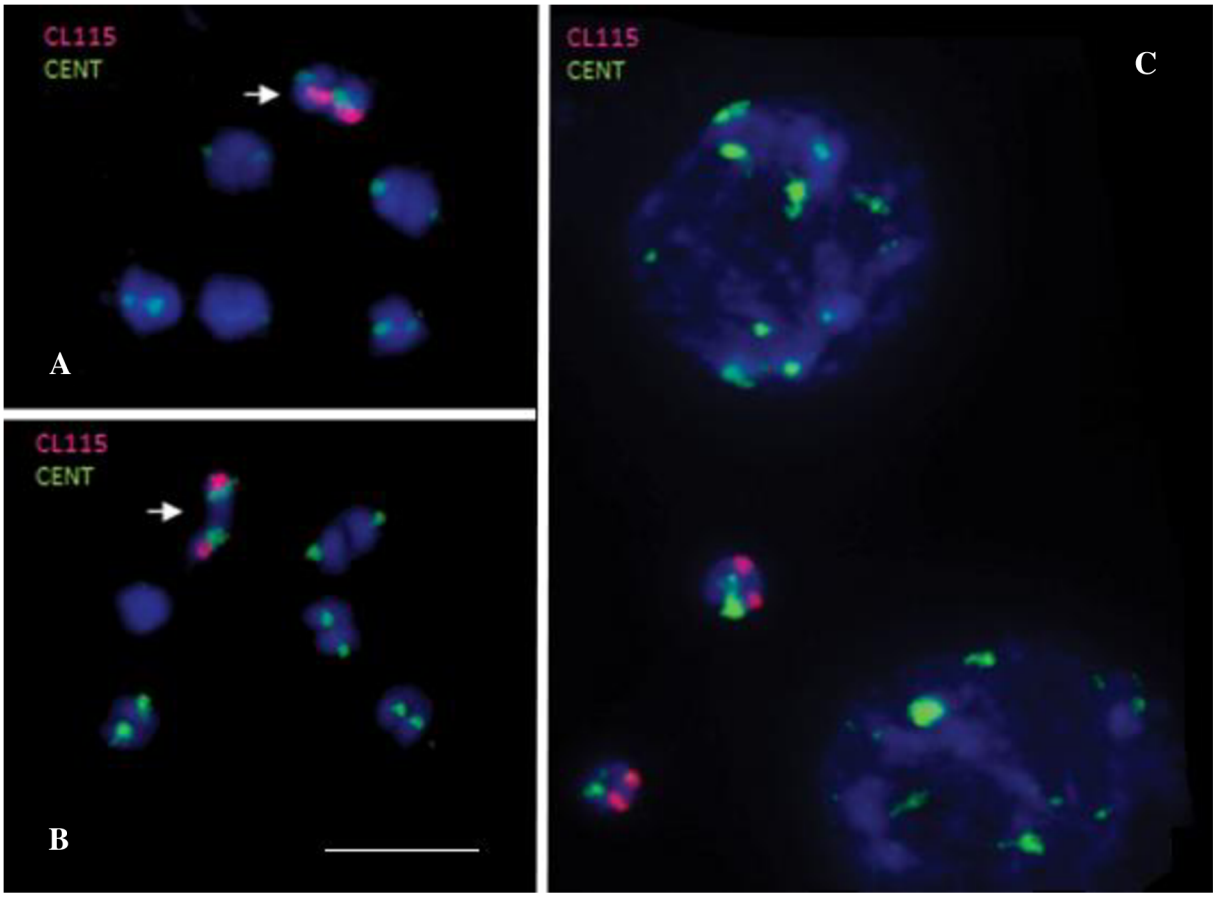


Obrázek 6: Efekt B chromozómů na fenotyp *S. purpureosericeum*. A – 4B rostlina s výrazně nižní vitalitou oproti 0B rostlině, B – květ 4B rostliny se sníženou fertilitou, C – květ 0B rostliny, D – produkce semen B0 rostlin a 2B rostlin.

Karafiátová et al., 2021 stanovili velikost genomu *S. purpureosericeum* na 2,21 Gb a cytometricky byla odhadnuta velikost B chromozómu na 421 Mb. *S. purpureosericeum* je jediným čirokem s B chromozómy, u kterého existují nějaké poznatky o jeho sekvenci. V rámci studie byla provedena analýza repetitivních sekvencí na platformě Illumina, která vedla k identifikaci shluků specifických pro B chromozomy u druhu *S. purpureosericeum* (Karafiátová et al., 2021). Na základě této analýzy odvozeny první B-specifické PCR a cytogenetické markery pro tento druh. Tyto nově vyvinuté markery lze nyní úspěšně používat pro další studium B chromozomů v *S. purpureosericeum* a pro rutinní screening. Například v rámci bakalářské práce byly využity 2 spolehlivé PCR markery (24, 27) pro

detekci přítomnosti B chromozómů v embryu a endospermu v F1 potomstvu linií 578 (Harnádková, 2022).

Studium B chromozómů u divokých čiroků je komplikováno jeho nestabilitou v pletivech. Z pletiv listu, kořene a stonku je B chromozóm zcela eliminován (Darlington et al., 1941; Karafiátová et al., 2021). Spolehlivě lze B chromozóm tohoto divokého čiroku detekovat pouze v květenství. Mechanismy eliminace B chromozómů u *S. purpureosericeum* prozatím nebyly více studovány a jeho objasnění je předmětem dalšího výzkumu.

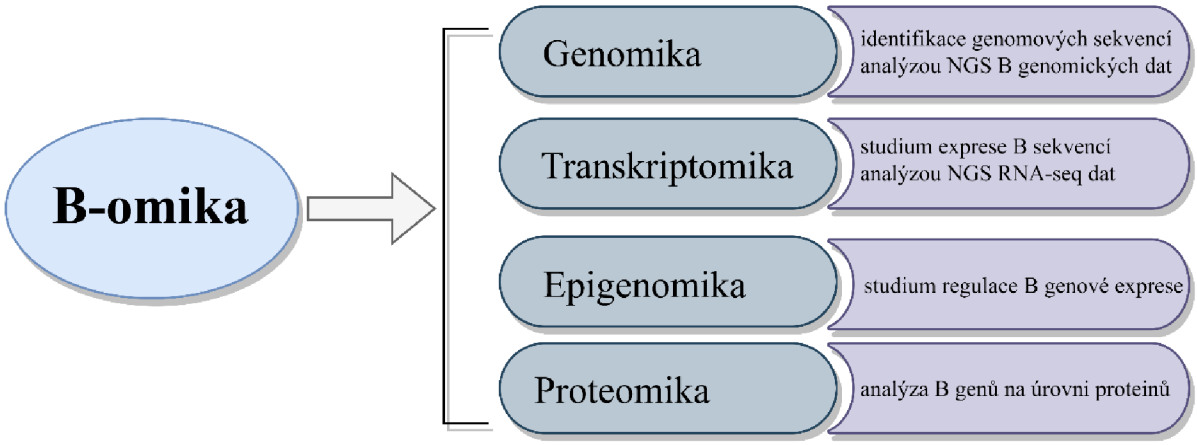


Obrázek 7: Lokalizace B chromozómu *S. purpureosericeum* v meiotické metafázi I pomocí FISH s B-specifickou repeticí. A, B – párování B chromozómů dělících se PMC. Šipkami naznačeny bivalenty B chromozómu. Červená znázorňuje B-specifickou sondu CL115 a zelená centromerickou sondu. Měřítko 5 μm. A – kruhové bivalenty, 2B rostliny, B – tyčinkové bivalenty 2B rostliny, C – detail tapetálních jader; mikrojádra s B chromozómy a jádra bez B chromozómů. Modifikováno z Karafiátová et al., 2021.

3.3 Sekvenační přístupy studia B chromozomů

Sekvenování nové generace (NGS) spolu s přístupy analýz dat výrazně přispěly k pokroku ve veškerých oblastech biologického výzkumu, a studium B chromozómů není výjimkou. V posledních letech se ve studiu B chromozómů využívá různých postupů, které se opírají

o základní „omics“ technologie. Tyto v současnosti velmi rozvíjené a využívané „multi-omics“ techniky (Obr. 8) se zasloužily o pochopení původu, evoluce, genomického složení a biologického významu B chromozómů. Jedná se zejména o genomiku, transkriptomiku, epigenomiku a proteomiku, které se podílely na vzniku nové oblasti výzkumu, která se nazývá „B-omika“ (z angl. B-omics).



Obrázek 8: Multi-omické přístupy probíhajícího výzkumu B chromozómů (B-omiky). Modifikováno z Ahmad & Martins, 2019.

Strategie studia strukturní genomiky B chromozomů lze rozdělit na přímý a nepřímý přístup (Ruban et al., 2017). Přímý přístup je založen na sekvenování DNA pocházejících pouze z B chromozómů (Obr. 9A). Hlavní výhoda přímého přístupu je významná redukce komplexity sekvenačních dat. B chromozómy lze izolovat mikrodisekcí či tříděním průtokovým cytometrem. Výhodou přímého přístupu je vysoká pravděpodobnost, že získané sekvence pochází z B chromozómu. Pomocí mikrodisekce byly izolovány B chromozómy několika živočišných druhů (Amorim et al., 2016; Bugrov et al., 2007; Valente et al., 2014) či kukuřice (Cheng & Lin, 2003) a žita (Sandery et al., 1991). Izolace B chromozómů průtokovou cytometrií je typická pro živočišné druhy (Graphodatsky et al., 2005; Ventura et al., 2015), nicméně byla úspěšně použita také při studie B chromozómu žita (Martis et al., 2012).

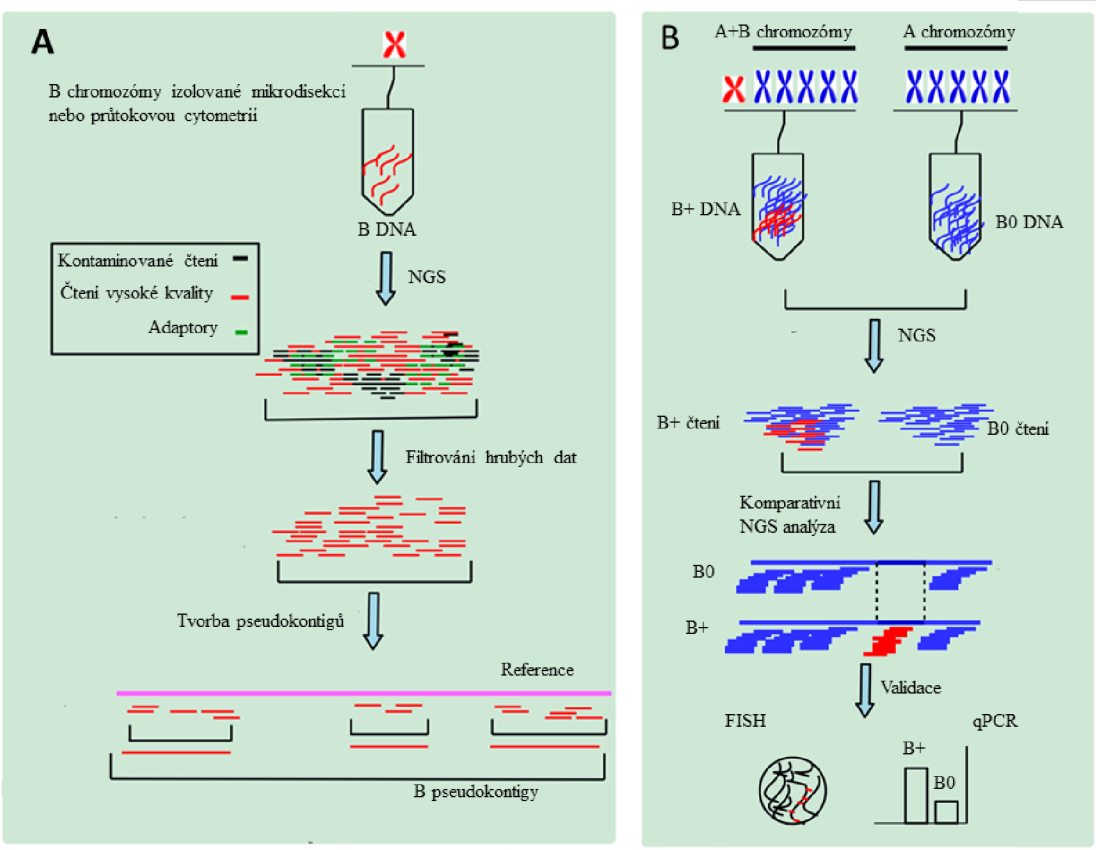
Ke kvalitnímu sestavení sekvence B chromozómu po mikrodisekci je zapotřebí dostatečná sekvenační hloubka spolu s aplikací pokročilých bioinformatických nástrojů. Takovou metodu například vypracovali (Thind et al., 2017). Data získaná z experimentů s nízkou sekvenační hloubkou nemohou poskytnout dostatečné pokrytí sekvence, které je potřebné pro zpětné sestavení sekvence celého chromozómu. Největší problém, který může vzniknout při sekvencování DNA z mikrodisektovaného materiálu, je šum generovaný kontaminací

z necílených chromozómů, necílených druhů (lidská kontaminace) a abundantních sekvencí vzniklých amplifikací (Ruban et al., 2017).

Nevýhodou pro třídění B chromozómů průtokovou cytometrii je nutnost existence vypracovaného protokolu synchronizace buněčného cyklu u daného druhu (Ruban et al., 2017). Důležitý faktor pro úspěšné separování B chromozómu je jeho dostatečně odlišná velikost od A chromozómů, což je u většiny druhů, kde B chromozóm nabývá stejné velikosti, problém. V neposlední řadě je limitací související s průtokovou cytometrií obtížné rozlišení mezi fragmentovanými B a A chromozómy, ale tomuto zdroji kontaminace lze některými postupy předejít (Doležel et al., 2014). Získaná DNA B chromozómů z mikrodisekce či průtokové cytometrie je následně sekvenována a bioinformatickými nástroji jsou detekovány B sekvence.

Nepřímé metody jsou založeny na porovnání sekvencí vzorků nesoucích B chromozómy (B+) a vzorků nenesoucích B chromozómy (B0) (Obr. 9B). Nadbytečné chromozómy jsou obvykle bohaté na repetitivní sekvence, které mohou být identifikovány shlukovou analýzou NGS sekvenačních čtení založené na podobnosti. Takového přístupu využívá nástroj ReapeatExplorer, který identifikuje shluky často překrývajících se čtení a interpretuje je jako repetitivní elementy (Novák et al., 2010). Tento nástroj také předpovídá počet kopií jednotlivých sekvencí na základě četnosti sekvenačních čtení. K identifikaci specifických repetitivních sekvencí přítomných na B chromozómech je možné provést analýzu v komparativním módu, kde proběhne simultánní shlukování čtení z B+ a B0 vzorků. Tento přístup byl využit u žita (Martis et al., 2012), *Plantago lagopus* (Kumke et al., 2016) a také u *S. purpureosericeum* (Karafiátová et al., 2021).

Nejpoužívanější softwary, které umožňují přiřazení sekvencí, je Burrows-Wheeler Alignment tool (BWA) (Li & Durbin, 2009) a Bowtie2 (Langmead & Salzberg, 2012). Následně jsou vzniklé SAM/BAM soubory prohledávány pro regiony s různým počtem přiřazených čtení například v softwaru Integrative Genomics Viewer (IVG) (Robinson et al., 2011).



Obrázek 9: Metody studia sekvence B chromozómu. A – Přímá metoda. B chromozóm je izolován mikrodisekcí nebo průtokovou cytometrií. Izolována DNA je sekvencována a přefiltrovaná hrubá data jsou bioinformatickými nástroji překrývány za tvorby pseudokontigů. B – Nepřímá metoda. DNA izolována z karyotypu s B chromozómy a DNA bez B chromozómů je sekvenována. Sekvence B chromozómu jsou získané komparativní analýzou NGS čtení obou datasetů. Modifikováno z Ahmad & Martins, 2019.

3.3.1 *De novo* referenční sekvence

Tvorba referenční sekvence je proces rekonstrukce genomu z jednotlivých sekvenačních čtení. Ta mohou být různých délek a pocházet z různých sekvenačních platforem. NGS technologie, které umožnily revoluci genomiky, jsou vysoce výkonné a poskytují miliony sekvenačních reakcí současně. Současně nejpoužívanější NGS platformy zahrnují Illumina/Solexa a Ion torrent (Bennett, 2004; Shen et al., 2005; Rothberg et al., 2011). NGS platformy mají však své nedostatky. Následkem krátké délky čtení (méně než 300 bp) mohou vznikat *de novo* hrubé genomy („assembly“ s velkým množstvím chybějících úseků, nepřesnostmi v sekvenci a fragmentace genů do mnoha kontigů (Denton et al., 2014). Tyto limitace lze částečně vyřešit použitím metod třetí generace (TGS z anglického „Third Generation Sequencing“), které oproti NGS technologiím produkují delší čtení, mají velkou přesnost v GC regionech a minimalizují sekvenační chyby vzniklé PCR amplifikací (Wee et

al., 2019). Mezi TGS technologie se řadí platformy Oxford Nanopore Technologies (ONT) a Single-Molecule Real-Time (SMRT).

Existující assemblery implementují jeden ze dvou přístupů – Overlap/Layout/Consensus (OLC) nebo de Bruijnovy grafy (Li et al., 2012). OLC sestavení sekvence využívá celé čtení, z jejichž překryvů je v prvním kroku vytvořen graf. Následně je grafem vyhledávaná ideální Hamiltonova cesta – taková cesta, která prochází všemi vrcholy grafu právě jednou. V posledním kroku se ze čtení, které se nachází podél této cesty, určuje finální konsensus sekvence (Pop, 2009). V současnosti je tento algoritmus používán programy pro sestavení genomu z dlouhých čtení (Sohn & Nam, 2018) jako například Canu (Koren et al., 2017), Flye (Kolmogorov et al., 2019), MaSuRCA (Zimin et al., 2017), SMARTdenovo (Liu et al., 2021) použitých v této práci.

De Bruijnovy grafy nevyužívají k sestavování sekvence celé čtení, ale množinu všech podřetězců z tohoto čtení o délce „k“, tedy z k-merů (Miller et al., 2010). K-mery mají však krátkou délku, což komplikuje zpětné sestavení repetitivních úseků. To může vést ke ztrátě velkého množství informací, jelikož algoritmus není schopen rozpoznat sousedící repetitivní sekvence (Sohn & Nam, 2018). Tento algoritmus je obvykle využíván nástroji, které sestavují *de novo* sekvenci z krátkých čtení produkované NGS technologiemi. Jedním z nejpoužívanějších nástrojů využívajících de Bruijnovy grafy je Abyss (Simpson et al., 2009).

Na rozdíl od genomů živočichů jsou rostlinné genomy větší a komplexnější. Jsou typické vyšší ploidií, heterozygozitou či repetitivními elementy (Gregory, 2005). Z tohoto důvodu není žádoucí skládat genomy pouze z krátkých čtení poskytnutých NGS technologiemi, protože vzniklé assembly můžou být vysoce fragmentované, nedokončené a obsahovat velký počet kontigů a scaffoldů.

3.3.2 Transkriptomická RNA-seq analýza

Již od objevu role RNA se identifikace transkriptů a kvantifikace genové exprese řadí mezi základní stavební pilíře molekulární biologie. Identita a kvantifikace může být zkoumaná v jednom experimentu zároveň, a to velmi výkonnou, robustní a přizpůsobivou metodou RNA-sekvencováním (RNA-seq), která poskytuje alternativu ke klasickým microarray analýzám transkriptomu (Wang et al., 2009). RNA-seq je standardně používaná metoda, nicméně neexistuje jeden optimální protokol zpracovávání dat. Postupy se odlišují na základě různých nástrojů kvantifikace transkriptů, normalizace a diferenciální exprese (Conesa et al.,

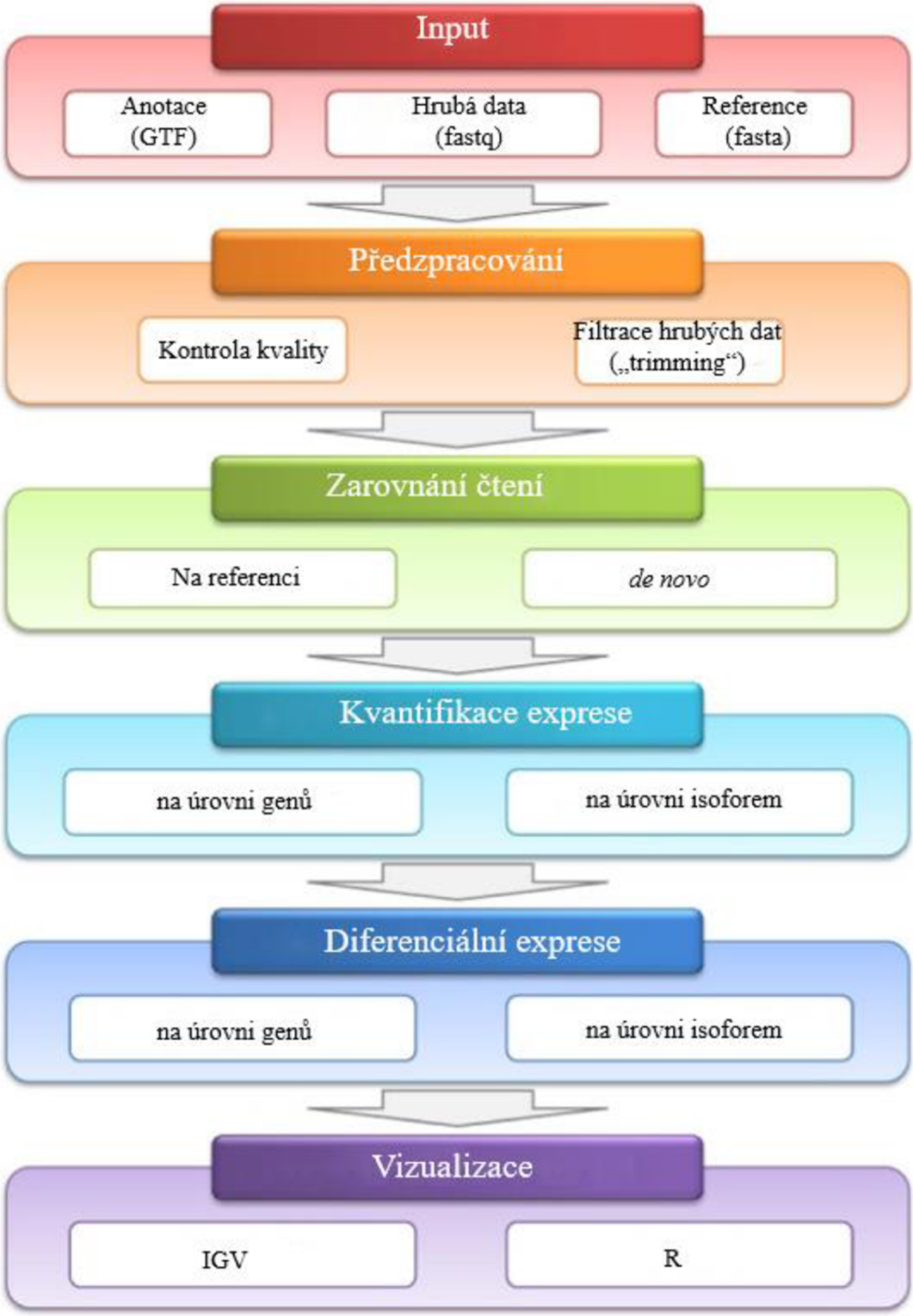
2016; Corchete et al., 2020). Analýza RNA-seq dat se typicky skládá z několika základních kroků – filtrace adaptorů a chybných dat („trimming“), překrytí transkriptomických čtení na referenční sekvenci („alignment,“¨), počítání transkriptů („counting“) a normalizace sekvencovaných čtení. Typické zpracování dat RNA-Seq experimentu je znázorněno na obrázku 10.

V důsledku přípravy sekvenačních knihoven i vlastní sekvence mohou sekvenační data formátu fastq obsahovat chyby a kontaminace, jakými jsou například adaptérové sekvence, nukleotidy určené s nízkou přesností a nadměrně zastoupené sekvence (k-mery). Pro dosažení lepších výsledků je vhodné nejdříve provést kontrolu kvality hrubých dat a na základě výsledků z této kontroly chyby eliminovat příslušným filtrováním (trimováním) (Del Fabbro et al., 2013). Délku čtení a jiné parametry k této filtraci je nutno volit obezřetně, aby nedošlo k nežádoucímu ovlivnění výsledků následné analýzy (MacManes, 2014; Williams et al., 2016). Pro mapování a kvantifikaci RNA-Seq čtení na genové úrovni není trimování dat nutné, jelikož některé moderní programy na mapování zahrnují jistou míru filtrace hrubých dat (Liao & Shi, 2020). Nejpoužívanější trimovací nástroje pro krátké RNA-Seq čtení jsou například Skewer (Jiang et al., 2014), fastp (Chen et al., 2018) nebo Trimmomatic (Bolger et al., 2014). Vhodným nástrojem pro kontrolu obsahu kontaminace, PCR artefaktů a sekvenačních chyb v Illumina datech před a po trimování jsou fastQC (Andrews, 2010) nebo NGSQC (Dai et al., 2010). Tyto programy poskytnou informace o kvalitě bází, GC obsahu, duplikovaných sekvencí či přítomnosti adaptérů.

Dalším krokem RNA-Seq analýzy bývá zpravidla mapování na referenční sekvenci. Existuje řada bioinformatických nástrojů k mapování RNA-Seq dat založených na různých algoritmech, z nichž nejpoužívanější bowtie2 (Langmead & Salzberg, 2012), HISAT2 (Kim et al., 2019) a STAR (Dobin et al., 2013). Srovnávací studie testovala 7 nástrojů pro mapování pro RNA-Seq dat *Arabidopsis thaliana*, a označila je všechny za vhodné k použití k RNA-Seq analýze, jelikož měly všechny nástroje srovnatelné výsledky (Schaarschmidt et al., 2020). Nicméně, srovnání mapovacích nástrojů na složitější genomy obsahující mnoho nekódujících sekvencí ještě nebylo provedeno. V případě nedostupnosti referenční sekvence je možné vytvořit transkriptom *de novo*. Jedním z nejpoužívanějších nástrojů na sestavení transkriptomu *de novo* je Trinity (Grabherr et al., 2011).

Čtení namapované na referenční sekvenci, musí být přiřazeny genu v procesu kvantifikace neboli countigu (Mortazavi et al., 2008). Tento krok RNA-Seq analýzy umožňují například programy Cufflings (Trapnell et al., 2010), HTSeq (Anders et al., 2015) a RSEM (Li &

Dewey, 2011). Přístup těchto nástrojů využívá GTF (gene transfer format) souboru, který obsahuje genomové souřadnice exonů a genů.



Obrázek 10: Schéma RNA-Seq experimentu. Vstupem je referenční sekvence (fasta formát) s její anotací (GTF formát) a transkriptomická data (fastq). Kvalita hrubých dat je před zarovnáním na

referenci kontrolována a může následovat jejich filtrace adaptorů a nukleotidů nízké kvality. Zarovnání probíhá na referenční sekvenci a v případě její absence je tvořen transkriptom *de novo*. Následuje kvantifikace a diferenciální exprese na úrovni genů či jejich isoforem. Výsledná data jsou statisticky zpracovávána a vizualizována v IGV nebo pomocí jazyka R. Modifikováno z Yang & Kim, 2015.

Po kvantifikaci čtení je genová exprese vyjádřena pomocí různých kvantifikátorů, jakou jsou počet transkriptů z milionu (TPM), fragmentů na kilobázi na milion mapovaných čtení (FPKM) nebo čtení na kilobázi na milion mapovaných čtení (RPKM). Tato normalizace je důležitá, protože výsledky kvantifikace čtení mohou být zkreslené biologickou variabilitou a technickými faktory, jako je například sekvenační hloubka a délka transkriptu. Zhodnocením vhodnosti použití různých kvantifikátorů genové exprese se zabývají různé studie (Conesa et al., 2016; Zhao et al., 2021).

Jeden z cílů transkriptomických studií je porovnávání počtu čtení (míry exprese genů) mezi různými vzorky či stejnými vzorky za různých podmínek. K tomu slouží různé algoritmy založené na pravděpodobnostních a statistických modelech. Modelu založeného na negativní binomické distribuci využívá edgeR (Robinson et al., 2010) a DeSeq2 (Anders & Huber, 2010). Oba nástroje jsou přístupné jako balíček v softwaru Bioconductor založeného na programovacím jazyce R (R Core Team, 2021). Nejčastější přístup k vyhodnocování pomocí těchto nástrojů je testování nulové hypotézy, že není rozdíl v expresi, tedy že změna exprese mezi vzorkem a kontrolou vyjádřená ve formě „logaritmic fold change“ ($\log_2$FC) je pro daný gen nulová. Cílem diferenciální analýzy je potom získat seznam genů, pro které je nulová hypotéza zamítnuta (Love et al., 2014).

3.3.2.1 Transkriptomické analýzy rostlinných genomů s B chromozómy

Navzdory předchozím domněnkám, že B chromozómy obsahují pouze nekódující sekvence, se v posledních letech nahromadily důkazy transkripčně aktivních sekvencí lokalizovaných na B chromozómech (Huang et al., 2016; Navarro-Domínguez et al., 2017; Aldrich et al., 2017) (Navarro-Domínguez et al., 2019; Hong et al., 2020). V současnosti je analýza genů za pomoci genomických a transkriptomických analýz na B chromozómech jedním z hlavních cílů studia B chromozómů. B chromozómy rostlin obvykle neposkytují hostitelům žádnou výhodu, a proto se transkriptomické analýzy soustředí zejména na odhalení genů a obecných mechanismů stojících za dědičnými mechanismy B chromozómů, případně eliminace B chromozómů z některých pletiv (Boudichevskaia et al., 2020).

Komparativní RNA-Seq analýza 2B a B0 vzorků žita (*Secale cereale* L.,) a pšenice (*Triticum aestivum* L., cv „Chinese Spring” s žitnými B chromozómy) potvrdila dřívější

výzkum ovlivnění genové exprese A chromozómů B chromozómy (Boudichevskaia et al., 2022). Data získané izolací RNA z 2B a B0 prašníků žita i pšenice odhalila 5-6% ovlivnění standardních transkriptů z A chromozómů v přítomnosti dvou B chromozómů původem z žita (Boudichevskaia et al., 2022). Mimo jiné studie odhalila 297 B-specifických transkriptů žita a 939 B-specifických transkriptů pšenice, z toho 29 % B-specifických transkriptů žita a pšenice byly sdíleny (Boudichevskaia et al., 2022).

Podobné transkripční efekty byly nalezeny i u B chromozómů kukuřice (Huang et al., 2016 Shi et al., 2022). Transkripčně aktivní geny B chromozómů kukuřice vykazují podobnost ke svým homologům na A chromozómech a za jejich přítomnosti je diferenciálně exprimováno až 130 genů související s buněčným metabolismem nebo vázáním nukleotidů (Huang et al., 2016). Komparativní RNA-Seq analýza kukuřic nesoucích 0B, 1B a 6B chromozómů, odhalila 758 protein kódujících sekvencí na B chromozómu, z nichž alespoň 88 bylo exprimováno (Blavet et al., 2021).

Další oblast výzkumu B chromozómu se zaměřuje na jeho nestabilitu v rostlinných pletivech. Pionýrem v této oblasti je B chromozóm *Aegilops speltoides,* který podléhá eliminaci v pletivech kořene již během brzké embryogeneze (Ruban et al., 2020). U tohoto druhu byla provedena laserová mikrodisekce B0 a B+ embryí v oblasti, kde dochází k eliminaci, a následně komparativní RNA-Seq analýza (Boudichevskaia et al., 2020). Bylo detekováno 1 457 upregulovaných genů v B+ vzorcích a 2 726 downregulovaných. Významně upregulované ($log_2FC > 1$, p-value < 0.05) byly 3 izoformy genu *Nuf2*, který je evolučně konzervován a jehož funkcí je kontrola regulace bipolárního napojení mikrotubulů sesterských chromatid před a po anafázi a jeho zvýšená exprese v savcích způsobuje defekty v chromozómové segregaci (Zhang et al., 2015). Mezi další slibné kandidáty podílející se na procesu eliminace B chromozómu *A. speltoide*s jsou geny kódující proteiny Sgo1 a Mis12, které se podílí na formaci kinetochoru (Boudichevskaia et al., 2020).

Na základě přísných parametrů bylo v této studii dále vyfiltrováno 341 B-specifických transkriptů, z toho 70 jsou funkčně anotované geny. Mezi geny se specifickou funkcí byly nalezeny dva geny kódující proteiny podobným kinázám. Jeden z nich vykazoval 94% identitu proteinu KIN-14C *Aegilops tauschii subsp tauschii*. V *Arabidopsis thaliana* je KIN-14C důležitý pro správnou akumulaci mikrotubulů na pólech dělicího vřeténka během profáze v mitóze (Mitsui et al., 1993). Komparativní RNA-Seq studie embryí *A. speltoides* zaznamenala mnoho dalších genů, které jsou jednoznačně napojeny na buněčné dělení a jsou tak vhodnými kandidáty na validaci eliminace B chromozómů (Boudichevskaia et al., 2022).

Důvod, proč B chromozómy *A. speltoides* podléhají eliminaci, zůstává prozatím dále neznámý. Autoři studie nicméně polemizují, že eliminace B chromozómů může být ochranným mechanismem pro udržení v populaci, poněvadž jeho přítomnost ve všech pletivech má negativní efekt na vitalitu a fertilitu rostlin (Boudichevskaia et al., 2020). Pletivově-specifická studie odhalila B-specifické transkripty kódující regulátory růstů a vývoje rostliny, které jsou v souladu s touto hypotézou (Boudichevskaia et al., 2020).

4 Materiál a metody

4.1 Biologický materiál

Pro experiment byly použity diploidní rostliny druhu *Sorghum purpureosericeum* nesoucí B chromozómy (2n=2x=10+B). Původní semena byla získána z ICRISAT (International Crops Research Institute for the Semi-Arid Tropics, položka IS 18947) a jeho F1 potomstvo bylo přemnoženo na ÚEB AV ČR a označeno jako linie 578. Rostliny byly pěstované ve fytotronu v režimu světlo/tma (10h 29 °C/14h 25 °C) při 50% vlhkosti.

4.2 Roztoky, chemikálie

Použité chemikálie

- Agaróza (Sigma-Aldrich, kat. č. A9539)
- Bromfenolová modř (Sigma-Aldrich, kat. č. B0126)
- Cresol Red (Sigma-Aldrich, kat. č. P5631)
- Dodecylsulfát sodný (SDS) (Sigma-Aldrich, kat. č. 817034)
- Ethanol 99,8% (Lach-Ner, kat. č. 20025-U99) • Ethidium bromid (Sigma-Aldrich, kat. č. E8751)
- Gene Ruler 100 bp DNA Ladder (Fermentas, kat. č.: SM0321)
- Glycerol (Sigma-Aldrich, kat. č.: G5516)
- Kyselina boritá (Lach-Ner, kat. č. 10017-AP0)
- Kyselina ethylendiamintetraoctové (EDTA) (Sigma-Aldrich, kat. č. E5134)
- Nukleotidy: dATP, dCTP, dGTP, dTTP, každý 100mM (VWR International, kat. č. 733-1364)
- Pufr pro *Taq* polymerázu (New England Biolabs, kat.č. B7002S)
- Taq DNA polymeráza (New England Biolabs, kat. č. M0209L)
- Tekutý dusík
- Tris base (Sigma-Aldrich, kat. č. 77-86-1)
- Xylencyanol (Sigma-Aldrich, kat. č.: X4126)

Použité roztoky

- 0,5M EDTA (1 l): 186,1 g dihydrátu disodné soli ethylendiamintetraoctové kyseliny rozpustit v 800 ml destilované vody, doplnit do 1 l a pH upravit na 8
- 0,5x TBE pufr (1 l): 100 ml zásobního 5x TBE pufru a doplnit na 1 l destilovanou vodou
- 5x TBE pufr (1 l): 54 g Tris báze a 27,5 g kyseliny borité rozpustit ve vodě, přidat 20 ml 0,5M EDTA, doplnit vodou na 1 l, upravit pH na 8
- 6x STOP C (10 ml): smíchat 5 mg bromfenolové modři a 5 mg xylencyanolu, přidat 2 ml 0,5M EDTA, 4,3 ml 99,9% glycerolu a 1 ml 10% SDS, doplnit vodou na 10 ml
- 0,05% Cresol Red (100 ml): 50 mg Cresol Red a 7,5 g sacharózy za stálého míchání a zahřívání na 50 °C rozpustit v 100 ml destilované vody, po rozpuštění přefiltrovat přes 0,2µm filtr a rozpipetovat do alikvotů po 1 ml
- Marker molekulového hmotnosti: smíchat 20 µl Gene Ruler 100 bp, 200 µl 6x STOP C a 300 µl vody, pečlivě zvortexovat
- Roztok ethidium bromidu (1 l): ve 100 ml destilované vody rozpustit 50 g ethidium bromidu, doplnit destilovanou vodou do 1 l
- Směs neznačených nukleotidů: do 5,25 µl vody napipetovat 0,5 µl 10mM dATP, 0,5 µl 10mM dCTP, 0,5 µl 10 mM dGTP, 1,65 µl 2mM dTTP a 1,6 1mM dUTP

4.3 Přístroje

- 2100 Bioanalyzer Instrument (Agilent)
- Centrifuga Mega Star 600R (VWR International)
- Elektroforetická horizontální aparatura Owl A6 (Thermo Fisher Scientific)
- Laminární box (VWR International)
- Laboratorní váha ViBRA AJ-820CE (Shinko Denshi)
- Stereomikroskop SZX16 (Olympus LS) s CCD kamerou
- Spektrofotometr NanoDrop One/OneC Microvolume UV-Vis (Thermo Fisher Scientific)
- Stolní centrifuga MiniStar silverline (VWR International)
- Termomixér Thermal Shake Touch (VWR International)
- Thermoycler C1000 touch (Bio-Rad)
- UV transluminátor InGenius3 (Syngene)
- Vortex REAX Top (Heidolph)

- Zdroj k elektroforéze Owl EC300XL2 (Thermo Fisher Scientific)

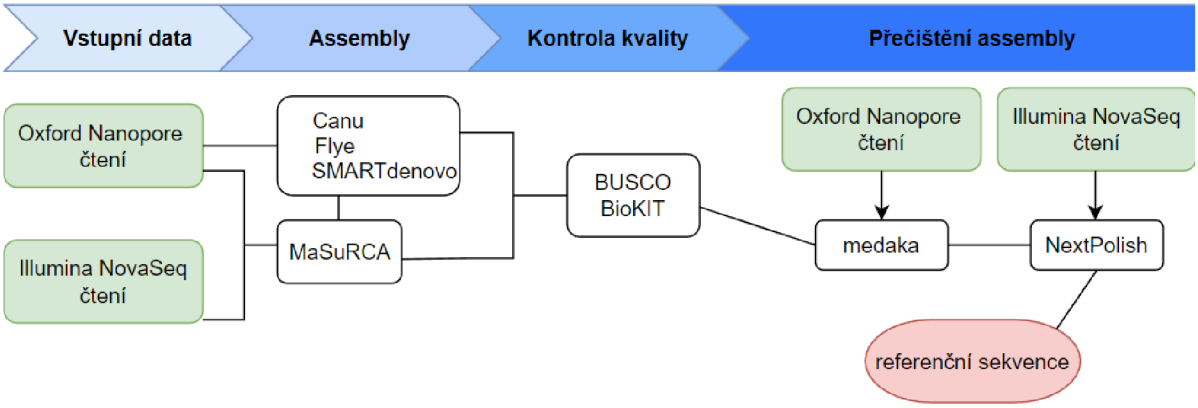
4.4 Použité kity

- Genomic DNA Purification Kit (Monarch® kat.č.: T3010S)
- NEBNext® Single Cell/Low Input RNA Library Prep Kit for Illumina® (kat.č. E6420L)
- RNA 6000 Nano Kit (Agilent, kat. č.: 5067-1511)
- RNA 6000 Pico kit (Agilent, kat. č.: 5067-1513)
- Total RNA Miniprep Kit (Monarch®, kat. č.: T2010)

4.5 Experimentální postupy

4.5.1 Tvorba a analýza referenční sekvence *S. purpureosericeum*

Vstupními daty byly dva NGS datasety z izolovaných květů *S. purpureosericeum* nesoucích B chromozóm – Illumina NovaSeq (krátké sekvenační čtení) a data z Oxford Nanopore platformy (ONT, dlouhé sekvenační čtení). Na sestavení prvotní referenční sekvence *S. purpureosericeum* byly použity bioinformatické nástroje (assemblery) Flye v2.9 (Kolmogorov et al., 2019), Canu v2.2 (Koren et al., 2017), a SMARTdenovo (Liu et al., 2021) pro data z dlouhých čtenía assembler MaSuRCA v4.0.7 (Zimin et al., 2017) pro hybridní assembly z krátkých i dlouhých čtení. Stručné schéma postupu při tvorbě referenční sekvence je znázorněna na obrázku 11.



Obrázek 11: Tvorba referenční sekvence. Vstupní data krátkého a dlouhého čtení byly použity čtyřmi nástroji k sestavení hrubé assembly. Na základě kontroly kvality byla ze čtyř verzí referenční sekvence vybrána ta nejlepší, která byla následně podrobena korekci hrubými daty. Vstupní data jsou znázorněny zeleně, výstup červeně a bíle použité nástroje. Výpočty probíhaly formou dávkovacích nebo interaktivních úloh v MetaCentru.

Úlohy potřebné k sestavování rostlinných genomů jsou náročné na výpočetní kapacitu a z tohoto důvodu bylo pro spuštění úloh všech assemblerů využito zdrojů MetaCentra (https://metavo.metacentrum.cz/). V unixovém příkazovém řádku byly vytvořeny skripty pro shell, který každé ze 4 úloh určil výpočetní kapacitu, vstupní a výstupní data a parametry assembleru. Příkazy pro úlohy a jejich parametry jsou uvedeny v Příloze 1.

Výstupy z těchto nástrojů byly podrobeny Benchmarking Universal Single-Copy Ortholog (BUSCO) v5.4.4 analýze (Manni et al., 2021), implementované na linuxovém serveru v prostředí Anaconda (*Anaconda Software Distribution*, 2016). V příkazu byl zahrnutý parametr pro OrthoDB soubor, který obsahuje dataset genů taxonu *Poales*. Výstupy textového

formátu se základními statistikami byly přesunuty do složky BUSCO_summaries.txt, pro kterou byl puštěn skript v rámci nástroje BUSCO pro grafické zobrazení:

Příkaz pro grafické zobrazení:

```
python /software/busco/3.0.2b/scripts/generate_plot.py -wd
BUSCO_summaries
```

Výstupním souborem je pak PNG soubor s grafem statistik pro všechny nástroje. Graf je vytvářen programovacím jazykem R s balíčkem ggplot2.

Kvalita genomových verzí nástrojů získaných pomocí Canu, Flye, MaSuRCA a SMARTdenovo se dále hodnotila na základě základních statistik genomů, jako je například celková délka referenční sekvence v jednotkách páru bází (bp) nebo počet kontigů. Dále se sledovaly parametry vypovídající o kontinuitě genomu – délka nejkratšího kontigu, jehož delší kontigy pokrývají alespoň 50 % genomu (N50), a nejmenší počet kontigů, jejichž součet délek utváří polovinu velikosti genomu (L50).

K výpočtu těchto parametrů byl použit nástroj BioKIT v0.5.0, implementovaný na linuxovém serveru v Anaconda prostředí (Steenwyk et al., 2022).

Na základě výsledku BUSCO a BioKIT analýzy byl vybrán výstup SMARTdenovo jako nejkvalitnější z vytvořených sekvencí. Následovala jeho korekce (přečišťování, z anglického „polishing“) hrubými ONT čtení nástrojem Medaka v1.6.0 a poté Illumina čtení nástrojem NextPolish v1.4.1. Příkazy pro korekci nástrojem Medaka a NextPolish a jejich parametry jsou přiloženy v Příloze 1.

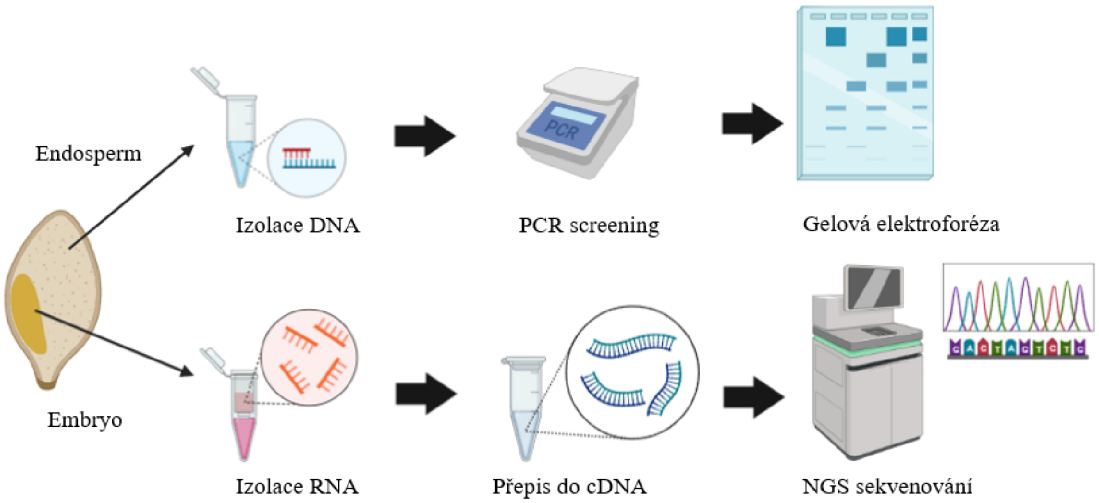
Anotace přečištěné verze SMARTdenovo referenční sekvence byla vytvořena mimo pracoviště.

4.5.2 RNA-Seq analýza

4.5.2.1 Izolace embryí

Vyvíjející se semena byla odebrána z B+ i B0 rostlin *S. purpureosericeum* linie 578_1 a 578_3 (2n = 2x = 10). Pod lupou byly při zvětšení 1,6x z každého vyvíjejícího se semene odstraněny vnější obaly a semeno bylo rozděleno na embryo a endosperm. Embryo bylo za pomocí kamery zdokumentováno k pozdějšímu vyhodnocení stáří embrya. Embryo i endosperm byly zvlášť vloženy do 1,5ml mikrozkumavek, zmraženy v tekutém dusíku a skladovány při -80 °C. Bylo odebráno 7 B+ i B0 vzorků od každé kategorie stáří –7-denních

embryí (7DAP), 14-denních embryí (14DAP), 21-denních embryí (21DAP) a 28-denních embryí (28DAP). Embryo i endosperm byly využity v následujících experimentech, jejichž průběh je znázorněn na obrázku 12.



Obrázek 12: Schéma RNA-Seq experimentu. První část zahrnuje izolaci DNA ze zmraženého endospermu. DNA byla dále použita pro PCR experiment s B-specifickými markery. Produkty byly ověřeny gelovou elektroforézou a dle výsledku (ne)přítomnosti byly vzorky rozřazeny do kategorií. Druhá část se skládá z izolace RNA ze zmraženého embrya. RNA byla použita pro RNA sekvenování NGS metodou Illumina.

4.5.2.2 Izolace DNA z endospermu.

Skladovaný endosperm byl po zmražení v tekutém dusíku homogenizován malým plastovým tloučkem. Z endospermu byla vyizolována DNA pomocí kitu pro malé množství pletiva (Monarch® Genomic DNA Purification Kit) dle protokolu výrobce v sekci „Animal Tissue“. Bylo použito 10 µl proteinázy K a inkubace probíhala po dobu 3 h při 1400 rpm a 56 °C, eluce byla prováděna do 50 µl. Koncentrace vyizolované DNA byla kontrolně změřena pomocí spektrofotometru Nanodrop.

4.5.2.3 PCR screening pro zjištění (ne)přítomnosti B chromozómů

Z důvodu nepravidelného přenosu B chromozómu a možného cizospřášení bylo nutné potvrdit (ne)přítomnost B chromozómů v embryích B pozitivních i B negativních rostlin. Nondisjunkce B chromozómu v první pylové mitóze dává vzniknout endospermu a embryu, které oba obsahují B chromozóm. (Ne)přítomnost B chromozómu v endospermu tedy svědčí i o stejném statusu (ne)přítomnosti B chromozómu v embryu.

Na izolované DNA ze všech endospermů byla provedena PCR reakce pro zjištění přítomnosti B chromozómů pomocí dvou B specifických markerů („utg312_gene155643_2“

a „utg5443_gene386712“) (Tab. 1). Reakční směsi obsahovaly: 50 ng DNA izolované z endospermu, 1x koncentrovaný pufr pro *Taq* Polymerázu, 1 U *Taq* DNA polymerázy, 100 mM směsi neznačených nukleotidů, 1x koncentrované barvivo Cresol Red a R a F primery jednoho z makerů o koncentraci 0,5 µM. Směs byla doplněna destilovanou vodou do 25 µl. Produkty byly amplifikovány za podmínek: 94 °C/3 min, 30 cyklů: 94 °C /30 s, 62 °C /30 s, 72 °C /45 s; 72 °C /10 min. Přítomnost PCR produktu byla ověřena elektroforeticky (1,2% agaróza v 0,5x TBE pufru).

Tabulka 1: Použité primery pro B chromozóm *S. purpureosericeum.*

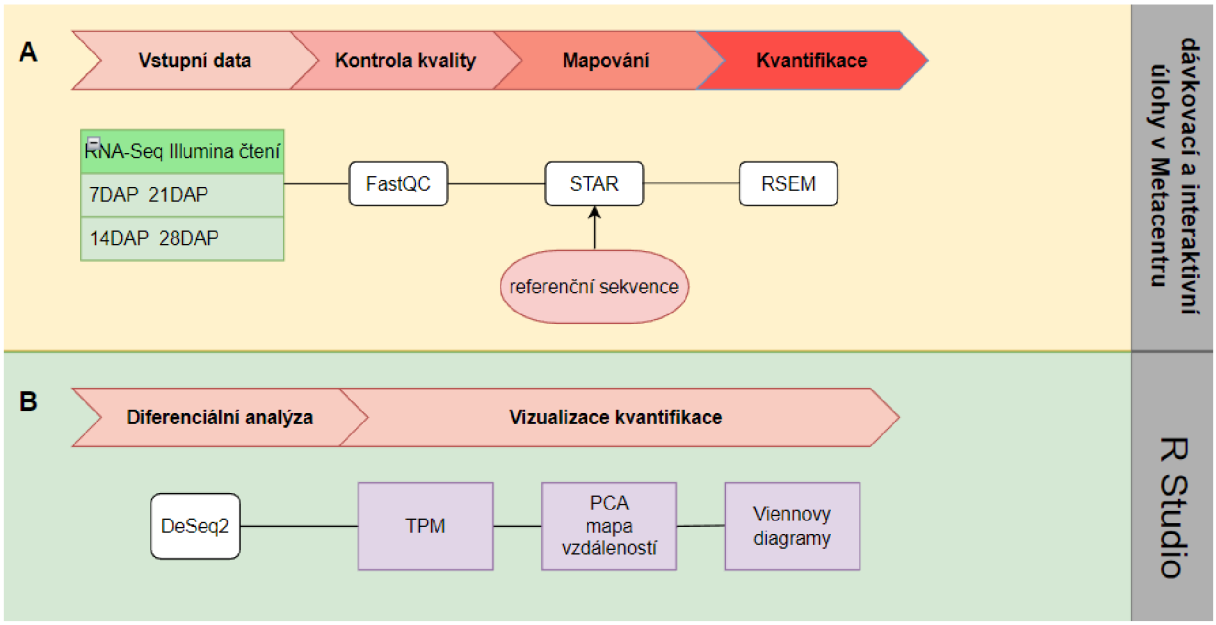
Označení	**Primer**	**Sekvence (5‘-3‘)**	**Velikost produktu (bp)**
24	utg312_gene155643_2Fa	AACAGCATTGCACACCGTTA	173
	utg312_gene155643_2Ra	GTTAGCGTTGCAACATGAGC	
27	utg5443_gene386712F	TTTCCAGCATAGACCCACT	193
	utg5443_gene386712R	TCCAAATAGCAGAGACAGGG	

4.5.2.4 Izolace RNA, kontrola kvality, sekvencování

Skladovaná embrya byla po zmražení v tekutém dusíku homogenizována tloučkem. Následovala izolace RNA pomocí kitu pro malé množství pletiva (Monarch® Total RNA Miniprep Kit) dle protokolu výrobce v sekci pro vzorky obtížně lyzovatelné. Eluce byla provedena do 30 µl elučního pufru. Kvalita a koncentrace byla stanovena elektroforeticky pomocí RNA 6000 Pico nebo Nano Kitu (Agilent) v automatizovaném přístroji 2100 Bioanalyzer Instrument (Agilent). Genomová knihovna byla vytvořena pomocí kitu pro malé množství RNA (NEBNext® Single Cell/Low Input RNA Library Prep Kit for Illumina®). Knihovny byly osekvenovány přístrojem Illumina NovaSeq (Illumina, Inc.).

4.5.2.5 Analýza RNA-Seq dat

Vstupními daty pro analýzu transkriptomu *S. purpureosericeum* byly Illumina NovaSeq párové čtení ve formátu fastq. Kvalita těchto hrubých dat ve formátu fastq byla zkontrolována pomocí nástroje fastQC v0.11.9. Hrubá data byla po kontrole kvality mapována nástrojem STAR na vytvořenou referenční sekvenci a překryté čtení byly následně kvantifikovány nástrojem RSEM (Obr. 13). Unixové příkazy pro nástroje STAR a RSEM jsou uvedeny v příloze 1.



Obrázek 13: Postup analýzy RNA-Seq dat. A – vstupní RNA seq data byla po kontrole kvality mapována na referenční sekvenci. Transkripty byly dále kvantifikovány. Výpočty probíhaly formou dávkovacích a interaktivních úloh v Metacentru. B – výstup z kvantifikace transkriptů byl podroben diferenciální analýze. Výsledky byly vizualizovány v prostředí R studia.

Výstupní soubory *isoforms.results nástroje RSEM byly importovány do R studia a následovala normalizace balíčkem Bioconductoru tximport (Soneson et al., 2015), který odhadované počty, délky a nadbytek transkritpů shrnuje do matic pro další použití v analýze genové exprese. Výstupem je TPM soubor, který byl využit pro explorativní analýzu jako je analýza hlavních komponent (PCA) a mapa vzdáleností podobnosti vzorků v programovacím jazyce R.

Pro diferenciální analýzu exprese byl použit balíček Bioconductoru DESeq2 (Love et al., 2014). Diferenciální analýza byla provedena vždy mezi 3 vzorky B+ a B0 embryí stejné kategorie stáří, přičemž byly stanoveny hranice pro geny se zvýšenou expresí na $\log_2 FC \geq 2$ (z anglického „$\log_2$ fold change“) a $p \leq 0.05$. Pro geny se sníženou expresí byly stanoveny hranice $\log_2 FC \leq -2$ a $p \leq 0.05$.

5 Výsledky

5.1 Referenční sekvence

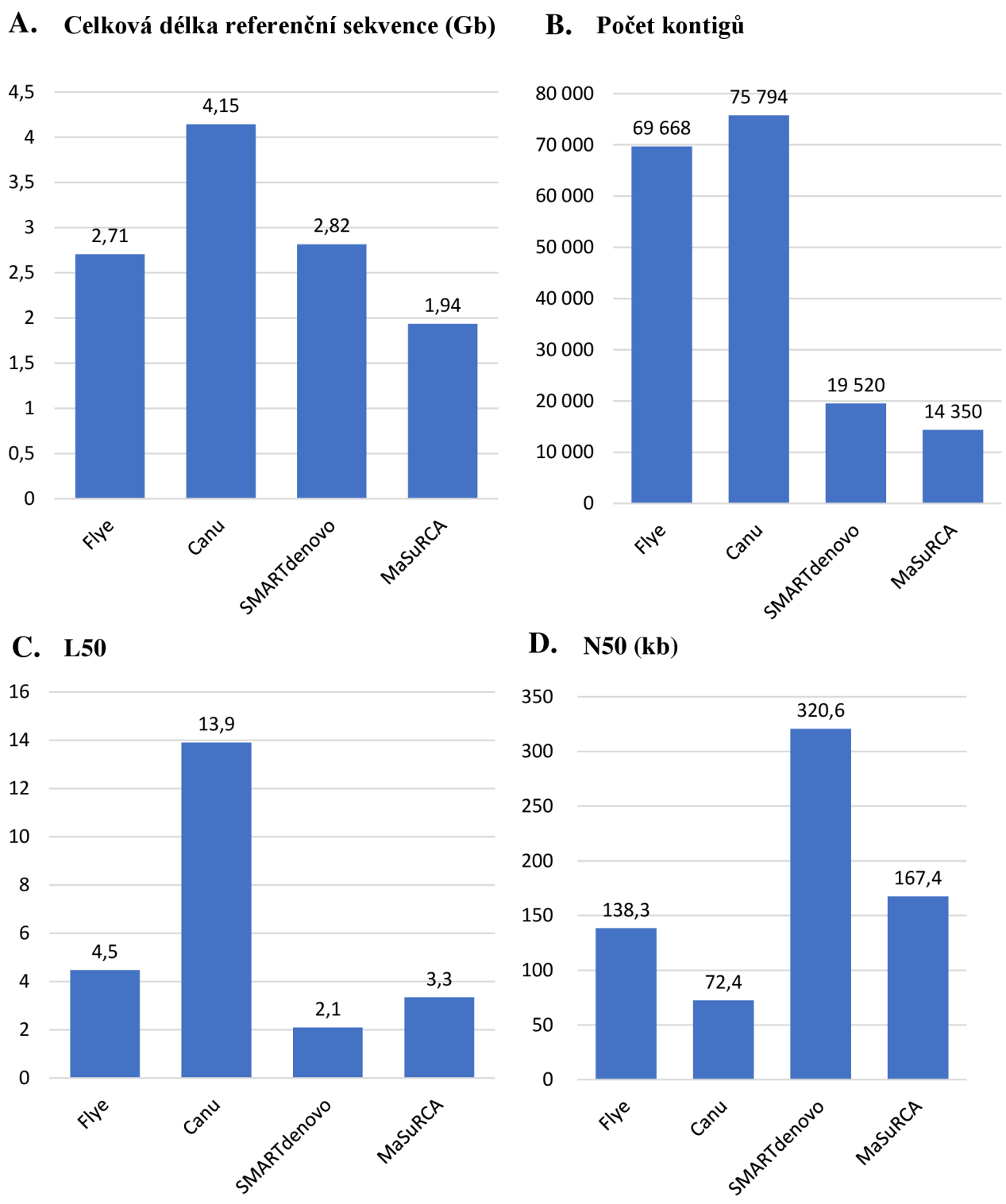
Pro sestavení referenční sekvence *de novo* byla použita ONT data dlouhého čtení a programy (assemblery) Canu, Flye, SMARTdenovo. Kombinace ONT dat dlouhého čtení a Illumina dat krátkého čtení byla použita pro hybridní assembler MaSuRCA.

Důležité parametry, které se sledují u vytvořených *de novo* genomů, je celková délka referenční sekvence, počet kontigů, N50 a L50 hodnoty. Výsledky statistik nástroje BioKIT s těmito základními parametry jsou zobrazeny na Obr. 14A-D. Nástroj Canu vytvořil nejdelší sestavení o velikosti přibližně 4,15 Gb a nástroj MaSuRCA nejkratšíverzi genomu o velikosti přibližně 1,94 Gb (Obr. 14A). Velikosti assembly vytvořených nástroji Flye a SMARTdenovo odpovídaly nejblíže předchozím odhadům velikosti genomu stanovených průtokovou cytometrií, s velikostmi 2,71 Gb a 2,82 Gb (Obr. 14A).

Nástroj Canu vyprodukoval 75 794 kontigů a nástroj Flye 69 668 kontigů (Obr. 14B). Genomová verze nástroje SMARTdenovo čítala 19 520 kontigů a nejmenší počet kontigů nástroj BioKIT spočítal v genomové verzi nástroje MaSuRCA, který čítal 14 350 kontigů (ObrXB).

Nejslibnější výsledek kontinuity genomu se ukázal u výstupu nástroje SMARTdenovo s hodnotami základních statistik L50 2,1 a N50 320,6 kb (Obr. 14C, 14D). Druhou nejlepší verzi genomu na základě kontinuity vyprodukoval nástroj MaSuRCA, který měl L50 3,3 a N50 167,4 kb (Obr. 14C, 14D). Podobný výsledek spočítal nástroj BioKIT u výstupního souboru nástroje Flye, kde bylo L50 stanoveno na 4,5 a N50 138,3 kb (Obr. 14C, 14D).

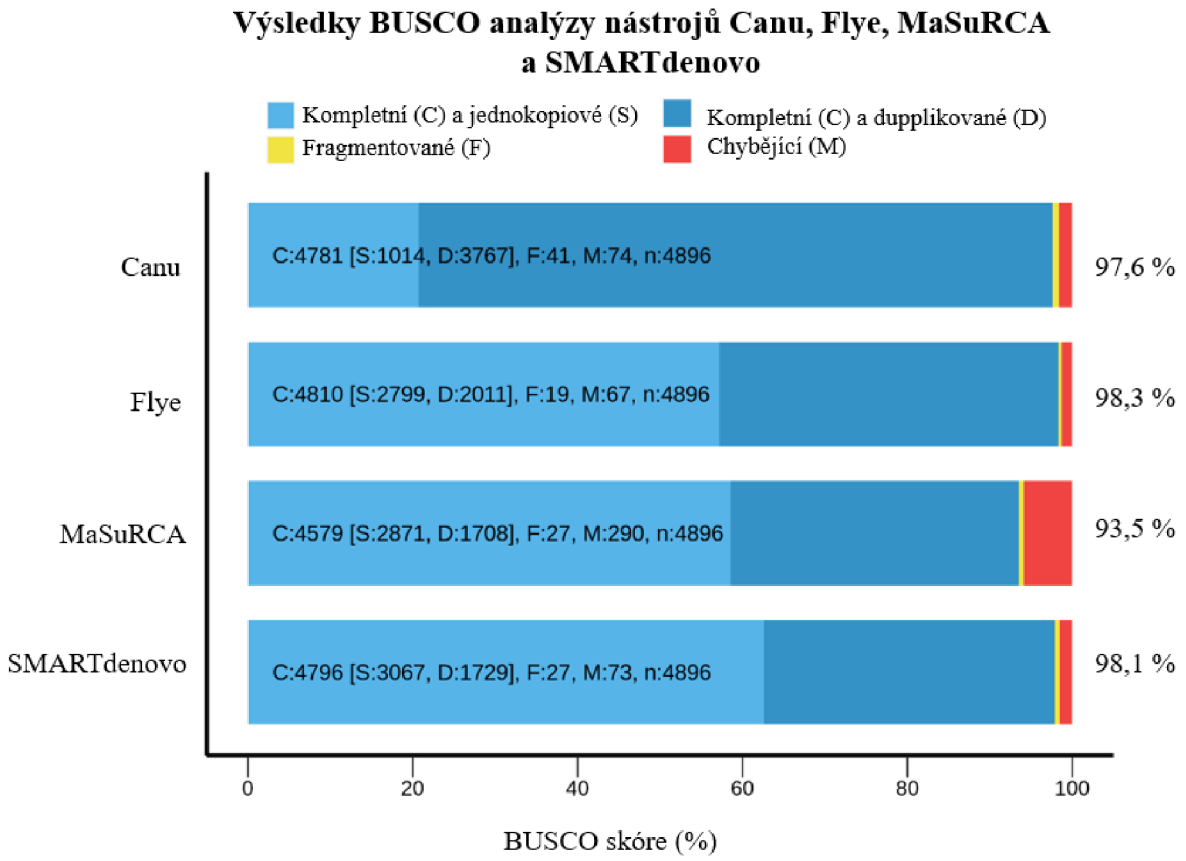
Jeden z metodických přístupů, který popisuje kvalitu sestavených referenčních sekvencí, je tzv. BUSCO analýza (Manni et al., 2021), která popisuje úplnost genomu na základě předpokládaného obsahu ortologních genů v dané taxonomické skupině. V tomto experimentu se pracovalo s taxonomickou skupinou lipnicotvarých *(Poales)*. Genomová verze assembleru Flye obsahovala 4810 kompletních genů z celkově 4 896 genů obsažených OrthoDB souboru pro taxon *Poales*, BUSCO skóre tedy činí 98,3 %, což byl nejlepší výsledek mezi čtyřmi assemblery (Obr. 15). Výstupní soubory SMARTdenovo a Canu obsahovalo 98,1 % a 97,6 % ortologním genů v uvedeném pořadí. Nejméně kompletních genů bylo BUSCO analýzou zjištěno ve výstupním souboru nástroje MaSuRCA, kde se nacházelo pouze 93,5 % genů (Obr. 15).



Obrázek 14: Výsledky nástroje BioKit pro Flye, Canu, SMARTdenovo a MaSuRCA. A – Celková délka referenční sekvence v gigabázích, B – Počet kontigů, C – L50, D – N50 v kilobázích.

Každá verze assembly jednotlivých programů obsahovala velký podíl duplikovaných genů oproti genům v jedné kopii. Nejvíce duplikovaných genů bylo vytvořeno nástrojem Canu, nejméně pak nástrojem MaSuRCA (Obr. 15). Pro výběr nejlepší genomové verze bylo nutné

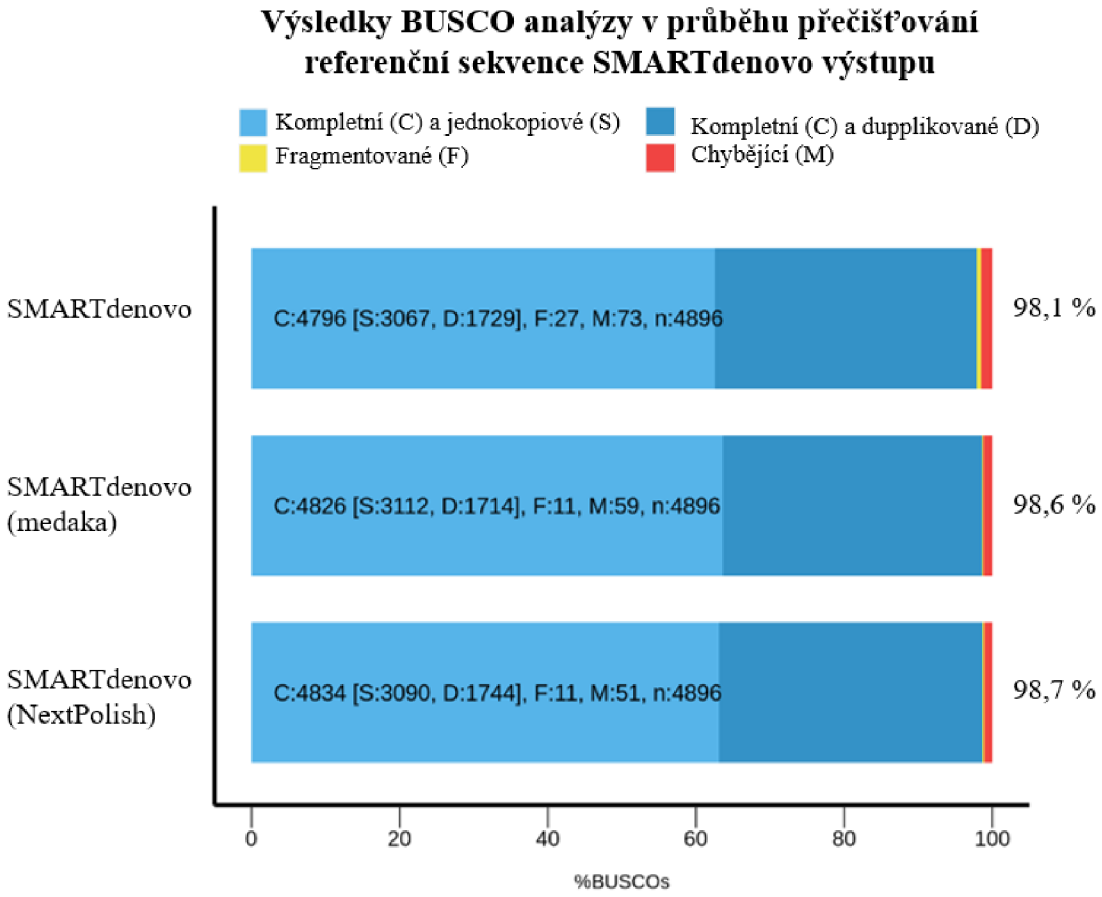
nalézt kompromis mezi vysokým procentuálním BUSCO skórem a co nejmenším počtem duplikovaných genů.



Obrázek 15: Grafické zobrazení BUSCO analýz výstupních souborů Canu, Flye, MaSuRCA a SMARTdenovo. Na ose x je uvedeno BUSCO skóre, které popisuje úplnost genomu na základě předpokládaného obsahu ortologních genů v taxonomické skupině *Poales*. Graf byl vytvořen pomocí skriptu generate_plot.py, který je součástí nástroje BUSCO.

Na základě výsledků analýz nástrojů BUSCO a BioKIT byla vybrána jako nejlepší verze genomu *S. purpureosericeum* výstupní soubor SMARTdenovo. Ten byl dále podroben korekci softwarem Medaka, který pro tento typ úloh využívá hrubých sekvenačních ONT dat dlouhého čtení. Výsledný fasta soubor byl opět podroben BUSCO analýze, ze kterých lze vyčíst, že se obsah kompletních genů zvýšil ze 4 796 na 4 826 (Obr. 16).

Výstup z nástroje Medaka byl dále přečištěn nástrojem NextPolish, který využívá hrubá Illumina data krátkého čtení. BUSCO analýza odhalila další navýšení kompletních genů na 4 834 kompletních genů. Přečištěná referenční sekvence má velikost 2,82 Gb a počet kontigů zůstává stále 19 520 (Tab. 1). Kontinuita se oproti referenci před přečištěním mírně zlepšila, N50 se navýšilo na 321,01 kb a L50 se snížilo na 2,09.



Obrázek 16: BUSCO analýza výstupních souborů nástroje SMARTdenovo a následně přečišťování daty dlouhého čtení (Medaka) a daty krátkého čtení (NextPolish). Graf byl vytvořen pomocí skriptu generate_plot.py, který je součástí nástroje BUSCO.

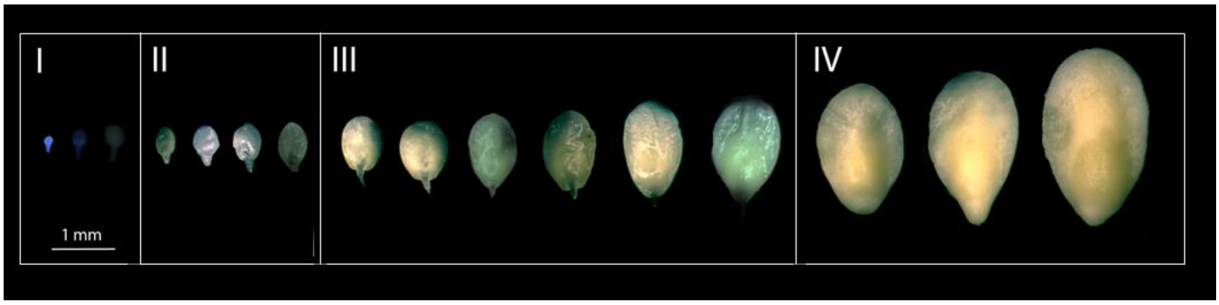
Tabulka 2: Finální parametry referenční sekvence.

	Finální referenční sekvence
Délka	2,82 Gb
L50	2,09
N50	321,01 kb
Počet kontigů	19 520

5.2 RNA-Seq analýza

5.2.1 Izolace embryí a endospermu

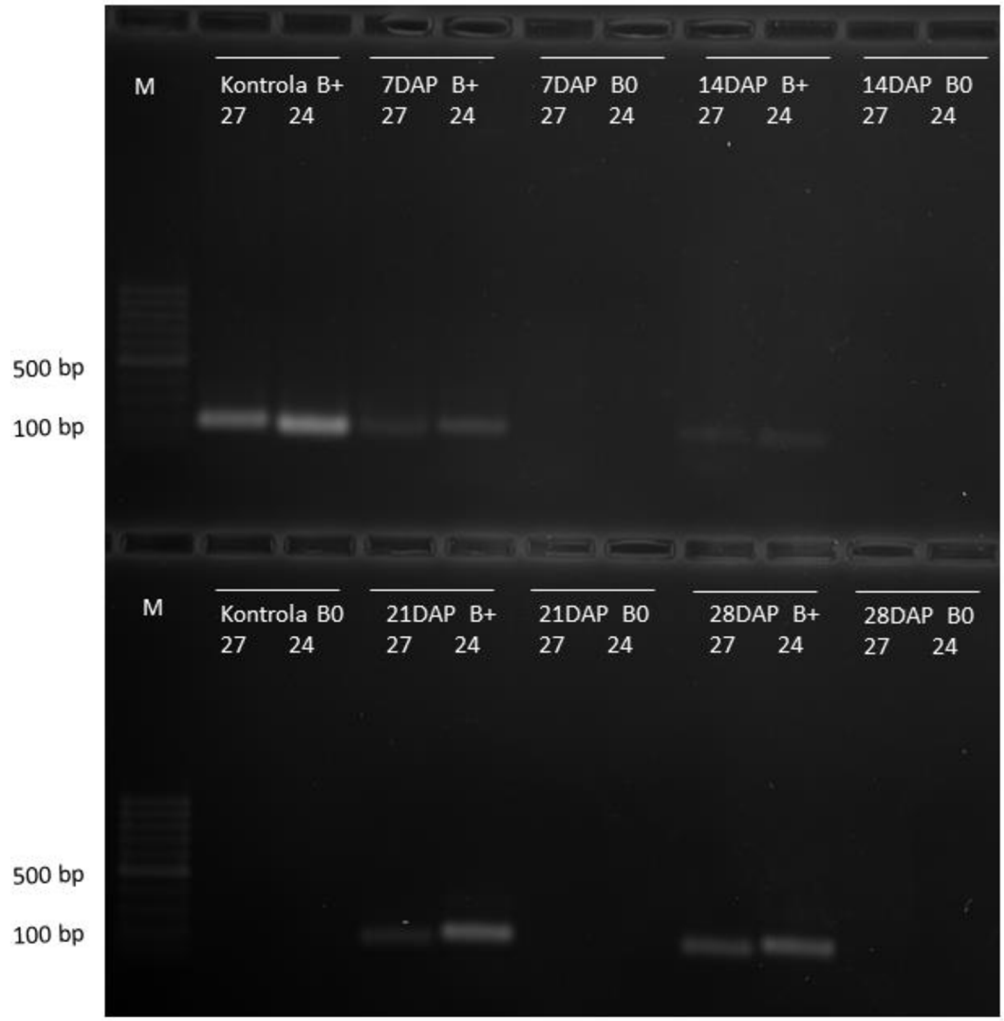
Ze zrajících semen čiroku *S. purpureosericeum* byla izolována embrya a jejich endospermy. Na základě velikosti byla embrya rozřazena do čtyř kategorií dle odhadovaného stáří ode dne opylení (DAP)(Obr. 17).



Obrázek 17: Vývojová řada embryí *S. purpureosericeum*. Členěno dle cílových skupin relevantních pro experimenty. 1. skupina – 7DAP, 2. skupina – 14DAP, 3. skupina – 21DAP, 4. skupina – 28 DAP Měřítko 1 mm.

5.2.2 Izolace DNA, screening z endospermů

Pro RNA-Seq experiment bylo vybráno 7 B+ a B0 izolovaných replik embryí a endospermů každé vývojové kategorie. Embryo i endosperm mají stejný B profil (viz Obr. 5C), čehož lze využít k určení B+ / B0 statusu embrya neinvazivní metodou. Zmražené endospermy byly využity k izolaci DNA endospermu. Koncentrace byla kontrolně změřena pomocí Nanodropu. B+ / B0 status endospermu (zároveň tedy i embrya) byl určen v PCR reakci na izolované DNA pomocí dvou specifických primerů. Výsledky PCR byly vyhodnoceny na základě výsledků elektroforézy (Obr. 18).



Obrázek 18: Gelová elektroforéza PCR produktů dvou markerů (24, 27) pro B chromozóm testovaných na endospermech B+ a B0 embryí. Jako pozitivní kontrola sloužila DNA izolovaná z prašníků a jako negativní kontrola DNA izolovaná z listu čiroku *S. purpureosericeum*. M – GeneRuler 100 bp DNA Ladder.

5.2.3 Izolace RNA, agilent

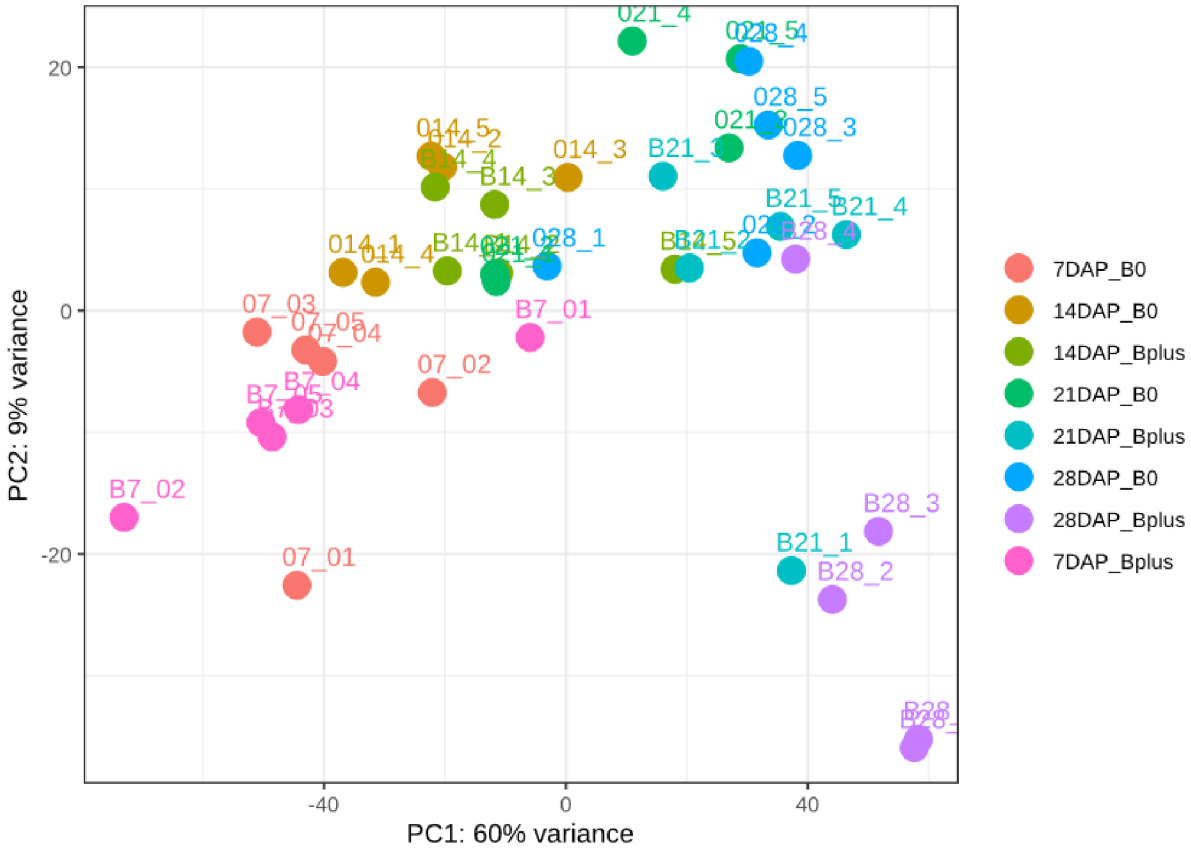
RNA byla izolována z 6 z celkových 7 odebraných kopií B0 i B+ embryí každé vývojové kategorie. Kvalita a koncentrace RNA byla změřena zařízením 2100 Bioanalyzer Instrument (Agilent) (Příloha 2). Koncentrace izolované RNA 7DAP a 14DAP vzorků se pohybovala v řádu stovek pikogramů na mikrolitr, zatímco 21DAP a 28DAP vzorků v řádu desítek nanogramů na mikrolitr (Příloha 2). 5 nejlepších vzorků každé kategorie bylo na základě změřené koncentrace vybráno pro sekvencování metodou Illumina.

5.2.4 Analýza RNA-Seq dat

Illumina sekvencování vygenerovalo celkem 1,86 miliard párových čtení (průměrně 30 milionů čtení na vzorek) (Příloha 2). Hrubá data byla namapována programem STAR na

genom sestavený pomocí SMARTdenovo nástroje. V příloze 2 je uvedeno množství unikátně namapovaných čtení na referenční sekvenci.

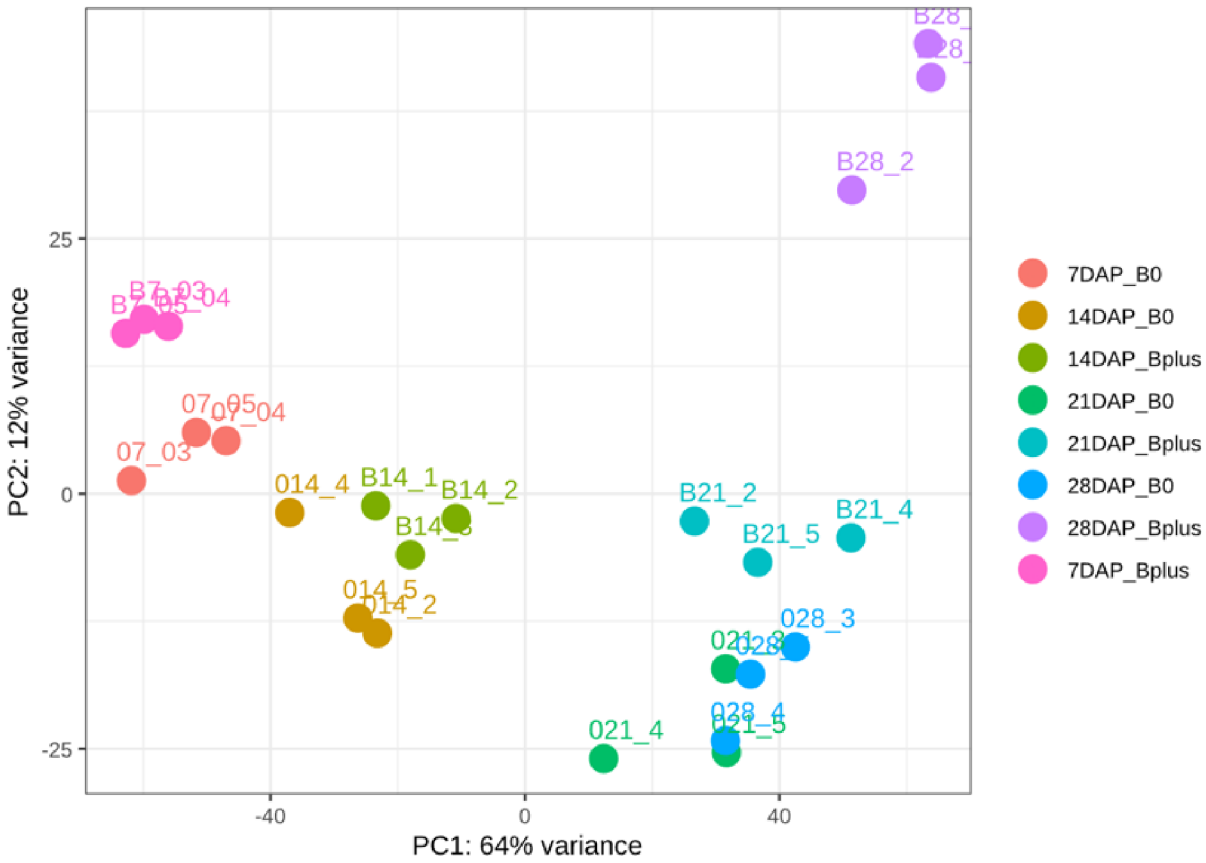
V programovacím jazyce R byla dále provedena vizualizace normalizovaných dat pomocí PCA analýzy (Obr. 19). PCA analýza je statistická metoda sloužící k určení hlavních faktorů ovlivňujících dataset RNA-Seq dat. Výsledky shlukování replik v PCA analýze (Obr. 19) odhalily nekvalitní repliky, které byly zařazeny do špatné kategorie stáří (např. replika B21_1). Z každé B+ i B0 kategorie byly pro následující analýzy zachovány 3 nejvíce konzistentní repliky na základě kvality čtení (Příloha 2) a shlukování v PCA prostoru.



Obrázek 19: PCA analýza 5 B+ a B0 replik 7DAP, 14DAP, 21DAP a 28DAP embryí. Na ose x je znázorněna PC1 variance, na ose y PC2 variance. Barevná legenda znázorňuje jednotlivé kategorie stáří.

Z grafu vygenerovaného PCA analýzou s 3 replikami je patrné, že vzorky sedmidenních embryí obsahující B chromozóm (B7) a vzorky sedmidenních embryí bez B chromozómů (B0) jsou v PCA prostoru jasně odděleny (Obr. 20). Skupiny B14 a 014 jsou od sedmidenních skupin posunuty v rámci PC1. Replika 014_4 do shluku nezapadá, ale nebylo možné ji nahradit jinou replikou z důvodu nekvalitních dat. Vzorky skupin B21, 021 a 028 jsou částečně překryté a nelze je jasně rozlišit, což může naznačovat podobnou genovou expresi či

již nevýrazné změny vývoje embrya mezi třetím a čtvrtým týdnem po opylení. Skupina dvaceti osmi denních embryí (B28) se od ostatních embryí podobného stáří výrazně odlišuje v rámci PC2 (Obr. 20). Toto shlukování replik B28 vzorků naznačuje největší odlišnost genové exprese v rámci sledovaných skupin. Z Obr. 20 je patrné, že zásadní roli v rozdílech mezi vzorky hraje stáří jednotlivých embryí. Nicméně je důležité poznamenat, že PCA analýza není přímým indikátorem genové exprese.

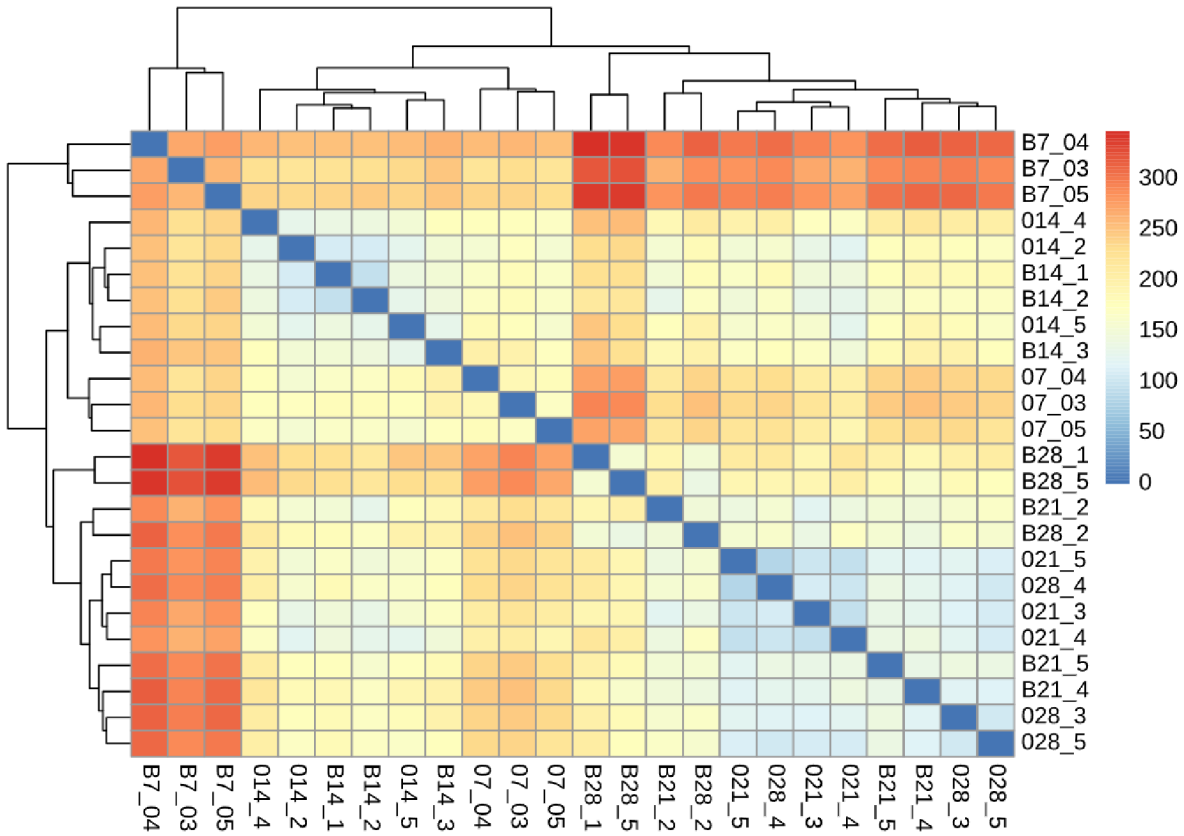


Obrázek 20: PCA analýza normalizovaných RNA-Seq dat po vyběru 3 nejvíce konzistentních replik. Vzorky s B chromozómy se shlukují vždy posunuty nahoru v rámci PC2. PC1 pravděpodobně znázorňuje vývojové kategorie embryí – nalevo se vyskytují nejmladší vzorky a napravo nejstarší. Nejvíce odlišná je skupina B28 (fialová). Barevná legenda znázorňuje jednotlivé kategorie stáří.

Po normalizaci RNA-Seq dat byla vytvořena také mapa vzdáleností vzorků („sample-to-sample distance heatmap“) (Obr. 21), která zobrazuje vzdálenosti mezi jednotlivými vzorky na základě genové exprese. Mapa vzdáleností vyjadřuje míru odlišnosti mezi vzorky pomocí barevného kódování. Čím sytější červená barva, tím větší odlišnost mezi vzorky a naopak modrá barva značí největší podobnost. Na diagonále grafu jsou modré buňky, které znázorňují porovnání stejných vzorků.

Z grafu lze jasně vidět dva sytě červené shluky, které ukazují, že tři vzorky skupiny sedmidenních embryí s B chromozómem (B7) jsou nejvíce odlišné od dvou replik skupiny nejstarších embryí s B chromozómem B28.

Některé vzorky vykazovaly větší podobnost s jinou skupinou než se svou vlastní skupinou. Tyto vzorky mohou být pro další analýzu problematické a mohou způsobit nepřenosti ve výsledcích a jejich interpretaci.

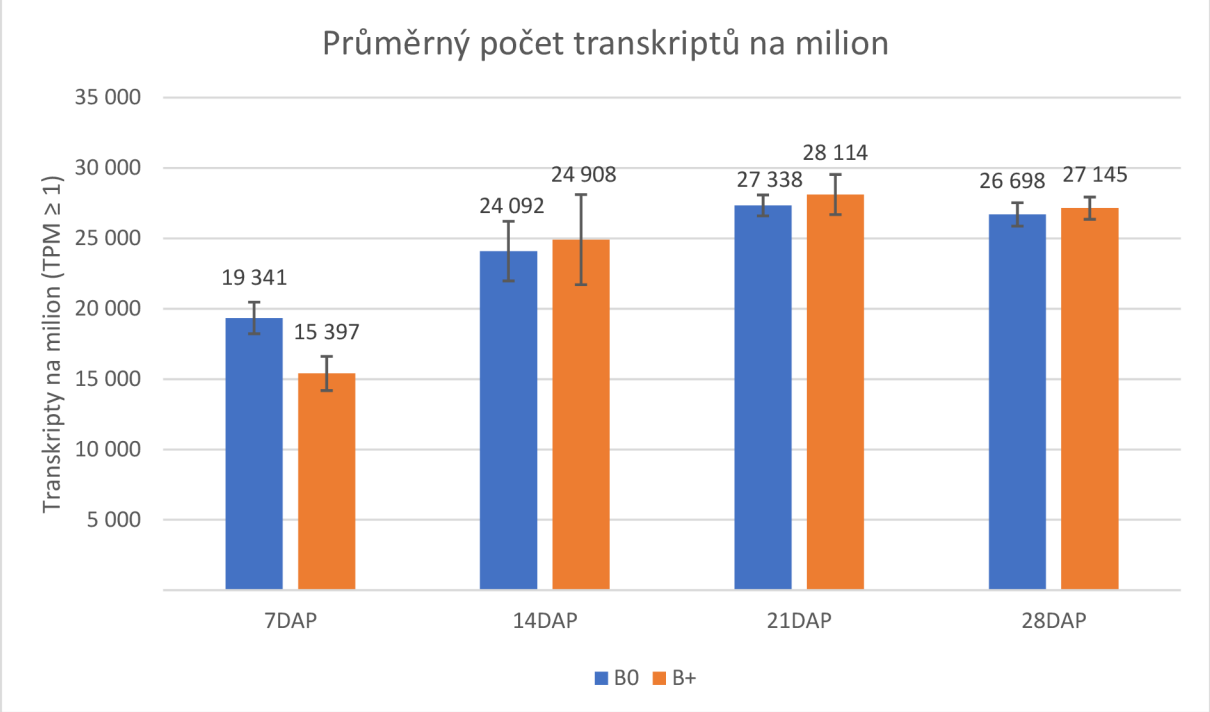


Obrázek 21: Vizualizace vzdáleností mezi RNA-Seq vzorky pomocí "sample-to-sample“ mapy vzdáleností. Nejvíce odlišné shluky (červená barva) jsou repliky skupiny nejmladších B7 embryí od replik skupiny nejstarších embryí B28.

Další informace o kvalitě dat a vztazích mezi kategoriemi stáří je možné také vyčíst z grafu průměrného počtu transkriptů na milion (TPM) (Obr. 22). Hranice pro všechny repliky byla stanovena na TPM ≥ 1. Z grafu lze vyčíst, že se počet genů exprimovaných nad touto úrovní pro B0 i B+ skupiny zvyšuje s rostoucím stářím embrya – pro 7DAP embrya je počet genů exprimovaných nad stanovenou hodnotu TPM až 19 431 transkriptů, pro 14DAP až 24 908, pro 21DAP až 28 114 (Obr. 22). Pro skupinu 28DAP je počet exprimovaných genů za podmínky TPM ≥ 1 o tisíc menší než u 21DAP skupiny, což však odpovídá již předchozím

analýzám, kde se skupiny 21DAP a 28DAP embryí částečně překrývají a z hlediska embryonálního vývoje pravděpodobně nejsou tak odlišná.

Výjimkou z pravidla většího počtu exprimovaných genů v B+ vzorcích je skupina 7DAP embryí, kde je počet genů s TPM ≥ 1 až o 4 tisíce nižší než B0 skupiny. To samotné však nemusí znamenat, že se transkripce B+ od B0 skupiny natolik liší, protože průběhu izolace RNA mohlo dojít ke ztrátě některých transkriptů kvůli nízké koncentraci izolované RNA.

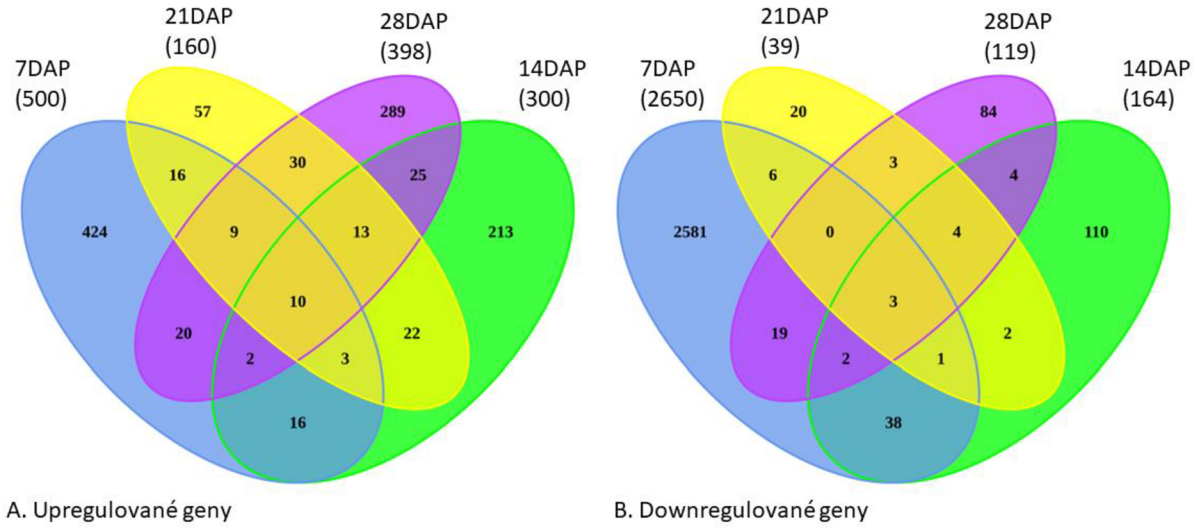


Obrázek 22: Průměrný počet transkriptů (TPM ≥ 1). Osa x znázorňuje kategorie stáří embryí a osa y průměrný počet transkriptů 3 replik v každé B0 (modře) i B+ (červeně) vývojové kategorii embrya.

5.2.4.1 Diferenciální analýza exprese

Analýza diferenciální exprese byla provedena v R skriptu implementací Deseq2 algoritmu. Bylo zjištěno, že v porovnání se vzorky bez B chromozómů došlo u B+ vzorků embryí k výraznému nárůstu exprese 500 genů v 7. den po oplození (7DAP), 300 genů v 14DAP, 160 genů ve 21DAP a 398 genů v 28DAP (Obr. 23A) ($\log_2$FC ≥ 2, p-hodnota < 0.05).

Dataset byl v experimentu diferenciální exprese prohledáván i pro geny se sníženou expresí ($\log_2$FC ≥ –2, p-value < 0.05). Nejvíce genů se sníženou expresí bylo nalezeno u 7DAP skupiny, kde počet takových genů činil 2650 v porovnání B+ proti B0 skupině. 164 genů se sníženou expresí bylo nalezeno v 14DAP B+ embryích, 39 genů ve 21DAP a 119 genů v 28DAP embryích oproti vzorkům bez B chromozómů (Obr. 23B).



Obrázek 23: Viennovy diagramy diferenciální exprese genů B+ embryo všech věkových skupin ($log_2FC \geq \pm 2$, p-value < 0.05). A – geny se zvýšenou expresí, B – geny se sníženou expresí. V závorce za názvy skupin je uveden počet celkových genů se zvýšenou či sníženou expresí, v barevných polích poté na průnicích počty sdílených genů různých skupin (7DAP – modrá, 14DAP – zelená, 21DAP – žlutá, 28DAP – fialová).

6 Diskuze

Divoký druh čiroku *Sorghum purpureosericeum* může ve svém genomu obsahovat nad rámec své obvyklé sady B chromozómy. Tyto B chromozómy se vyskytují stabilně pouze v květenství a z ostatních pletiv jsou eliminovány (Darlington et al., 1941). K eliminaci dochází pravděpodobně již v raných stádiích embrya, podobně jako u *Aegilops speltoides*, kde byla eliminace B chromozómů z kořenů zaznamenána již v 6–8 denním embryu (Ruban et al., 2020). Cílem této diplomové práce bylo vytvoření první referenční sekvence čiroku *Sorghum purpureosericeum* a analýza RNA-Seq dat embryí *S. purpureosericeum* s B chromozómem i bez B chromozómu v průběhu vývoje (7DAP, 14DAP, 21DAP a 28DAP) za účelem zisku diferenciálně exprimovaných genů, které by mohly být zodpovědné za eliminaci B chromozómu v pletivech.

Pro vytvoření referenční sekvence byly zvoleny čtyři nástroje – Flye, Canu, MaSuRCA a SMARTdenovo. Bylo provedeno vyhodnocení kvality výstupních souborů z těchto nástrojů na základě BUSCO skóre a základních parametrů (délka celkové referenční sekvence, počet kontigů, N50 a L50), které sloužilo k porovnání vzniklých sekvencí a k výběru nejlepší verze referenční sekvence. V souvislosti s výsledky porovnávaní kvality vzniklých assembly, byla pro další experimenty zvolena verze vytvořená programem SMARTdenovo. Výstup ze SMARTdenovo byl ještě očištěn nástrojem Medaka (dlouhé čtení) a NextPolish (krátké čtení). SMARTdenovo nástroj byl úspěšně použit ve studii, která sestavila sekvenci genomu čiroku *Sorghum bicolor* Tx430 z dlouhých čtení ONT MiniON platformy (Deschamps et al., 2018) a dalších rostlinných genomů (Lin et al., 2018) (Schmidt et al., 2017).

Výsledná referenční sekvence *S. purpureosericeum* nabývá délky 2,82 Gb. Tato délka je větší než reálná velikost genomu *S. purpureosericeum,* která byla již dříve průtokovou cytometrií stanovena na 2.21 Gb (Karafiátová et al., 2021). Problémem ve vytvořené referenci jsou duplikované kontigy, kterých výsledná verze vytvořená v této práci obsahuje přibližně třetinu. Duplikáty vznikaly v heterozygotických regionech, kde bioinformatické nástroje vytvořily namísto jedné kopie genu dvě kopie odpovídající mateřské a otcovské kopii daného lokusu. Tyto duplikáty narušují kontinuitu sekvence a vytváří nepřesnosti v transkriptomické analýze. Pro další vylepšení referenční sekvence v následujících verzích je nutno duplikované geny (také „haplotigy“) odstranit pomocí nástrojů jako je purge_haplotigs (Roach et al., 2018) nebo purge_dups (Guan et al., 2020). V této práci prozatím nebyly duplikáty odstraněny z důvodu možné ztráty genů nacházejících se na B chromozómu.

SMARTdenovo výstup má 19 520 kontigů, což je také stále velké číslo, a bylo by vhodné využít dalších přístupů pro sestavení kontigů do superkontigů (tzv. „scaffoldy“) o délce samotných chromozómů. Takovými přístupy může být například kombinace NGS/TGS a Hi-C techniky, což bylo úspěšně provedeno u ječmene s finálním počtem 6 347 superkontigů (Mascher et al., 2017). Další možností vylepšení kontinuity sekvence je použití metody fyzikálních map (optická mapa, restrikční mapa). Tato metoda byla použita při sestavení verze genomu *Arabidopsis thaliana* (Michael et al., 2018). Výsledná sekvence genomu získaná tímto způsobem se skládala z 62 kontigů.

Na referenci byla úspěšně mapována transkriptomická data. Shluková PCA analýza rozdělila skupiny embryí do PCA prostoru na základě PC1 variance (Obr. 20, osa x) a PC2 variance (Obr. 20 osa y). PC1 64% variance jasně odpovídá stáří, kde nejmladší vzorky se nachází vlevo a nejstarší vpravo. Vzorky s B chromozómy se nacházely v PCA prostoru vždy nad B chromozómy. Rozdíl v PC2 mezi B+ a B0 vzorky byl u 7DAP, 14DAP a 21DAP přibližně stejný, zato u 28DAP embryí byl zaznamenán největší rozdíl mezi B+ a B0 vzorky. Tento vysoký rozdíl v PC2 u nejstarších embryí může znamenat vysoký transkriptomický vliv B chromozómů. Dle PCA analýzy mělo větší efekt na transkriptom embryí *S. purpureosericeum* jasně jejich vývojové stáří, a ne přítomnost B chromozómu.

Počet transkriptů se v průběhu vývoje zvyšoval a obecně platilo, že vzorky s B chromozómy měly větší počet transkriptů (Obr. 22). Tomuto trendu neodpovídá statistika nejmladších embryí s B chromozómy, jejichž TPM bylo nižší než B0 vzorky. Je třeba brát v potaz, že izolace RNA je ztrátová, a jisté množství transkriptů se v průběhu zpracování vzorků ztrácí. Obezřetně je nutno přistupovat k získanému počtu transkriptů zejména u sedmidenních embryí, které jsou velmi malé, je s nimi obtížná manipulace a ztráty RNA budou znatelné.

Analýza diferenciální exprese B+ embryí oproti B0 vzorkům odhalila 500 genů se zvýšenou expresí v 7DAP embryích, 160 genů v 14DAP embryích, 300 v 21DAP embryích a 398 v 28DAP embryích. Pomocí průniků všemi kategoriemi (Obr. 23A) bylo vyfiltrováno 424 unikátních genů, které jsou nadexprimovány pouze v sedmidenních embryích a které by potenciálně mohly obsahovat geny zodpovědné za eliminaci B chromozómů. Obdobně tomu bylo u genů se sníženou expresí (Obr. 23B), kde bylo zachyceno až 2 650 genů u 7DAP embryí. Nicméně je nutno brát v potaz předchozí výsledky, ze kterých je pravděpodobné, že transkriptomy těchto nejmladších embryí nejsou kompletní, a tudíž může být analýza diferenciální exprese zkreslená. Další postup v RNA-Seq analýze, která již není v této práci zahrnuta, je anotace kandidátních transkriptů. Nejužívanějším nástrojem pro anotaci je

Blast2GO, který rozpoznává geny na základě podobností v sekvencích uložených v mnoha databázích (Conesa et al., 2005).

Již zmíněná RNA-Seq studie eliminace B chromozómů v kořenech *Aegilops speltoides* se zabývala vzorky mezi 17–20 dny embryogeneze a nalezla 341 diferenciálně exprimovaných B+ genů (Boudichevskaia et al., 2020), zatímco v předkládané práci bylo nalezeno u 21DAP embryí pouze 199 diferenciálně exprimovaných genů. Transkriptomická studie *Aegilops speltoides* dále anotovala diferenciálně exprimované geny. Nejvíce zastoupenými pojmy v kategorii „molekulární procesy“ byly „buněčné procesy“ a „metabolické procesy dusíkatých sloučenin“. Dále také „genová exprese“, „procesy související s mikrotubuly“, „pohyb založený na mikrotubulech“, „buněčný nebo organelový pohyb“ a „organelová fúze“. V kategorii molekulární funkce byly nejvíce zastoupeny pojmy „vazba GTP“, „vazba cytoskeletárních proteinů“, „vazba mikrotubulů“, „aktivita mikrotubulových motorů“ a „vazba tubulinu“ (Boudichevskaia et al., 2020). Průnik výsledků anotace transkriptů *S. purpureosericeum* s *A. speltoides* by mohl poukázat na podobné mechanismy, které jsou zodpovědné za eliminaci B chromozómů v obou rostlinách.

7 Závěr

Diplomová práce se zabývala studiem transkriptomu v průběhu vývoje embryí divokého čiroku *S. purpureosericeum* za účelem identifikace skupiny kandidátních genů, které jsou zodpovědné za eliminaci B chromozómu v somatických pletivech.

Pro vytvoření referenční sekvence byly použity data krátkého i dlouhého čtení a nástroje Flye, Canu, MaSuRCA a SMARTdenovo. Nejlepší výstupní soubor poskytl nástroj SMARTdenovo, který byl dále hrubými daty přečišťován. Finální referenční sekvence nabývá velikosti 2,82 Gb a kompletnosti BUSCO skóre 98,7 %.

Následná analýza diferenciální exprese stanovila potenciálně zajímavé geny se zvýšenou expresí 7DAP embryí, u kterých je pravděpodobně efekt eliminace B chromozómu nejsilnější, a které by mohly být zapojeny v procesu eliminace B chromozómu u studovaného druhu *S. purpureosericeum.*

8 Bibliografie

Ahmad, S., & Martins, C. (2019). The Modern View of B Chromosomes Under the Impact of High Scale Omics Analyses. *Cells*, *8*(2). https://doi.org/10.3390/cells8020156

Aldrich, J., Leibholz, A., Cheema, M., Ausió, J., & Ferree, P. (2017). A 'selfish' B chromosome induces genome elimination by disrupting the histone code in the jewel wasp Nasonia vitripennis. *Scientific Reports*, *7*(1). https://doi.org/10.1038/srep42551

Amorim, I., Milani, D., Cabral-de-Mello, D., Rocha, M., Moura, R., & Puertas, M. (2016). Possible origin of B chromosome in Dichotomius sericeus (Coleoptera). *Genome*, *59*(8), 575-580. https://doi.org/10.1139/gen-2016-0048

Anaconda Software Distribution: Computer software. Vers. 2-2.4.0. Anaconda. (2016). Retrieved 2023-03-04, from <https://anaconda.com>

Ananda, G., Myrans, H., Norton, S., Gleadow, R., Furtado, A., & Henry, R. (2020). Wild Sorghum as a Promising Resource for Crop Improvement. *Frontiers in Plant Science*, *11*(1108). https://doi.org/10.3389/fpls.2020.01108

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10). https://doi.org/10.1186/gb-2010-11-10-r106

Anders, S., Pyl, P., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166-169. https://doi.org/10.1093/bioinformatics/btu638

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Babraham Bioinformatics. Retrieved 2023-02-18, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bajpai, B. (2014). High Capacity Vectors. *Advances in Biotechnology*, 1-10. https://doi.org/10.1007/978-81-322-1554-7_1

Banaei-Moghaddam, A., Meier, K., Karimi-Ashtiyani, R., & Houben, A. (2013). Formation and Expression of Pseudogenes on the B Chromosome of Rye. *The Plant Cell*, *25*(7), 2536–2544. https://doi.org/10.1105/tpc.113.111856

Barnaud, A., Deu, M., Garine, E., Chantereau, J., Bolteu, J., Koïda, E., McKey, D., & Joly, H. (2009). A weed-crop complex in sorghum: The dynamics of genetic diversity in a traditional farming system. *American Journal of Botany*, *96*(10), 1869-1879. https://doi.org/10.3732/ajb.0800284

Bennett, S. (2004). Solexa Ltd. *Pharmacogenomics*, *5*(4), 433-438. https://doi.org/10.1517/14622416.5.4.433

Birchler, J. (1991). Chromosome manipulations in maize. In P. Gupta & T. Tsuchiya (eds.), *Chromosome engineering in plants: genetics, breeding, evolution* (pp. 531–559). Elsevier.

Blavet, N., Yang, H., Su, H., Solanský, P., Douglas, R., Karafiátová, M., Šimková, L., Zhang, J., Liu, Y., Hou, J., Shi, X., Chen, C., El-Walid, M., McCaw, M., Albert, P., Gao, Z., Zhao, C., & Ben-Zvi, G. (2021). *Proceedings of the National Academy of Sciences*, *118*(23). https://doi.org/10.1073/pnas.2104254118

Bolger, A., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Boudichevskaia, A., Fiebig, A., Kumke, K., Himmelbach, A., & Houben, A. (2022). Rye B chromosomes differently influence the expression of A chromosome–encoded genes depending on the host species. *Chromosome Research*, *30*(4), 335-349. https://doi.org/10.1007/s10577-022-09704-6

Boudichevskaia, A., Ruban, A., Thiel, J., Fiebig, A., & Houben, A. (2020). Tissue-Specific Transcriptome Analysis Reveals Candidate Transcripts Associated with the Process of

Programmed B Chromosome Elimination in Aegilops speltoides. *International Journal of Molecular Sciences*, *21*(20). https://doi.org/10.3390/ijms21207596

Bramel-Cox, T. (1988). Use of wild sorghums in sorghum improvement.

Bugrov, A., Karamysheva, T., Perepelov, E., Elisaphenko, E., Rubtsov, D., Warchałowska-Śliwa, E., Tatsuta, H., & Rubtsov, N. (2007). DNA content of the B chromosomes in grasshopper Podisma kanoi Storozh. (Orthoptera, Acrididae). *Chromosome Research*, *15*, 315–325. https://doi.org/10.1007/s10577-007-1128-z

Camacho, J. (ed.). (2005). B chromosomes. In T. Gregory, *The Evolution of the Genome* (pp. 223–286). Elsevier.

Carlson, W. (1969). Factors Affecting Preferential Fertilization in Maize. *Genetics*, *62*(3), 543-554. https://doi.org/10.1093/genetics/62.3.543

Conesa, A., Gotz, S., Garcia-Gomez, J., Terol, J., Talon, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674-3676. https://doi.org/10.1093/bioinformatics/bti610

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M., Gaffney, D., Elo, L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-0881-8

Corchete, L., Rojas, E., Alonso-López, D., De Las Rivas, J., Gutiérrez, N., & Burguillo, F. (2020). Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-76881-x

Dai, M., Thompson, R., Maher, C., Contreras-Galindo, R., Kaplan, M., Markovitz, D., Omenn, G., & Meng, F. (2010). NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics*, *11*(4). https://doi.org/10.1186/1471-2164-11-S4-S7

Damania, A. (2002). The Hindustan centre of origin of important plants. *Asian Agri-history*, *6*(4), 333–341.

D'Ambrosio, U., Pilar Alonso-Lifante, M., Barros, K., Kovařík, A., Mas de Xaxars, G., & Sònia Garcia, G. (2017). B-chrom: a database on B-chromosomes of plants, animals and fungi. *New Phytologist*, *216*, 635-642. https://doi.org/10.1111/nph.14723

Darlington, C., Thomas, T., & Thomas, P. (1941). Morbid mitosis and the activity of inert chromosomes in Sorghum. *Proceedings of the Royal Society B: Biological Sciences*, *130*, 127–150.

de Morais Cardoso, L., Pinheiro, S., Duarte Martino, H., & Pinheiro-Sant'Ana, H. (2017). Sorghum (Sorghum bicolor L.): Nutrients, bioactive compounds, and potential impact on human health. *Critical Reviews in Food Science and Nutrition*, *57*(2), 372-390. https://doi.org/10.1080/10408398.2014.887057

Del Fabbro, C., Scalabrin, S., Morgante, M., Giorgi, F., & Seo, J. (2013). An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLoS ONE*, *8*(12). https://doi.org/10.1371/journal.pone.0085024

Denton, J., Lugo-Martinez, J., Tucker, A., Schrider, D., Warren, W., Hahn, M., & Guigo, R. (2014). Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLoS Computational Biology*, *10*(12). https://doi.org/10.1371/journal.pcbi.1003998

Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., May, G., & Lin, H. (2018). A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-07271-1

Dobin, A., Davis, C., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21. https://doi.org/10.1093/bioinformatics/bts635

Doležel, J., Vrána, J., Cápal, P., Kubaláková, M., Burešová, V., & Šimková, H. (2014). Advances in plant chromosome genomics. *Biotechnology Advances*, *32*(1), 122-136. https://doi.org/10.1016/j.biotechadv.2013.12.011

Douglas, & Birchler, R. (2017). In Chromosome Structure and Aberrations. In *Chromosomes* (1st edition, pp. 13–39.). Springer.

Evans, G., Rees, H., Snell, C., & Sun, S. (1972). The relationship between nuclear DNA amount and the duration of the mitotic cycle. *Chromosomes Today*, *3*, 24–31.

FAO, 2009. Food and Agriculture Organization of the United Nations(FAO), State of Food Insecurity in the World 2009, Rome.

Ghurye, J., Pop, M., & Segata, N. (2019). Modern technologies and algorithms for scaffolding assembled genomes. *PLOS Computational Biology*, *15*(6). https://doi.org/10.1371/journal.pcbi.1006994

Gotoh, K. (1924). Uber die chromosomenzahl von Secale cereale L. *Botanical Magazine*, *38*, 135–152.

Grabherr, M., Haas, B., Yassour, M., Levin, J., Thompson, D., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B., Nusbaum, C., Lindblad-Toh, K. et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644-652. https://doi.org/10.1038/nbt.1883

Graphodatsky, A., Kukekova, A., Yudkin, D., Trifonov, V., Vorobieva, N., Beklemisheva, V., Perelman, P., Graphodatskaya, D., Trut, L., Yang, F., Ferguson-Smith, M., Acland, G., & Aguirre, G. (2005). The proto-oncogene C-KIT maps to canid B-chromosomes. *Chromosome Research*, *13*(2), 113-122. https://doi.org/10.1007/s10577-005-7474-9

Gregory, T. (2005). The C-value Enigma in Plants and Animals: A Review of Parallels and an Appeal for Partnership. *Annals of Botany*, *95*(1), 133-146. https://doi.org/10.1093/aob/mci009

Guan, D., McCarthy, S., Wood, J., Howe, K., Wang, Y., Durbin, R., & Valencia, A. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896-2898. https://doi.org/10.1093/bioinformatics/btaa025

Harlan, J. (1992). Indigenous African agriculture", in Crops Man. Ed. Harlan, J. R. (Wisconsin, USA: American Society of Agronomy) doi: 10.2135/1992.cropsandman.c9. In *Crops Man)* (., pp. 177-191). American Society of Agronomy.

Harnádková, N. (2022). *Odvodenie PCR markerov špecifických pre B chromozóm ciroku Sorghum purpureo-sericeum a ich využitie pre selekciu B-pozitívnych jedincov* [bakalářská práce]. Univerzita Palackého v Olomouci.

Hasegawa, N. (1934). A cytological study on 8-chromosome rye. *Cytologia*, *6*(1), 68–77.

Henriques-Gil, N., Santos, J., & Arana, P. (1984). Evolution of a complex B-chromosome polymorphism in the grasshopper Eyprepocnemis plorans. *Chromosoma*, *89*(4), 290–293. https://doi.org/10.1007/bf00292477

Hewitt, G. (1973). Variable transmission rates of a B-chromosome in Myrmeleotettix maculatus (Thumb.) (Acrididae: Orthoptera). *Chromosoma*, *40*(1), 83-106. https://doi.org/10.1007/BF00319837

Hong, Z., Xiao, J., Peng, S., Lin, Y., & Cheng, Y. (2020). Novel B-chromosome-specific transcriptionally active sequences are present throughout the maize B chromosome. *Molecular Genetics and Genomics*, *295*(2), 313-325. https://doi.org/10.1007/s00438-019-01623-2

Houben, A. (2017). B Chromosomes – A Matter of Chromosome Drive. *Front. Plant Sci.*, *8*(210). https://doi.org/10.3389/fpls.2017.00210

Huang, W., Du, Y., & Zhao, X. (2016). B chromosome contains active genes and impacts the transcription of A chromosomes in maize (Zea mays L.). *BMC Plant Biol*, *16*(1). https://doi.org/10.1186/s12870-016-0775-7

Huskins, C., & Smith, S. (1934). A cytological study of the genus Sorghum Pers. II. The meiotic chromosomes. *Journal of Genetics*, *28*, 387–395.

Cheng, Y., & Lin, B. (2003). Cloning and Characterization of Maize B Chromosome Sequences Derived From Microdissection. *Genetics*, *164*(1), 299-310. https://doi.org/10.1093/genetics/164.1.299

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

Jackson, R., & Newmark, P. (1960). Effects of Supernumerary Chromosomes on Production of Pigment in Haplopappus gracilis. *Science*, *132*(3436), 1316-1317. https://doi.org/10.1126/science.132.3436.1316

Janaki-Amma, E. (1939). Supernumerary chromosomes in para-Sorghum. *Current Science*, *8*, 210–211.

Janaki-Ammal, E. (1940). Chromosome diminuition in a plant. *Nature*, *146*, 839–840.

Jiang, H., Lei, R., Ding, S., & Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, *15*(1). https://doi.org/10.1186/1471-2105-15-182

Johnson Pokorná, M., & Reifová, R. (2021). Evolution of B Chromosomes: From Dispensable Parasitic Chromosomes to Essential Genomic Players. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.727570

Jones, R. (1991). B-Chromosome Drive. *The American Naturalist*, *137*(3), 430–442.

Jones, R., Viegas, W., & Houben, A. (2008). A Century of B Chromosomes in Plants: So What?. *Annals of Botany*, *101*(6), 767-775. https://doi.org/10.1093/aob/mcm167

Jones, R. (1995). B chromosomes in plants. *New Phytologist*, *131*(4), 411–434. https://doi.org/10.1111/j.1469-8137.1995.tb03079.x

Jones, R., & Rees, H. (1982). *B chromosomes*. Academic Press.

Kamala, V., Sharma, H., Manohar Rao, D., Varaprasad, K., & Bramel, P. (2009). Wild relatives of sorghum as sources of resistance to sorghum shoot fly, Atherigona soccata. *Plant Breeding*, *128*(2), 137-142. https://doi.org/10.1111/j.1439-0523.2008.01585.x

Kamala, V., Singh, S., Bramel, P., & Rao, D. (2002). Sources of Resistance to Downy Mildew in Wild and Weedy Sorghums. *Crop Science*, *42*(4), 1357-1360. https://doi.org/10.2135/cropsci2002.1357

Karafiátová, M., Bednářová, M., Said, M., Čížková, J., Holušová, K., Blavet, N., & Bartoš, J. (2021). The B chromosome of Sorghum purpureosericeum reveals the first pieces of its sequence. *Journal of Experimental Botany*, *72*(5), 1606–1616. https://doi.org/10.1093/jxb/eraa548

Kayano, H. (1956). Cytogenetic studies in Lilium callosum. II. Preferential segregation of a supernumerary chromosome. *Mem. Fac. Sci.*, *2*, 53-60.

Kim, D., Paggi, J., Park, C., Bennett, C., & Salzberg, S. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907-915. https://doi.org/10.1038/s41587-019-0201-4

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540-546. https://doi.org/10.1038/s41587-019-0072-8

Koren, S., Walenz, B., Berlin, K., Miller, J., Bergman, N., & Phillippy, A. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*, *27*(5), 722-736. https://doi.org/10.1101/gr.215087.116

Kour, G., Kaul, S., & Dhar, M. (2013). Molecular characterization of repetitive DNA sequences from B chromosome in Plantago lagopus L. *Cytogenet. Genome Res.*, *142*, 121–128. https://doi.org/10.1159/000356472.

Kumke, K., Macas, J., Fuchs, J., Altschmied, L., Kour, J., Dhar, M., & Houben, A. (2016). Plantago lagopus B chromosome is enriched in 5S rDNA-derived satellite DNA. *Cytogenet. Genome Res.*, *148*, 68–73. https://doi.org/10.1159/000444873

Kuwada, Y. (1925). On the number of chromosomes in maize. *Botanical Magazine*, *39*, 227–234.

Lamb, J., Kato, A., & Birchler, J. (2005). Sequences associated with A chromosome centromeres are present throughout the maize B chromosome. *Chromosoma*, *113*, 337–349.

Lamb, J., Riddle, N., Cheng, Y., Theuri, J., & Birchler, J. (2007). Localization and transcription of a retrotransposon-derived element on the maize B chromosome. *Chromosome Res*, *15*, 383–398. https://doi.org/10.1007/s10577-007-1135-0

Langmead, B., & Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357-359. https://doi.org/10.1038/nmeth.1923

Lazarides, M., Hacker, J., & Andrew, M. (1991). Taxonomy, cytology and ecology of indigenous Australian sorghums (Sorghum Moench: Andropogoneae: Poaceae). *Australian Systematic Botany*, *4*(4), 591–635. https://doi.org/10.1071/sb9910591

Levin, D., Palestis, B., Jones, R., & Trivers, R. (2005). Phyletic hot spots for B chromosomes in angiosperms. *Evolution*, *59*(5), 962-9.

Lewis, H. (1951). The origin of supernumerary chromosomes in natural populations of Clarkia elegans. *Evolution*, *5*(2), 142–157. https://doi.org/10.2307/2405765

Liao, Y., & Shi, W. (2020). Read trimming is not required for mapping and quantification of RNA-seq reads at the gene level. *NAR Genomics and Bioinformatics*, *2*(3). https://doi.org/10.1093/nargab/lqaa068

Li, B., & Dewey, C. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1). https://doi.org/10.1186/1471-2105-12-323

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Lin, T., Xu, X., Ruan, J., Liu, S., Wu, S., Shao, X., Wang, X., Gan, L., Qin, B., Yang, Y., Cheng, Z., Yang, S., Zhang, Z., Xiong, G., Huang, S., Yu, H., & Li, J. (2018). Genome analysis of Taraxacum kok-saghyz Rodin provides new insights into rubber biosynthesis. *National Science Review*, *5*(1), 78-87. https://doi.org/10.1093/nsr/nwx101

Liu, H., Wu, S., Li, A., & Ruan, J. (2021). SMARTdenovo: a de novo assembler using long noisy reads. *Gigabyte*, *2021*, 1-9. https://doi.org/10.46471/gigabyte.15

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., & Fan, W. (2012). Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Briefings in Functional Genomics*, *11*(1), 25-37. https://doi.org/10.1093/bfgp/elr035

Longley, A. (1927). Supernumerary chromosomes in Zea mays. *Journal of Agriculture Research*, *35*, 769–784.

López-León, M., Cabrero, J., Pardo, M., Viseras, E., Camacho, J., & Santos,, J. (1993). Generating high variability of B chromosomes in Eyprepocnemis plorans (grasshopper). *Heredity*, *71*(4), 352–362. https://doi.org/10.1038/hdy.1993.149

López-León, M., Neves, N., Schwarzacher, T., Heslop-Harrison, J., Hewitt, G., & Camacho, J. (1994). Possible origin of a B chromosome deduced from its DNA composition using double FISH technique. *Chromosome Res*, *2*, 87–92. https://doi.org/10.1007/BF01553487

Love, M., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12). https://doi.org/10.1186/s13059-014-0550-8

MacManes, M. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*, *5*(13). https://doi.org/10.3389/fgene.2014.00013

Makunin, A., Romanenko, S., Beklemisheva, V., Perelman, P., Druzhkova, A., Petrova, K., Prokopov, D., Chernyaeva, E., Johnson, J., Kukekova, A., Yang, F., Ferguson-Smith, M., & Trifonov, V. (2018). Sequencing of supernumerary chromosomes of red fox and raccoon dog confirms a non-random gene acquisition by B chromosomes. *Genes*, *9*(8). https://doi.org/10.3390/genes9080405.

Manni, M., Berkeley, M., Seppey, M., & Zdobnov, E. (2021). BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols*, *1*(12). https://doi.org/10.1002/cpz1.323

Mann, J., Kimber, C., & Miller, F. (1983). The origin and early cultivation of sorghums in Africa. *Bulletin (Texas agricultural experiment station)*, (1454), 1-21.

Martis, M., Klemme, S., Banaei-Moghaddam, A., Blattner, F., Macas, J., Schmutzer, T., Scholz, U., Gundlach, H., Wicker, T., Simkova, H., Novák, P., Neumann, P., Bauer, E., Kubaláková, M., Haseneyer, G., Fuchs, J., Doležel, J., Stein, N., Mayer, K. et al. (2012). Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proc. Natl. Acad. Sci.*, *109*(33), 13343–13346. https://doi.org/10.1073/pnas.1204237109.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S., Wicker, T., Radchuk, V., Dockter, C., Hedley, P., Russell, J., Bayer, M., Ramsay, L., Liu, H., Haberer, G., Zhang, X., Zhang, Q., Barrero, R., Li, L., Taudien, S. et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature*, *544*(7651), 427-433. https://doi.org/10.1038/nature22043

McAllister, B., & Werren, J. (1997). Hybrid origin of a B chromosome (PSR) in the parasitic wasp Nasonia vitripennis. *Chromosoma*, *106*(4), 243–253. https://doi.org/10.1007/s004120050245

Meilleur, B., & Hodgkin, T. (2004). In situ conservation of crop wild relatives: status and trends. *Biodiversity and Conservation*, *13*(4), 663-684. https://doi.org/10.1023/B:BIOC.0000011719.03230.17

Mendelson, D., & Zohary, D. (1972). Behaviour and transmisson of supernumerary chromosomes in Aegilops speltoides. *Heredity*, *29*, 329–339.

Miao, V., Covert, S., & Van Etten, H. (1991). A fungal gene for antibiotic resistance on a dispensable ("B") chromosome. *Science*, *254*(5039), 1773–1776. https://doi.org/10.1126/science.1763326.

Michael, T., Jupe, F., Bemm, F., Motley, S., Sandoval, J., Lanz, C., Loudet, O., Weigel, D., & Ecker, J. (2018). High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-03016-2

Miller, J., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315-327. https://doi.org/10.1016/j.ygeno.2010.03.001

Mitsui, H., Yamaguchi-Shinozaki, K., Shinozaki, K., Nishikawa, K., & Takahashi, H. (1993). Identification of a gene family (kat) encoding kinesin-like proteins in Arabidopsis thaliana and the characterization of secondary structure of KatA. *Molecular and General Genetics MGG*, *238*(3), 362-368. https://doi.org/10.1007/BF00291995

Mochizuki, A. (1957). B chromosomes in Aegilops mutica Boiss. *Wheat Information Service*, *5*, 9–11.

Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621-628. https://doi.org/10.1038/nmeth.1226

Navarro-Domínguez, B., Martín-Peciña, M., Ruiz-Ruano, F., Cabrero, J., Corral, J., López-León, M., Sharbel, T., & Camacho, J. (2019). Gene expression changes elicited by a parasitic B chromosome in the grasshopper Eyprepocnemis plorans are consistent with its phenotypic effects. *Chromosoma*, *128*(1), 53-67. https://doi.org/10.1007/s00412-018-00689-y

Navarro-Domínguez, B., Ruiz-Ruano, F., Cabrero, J., Corral, J., López-León, M., Sharbel, T., & Camacho, J. (2017). Protein-coding genes in B chromosomes of the grasshopper Eyprepocnemis plorans. *Scientific Reports*, *7*(1). https://doi.org/10.1038/srep45200

Novák, P., Neumann, P., & Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, *11*(1), 378. https://doi.org/10.1186/1471-2105-11-378

Nur, U. (1962). Sperms, sperm bundles and fertilization in a mealy bug, Pseudococcus obscurus Essig. (Homoptera:Coccoidea). *J. Morphol.*, *111*, 173-199. https://doi.org/10.1002/jmor.1051110204

Nur, U. (1969). Mitotic instability leading to an accumulation of B-chromosomes in grasshoppers. *Chromosoma*, *27*, 1-19. https://doi.org/10.1007/BF00326108

Nygren, A. (1957). Poa timoleontis Heldr., a new diploid species of the section Bolbophorum and Gr. with accessory chromosomes only in meiosis. *LantbrHögsk. Annual Report Agricultural College of Sweden*, *23*, 489–495.

Palestis, B., Burt, A., & Jones, R. (2004). The distribution of B chromosomes across species. *Cytogenetic and Genome Research*, *106*(2-4), 151-158. https://doi.org/10.1159/000079281

Pansonato-Alves, J., Serrano, É., Utsunomia, R., Camacho, J., da Costa Silva, G., Vicari, M., Artoni, R., Oliveira, C., & Foresti, F. (2014). Single origin of sex chromosomes and multiple origins of B chromosomes in fish genus Characidium. *PloS one*, *9*(9). https://doi.org/10.1371/journal.pone.0107169

Papa, R., Acosta, J., Delgado-Salinas, A., & Gepts, P. (2005). A genome-wide analysis of differentiation between wild and domesticated Phaseolus vulgaris from Mesoamerica. *Theoretical and Applied Genetics*, *111*(6), 1147-1158. https://doi.org/10.1007/s00122-005-0045-9

Parker, J., Taylor, S., & Ainsworth, C. (1982). The B-chromosome system of Hypochoeris maculata. *Chromosoma*, *85*(2), 299–310. https://doi.org/10.1007/bf00294973

Parker, J., Taylor, S., & Ainsworth, C. (1982). The B-chromosome system of Hypochoeris maculata. *Chromosoma*, *85*(2), 299-310. https://doi.org/10.1007/BF00294973

Perfectti, F., & Werren, J. (2001). The interspecific origin of B chromosomes: experimental evidence. *Evolution; international journal of organic evolution*, *55*(5), 1069–1073. https://doi.org/10.1554/0014-3820(2001)055[1069:tioobc]2.0.co;2

Pontieri, P., Mamone, G., De Caro, S., Tuinstra, M., Roemer, E., Okot, J., De Vita, P., Ficco, D., Alifano, P., Pignone, D., Massardo, D., & Del Giudice, L. (2013). Sorghum, a healthy and gluten-free food for celiac patients as demonstrated by genome, biochemical, and immunochemical analyses. *Journal of Agricultural and Food Chemistry*, *61*(10), 2565-71. https://doi.org/10.1021/jf304882k

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics*, *10*(4), 354-366. https://doi.org/10.1093/bib/bbp026

R Core Team. (2021). *R: A language and environment for statistical computing*. Retrieved 2023-03-31, from https://www.R-project.org/

Rajendra Kumar, R., & Patil, J. (2015). Sorghum: Origin, Classification, Biology and Improvement. In *Sorghum Molecular Breeding* (1st edition, pp. 3-20). Springer.

Rakshit, S., & Bellundagi, A. (2019). Conventional Breeding Techniques in Sorghum. In *Breeding Sorghum for Diverse End Uses* (pp. 77-91). Elsevier. https://doi.org/10.1016/B978-0-08-101879-8.00005-X

Raman, V., & Krishnaswami, D. (1959). Accessory chromosomes in Sorghum nitidum. *Pers. Journal of Indian Botanical Society*, *34*, 278–280.

Raman, V., & Krishnaswami, D. (1960). Accessory chromosomes in Sorghum nitidum. *Pers. Journal of the Indian Botanical Society*, *39*, 278–280.

Raman, V., Meenakshi, K., Thangam, M., & Sivagnanam, L. (1964). The cytogenetical behaviour of B-chromosomes in Sorghum halepense. *Madras Agricultural Journal*, *51*, 72–73.

Randolph, L. (1941). Genetic characteristics of the B chromosomes in maize. *Genetics*, *26*(6), 608-631. https://doi.org/10.1093/genetics/26.6.608

Randolph, L. (1928). Types of supernumerary chromosomes in maize. *The Anatomical Record*, *41*, 102.

Reddy, N., & Yang, Y. (2005). Biofibers from agricultural byproducts for industrial applications. *Trends in Biotechnology*, *23*(1), 22-27.

Rice, A., Glick, L., Abadi, S., Einhorn, M., Kopelman, N., Salman-Minkov, A., Mayzel, J., Chay, O., & Mayrose, I. (2015). The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. *New Phytologist*, *206*(1), 19-26. https://doi.org/10.1111/nph.13191

Roach, M., Schmidt, S., & Borneman, A. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2485-7

Robinson, J., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E., Getz, G., & Mesirov, J. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24-26. https://doi.org/10.1038/nbt.1754

Robinson, M., McCarthy, D., & Smyth, G. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140. https://doi.org/10.1093/bioinformatics/btp616

Rothberg, J., Hinz, W., Rearick, T., Schultz, J., Mileski, W., Davey, M., Leamon, J., Johnson, K., Milgrew, M., Edwards, M., Hoon, J., Simons, J., Marran, D., Myers, J., Davidson, J., Branting, A., Nobile, J., Puc, B., Light, D. et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348-52. https://doi.org/10.1038/nature10242

Ruban, A., Schmutzer, T., Scholz, U., & Houben, A. (2017). How Next-Generation Sequencing Has Aided Our Understanding of the Sequence Composition and Origin of B Chromosomes. *Genes*, *8*(1), 294. https://doi.org/10.3390/genes8110294

Ruban, A., Schmutzer, T., Wu, D., Fuchs, J., Boudichevskaia, A., Rubtsova, M., Pistrick, K., Melzer, M., Himmelbach, A., Schubert, V., Scholz, U., & Houben, A. (2020). Supernumerary B chromosomes of Aegilops speltoides undergo precise elimination in roots early in embryo development. *Nature Communications*, *11*(2764). https://doi.org/10.1038/s41467-020-16594-x

Rutishauser, A., & Röthlisberger, E. (1966). Boosting mechanism of B chromosomes in Crepis capillaris. *Chromosomes Today*, *1*, 28-30.

Sandery, M., Forster, J., Macadam, S., Blunden, R., Jones, R., & Brown, S. (1991). Isolation of a sequence common to A- and B-chromosomes of rye (Secale cereale) by microcloning. *Plant Molecular Biology Reporter*, *9*(1), 21-30. https://doi.org/10.1007/BF02669286

Serrano, É., Utsunomia, R., Sobrinho Scudeller, P., Oliveira, C., & Foresti, F. (2017). Origin of B chromosomes in Characidium alipioi (Characiformes, Crenuchidae) and its relationship with supernumerary chromosomes in other Characidium species. *Comparative Cytogenetics*, *11*(1), 81-95. https://doi.org/10.3897/CompCytogen.v11i1.10886

Shen, R., Fan, J., Campbell, D., Chang, W., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Wickham Garcia, E., McBride, C., Steemers, F., Garcia, F., Kermani, B., Gunderson, K., & Oliphant, A. (2005). High-throughput SNP genotyping on universal bead arrays. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *573*(1-2), 70-82. https://doi.org/10.1016/j.mrfmmm.2004.07.022

Shi, X., Yang, H., Chen, C., Hou, J., Ji, T., Cheng, J., & Birchler, J. (2022). Effect of aneuploidy of a non-essential chromosome on gene expression in maize. *The Plant Journal*, *110*(1), 193-211. https://doi.org/10.1111/tpj.15665

Schaarschmidt, S., Fischer, A., Zuther, E., & Hincha, D. (2020). Evaluation of Seven Different RNA-Seq Alignment Tools Based on Experimental Data from the Model Plant Arabidopsis thaliana. *International Journal of Molecular Sciences*, *21*(5). https://doi.org/10.3390/ijms21051720

Schmid, M., Ziegler, C., Steinlein, C., Nanda, I., & Schartl, M. (2006). Cytogenetics of the bleak (Alburnus alburnus), with special emphasis on the B chromosomes. *Chromosome Research*, *14*(3), 231-242. https://doi.org/10.1007/s10577-006-1038-5

Schmidt, M., Vogel, A., Denton, A., Istace, B., Wormit, A., van de Geest, H., Bolger, M., Alseekh, S., Maß, J., Pfaff, C., Schurr, U., Chetelat, R., Maumus, F., & Aury, J. (2017). De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *The Plant Cell*, *29*(10). https://doi.org/10.1105/tpc.17.00521

Silva, T., Thomas, J., Dahlberg, J., Rhee, S., & Mortimer, J. (2022). Progress and challenges in sorghum biotechnology, a multipurpose feedstock for the bioeconomy. *Journal of Experimental Botany*, *73*(3), 646–664. https://doi.org/10.1093/jxb/erab450

Simpson, J., Wong, K., Jackman, S., Schein, J., Jones, S., & Birol, İ. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117-1123. https://doi.org/10.1101/gr.089532.108

Slavin, J. (2004). Whole grains and human health. Nutr Res Rev. *Nutrition Research Reviews*, *17*(1), 99-110. https://doi.org/10.1079/NRR200374

Snowden, J. (1936). *Cultivated Races of Sorghum* (1st edition). Royal Botanic Gardens, Kew.

Sohn, J., & Nam, J (2018). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbw096

Soneson, C., Love, M., & Robinson, M. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, *4*. https://doi.org/10.12688/f1000research.7563.1

Staub, R. (1987). Leaf striping correlated with the presence of B chromosomes in maize. *Journal of Heredity*, *78*(2), 71-74. https://doi.org/10.1093/oxfordjournals.jhered.a110339

Steenwyk, J., Buida, T., Gonçalves, C., Goltz, D., Morales, G., Mead, M., LaBella, A., Chavez, C., Schmitz, J., Hadjifrangiskou, M., Li, Y., Rokas, A., & Stajich, J. (2022). BioKIT: a versatile toolkit for processing and analyzing diverse types of sequence data. *Genetics*, *221*(3). https://doi.org/10.1093/genetics/iyac079

Taylor, P. (1998). Chromosomal drive and the evolution of meiotic nondisjunction and trisomy in humans. *Proceedings of the National Academy of Sciences*, *95*(5), 2361-2365. https://doi.org/10.1073/pnas.95.5.2361

Thind, A., Wicker, T., Šimková, H., Fossati, D., Moullet, O., Brabant, C., Vrána, J., Doležel, J., & Krattinger, S. (2017). Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly. *Nature Biotechnology*, *35*(8), 793-796. https://doi.org/10.1038/nbt.3877

Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M., Salzberg, S., Wold, B., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511-515. https://doi.org/10.1038/nbt.1621

Tripathi, A., Mishra, R., Maurya, K., Singh, R., & Wilson, D. (2019). Chapter 1 - Estimates for World Population and Global Food Availability for Global Health. In *he Role of Functional Food Security in Global Health* (., pp. 3-24). Academic Press. https://doi.org/10.1016/B978-0-12-813148-0.00001-3

Trivers, R., Burt, A., & Palestis, B. (2004). B chromosomes and genome size in flowering plants. *Genome*, *47*(1), 1–8. https://doi.org/10.1139/g03-088

Uhl., C., & Moran, R. (1973). The Chromosomes of Pachyphytum (Crassulaceae). *American Journal of Botany*, *60*(7), 648–656. https://doi.org/10.2307/2441442

USDA: Agricultural Research Service, National Plant Germplasm System. (2022). Germplasm Resources Information Network (GRIN-Taxonomy). Retrieved 2022-11-17, from

Valente, G., Conte, M., Fantinatti, B., Cabral-de-Mello, D., Carvalho, R., Vicari, M., Kocher, T., & Martins, C. (2014). Origin and Evolution of B Chromosomes in the Cichlid Fish Astatotilapia latifasciata Based on Integrated Genomic Analyses. *Molecular Biology and Evolution*, *31*(8), 2061-2072. https://doi.org/10.1093/molbev/msu148

Ventura, K., O'Brien, P., do Nascimento Moreira, C., Yonenaga-Yassuda, Y., Ferguson-Smith, M., & Houben, A. (2015). On the Origin and Evolution of the Extant System of B Chromosomes in Oryzomyini Radiation (Rodentia, Sigmodontinae). *PLOS ONE*, *10*(8), 618. https://doi.org/10.1371/journal.pone.0136663

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57-63. https://doi.org/10.1038/nrg2484

Wee, Y., Bhyan, S., Liu, Y., Lu, J., Li, X., & Zhao, M. (2019). The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. *Briefings in Functional Genomics*, *18*(1), 1-12. https://doi.org/10.1093/bfgp/ely037

Williams, C., Baccarella, A., Parrish, J., & Kim, C. (2016). Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics*, *17*(1). https://doi.org/10.1186/s12859-016-0956-2

Wilson, E. (1907). The supernumerary chromosomes of Hemiptera. *Science*, *26*, 870-871.

Wu, T. (1992). B-chromosomes in Sorghum stipoideum. *Heredity*, *68*, 457–463.

Xie, Q., & Xu, Z. (2019). Sustainable Agriculture: From Sweet Sorghum Planting, Ensiling to Ruminant Feeding. *Molecular Plant*, *12*, 603–606. https://doi.org/10.1016/j.molp.2019.04.001

Yang, I., & Kim, S. (2015). Analysis of Whole Transcriptome Sequencing Data: Workflow and Software, *13*(4). https://doi.org/10.5808/GI.2015.13.4.119

Yoshida, K., Terai, Y., Mizoiri, S., Aibara, M., Nishihara, H., Watanabe, M., Kuroiwa, A., Hirai, H., Hirai, Y., Matsuda, Y., & Okada, N. (2011). B chromosomes have a functional effect on female sex determination in Lake Victoria cichlid fishes. *PLoS Genet*, *7*(8). https://doi.org/10.1371/journal.pgen.1002203.

Zhang, T., Zhou, Y., Qi, S., Wang, Z., Qian, W., Ouyang, Y., Shen, W., Schatten, H., & Sun, Q. (2015). Nuf2 is required for chromosome segregation during mouse oocyte meiotic maturation. *Cell Cycle*, *14*(16), 2701-2710. https://doi.org/10.1080/15384101.2015.1058677

Zhao, Y., Li, M., Konaté, M., Chen, L., Das, B., Karlovich, C., Williams, P., Evrard, Y., Doroshow, J., & McShane, L. (2021). TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine*, *19*(1). https://doi.org/10.1186/s12967-021-02936-w

Ziegler, C., Lamatsch, D., Steinlein, C., Engel, W., Schartl, M., & Schmid, M. (2003). The Giant B Chromosome of the Cyprinid Fish Alburnus alburnus Harbours a Retrotransposon-Derived Repetitive DNA Sequence. *Chromosome Res*, *11*(1), 23–35. https://doi.org/10.1023/a:1022053931308

Zimin, A., Puiu, D., Luo, M., Zhu, T., Koren, S., Marçais, G., Yorke, J., Dvořák, J., & Salzberg, S. (2017). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii , a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research*, *27*(5), 787-792. https://doi.org/10.1101/gr.213405.116

Zimin, A., Puiu, D., Luo, M., Zhu, T., Koren, S., Marçais, G., Yorke, J., Dvořák, J., & Salzberg, S. (2017). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii , a

progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research*, *27*(5), 787-792. https://doi.org/10.1101/gr.213405.116

9 Seznam zkratek

BAM	binární verze souboru SAM
CCDB	databáze počtů chromozómů (The Chromosome Counts Database)
CENH3	centromerická varianta histonového proteinu 3 (Centromere Protein H3)
CPU	centrální procesorová jednotka (Central Processing Unit)
DAP	dny po opylení (Days After Pollination)
dXTP	deoxy–X–trifosfát (X = A (adenosin), C (cytidin), G (guanosin), T (thymidin))
EDTA	kyselina ethylendiamintetraoctová
FAO	Organizace pro výživu a zemědělství Spojených národů (The Food and Agriculture Organization of the United Nations)
FISH	fluorescenční *in situ* hybridizace
FPKM	fragmenty na kilobázi exonů na milion namapovanýach čtení (fragments per kilobase of exon per million mapped fragments)
GPU	grafický procesor (Graphics Processing Unit)
GTF	formát trasnferu genů (Gene Transfer Format)
Hi-C	zachycení konformace chromatinu (Chromatin Conformation Capture)
ICRISAT	Mezinárodní institut pro výzkům plodin v polosuchých oblastech (International Crops Research Institute for the Semi-Arid Tropics)
mobDNA	mobilní DNA
NATE	skupina retrotranspozónů vázajících se do specifických míst genomu (Nucleotide Addition Target Site)
NGS	sekvenování nové generace (Next Generation Sequencing)
OLC	algoritmus pro sestavení na základě překryvu a konsensu sekvencí (Overlap Layout Consensus)
ONT	Oxford Nanopore Technologies

PCA	analýza hlavních komponent (Principal Component Analysis)
PCR	polymerázová řetězová reakce (Polymerase Chain Reaction)
PNG	přenosná síťová grafika (Portable Network Graphics)
PMC	mateřská buňka pylu (Pollen Mother Cell)
qPCR	kvantitativní polymerázová řetězová reakce
rDNA	ribozomální DNA
RNA-Seq	sekvenování RNA
RPKM	fragmenty na kilobázi exonů na milion namapovanýach čtení (reads per kilobase of exon per million reads mapped)
SAM	textový soubor výsledků zarovnání sekvencí vůči referenčním sekvencím (Sequence Alignment Map)
satDNA	satelitní DNA
SMRT	jednomolekulové sekvenování v reálném čase (Single Molecule Real-Time)
TBE	Tris/Borát/EDTA pufr
TGS	sekvenování třetí generace (Third Generation Sequencing)
TPM	transkripty na milion (Transcipt Per Million)

10 Přílohy

Příloha 1: Unixové příkazy pro nástroje tvorby referenční sekvence a transkriptomu

Příkaz pro běh assembleru Flye:

```
flye --nano-hq Spu_ONT_Bs_newguppy.fq --out-dir $SCRATCHDIR –
threads 12
```

kde –nano-hq označuje mód pro dekodéry ONT Guppy5+, dále je zadána výstupní složka a počet vláken CPU.

Příkaz pro běh assembleru Canu:

```
canu -s Spu_canu_options.sh -p Spu_canu_assembly -d assembly
genomeSize=2210m java=/packages/run/jdk-8/current/bin/java
-nanopore Spu_ONT_Bs_newguppy.fq
```

kde Spu_canu_options.sh je konfigurační soubor, který řídí běh Canu. Dále jsou zadány vstupní data (-nanopore) a výstupní složka (-p), cesta k java balíčku a přibližná velikost genomu.

Příkaz pro běh assembleru MaSuRCA:

```
masurca -t 32 -i Bplus_S1_R1.fastq Bplus_S1_R3.fastq
-r Spu_ONT_Bs_newguppy.fq
```

kde -i určuje vstupní párové Illumina soubory a -r Oxford Nanopore soubor ve formátu fastq. -t udává počet vláken CPU.

Příkaz pro běh assembleru SMARTdenovo:

```
smartdenovo.pl -p SpuB_assembly_smartdenovo_corr -e dmo -t 32 -k
23 -c 1 ../cns_final.fasta.gz
```

kde jsou zadány vstupní a výstupní destinace, -t udává opět počet vláken CPU, -e algoritmus dmo, -k délku k-merů a -c stupeň korekce (přečištění reference hrubými daty).

Korekce pomocí nástroje Medaka se skládá ze tří úkonů:

1. Mini_align:

```
mini_align -i SPS_ONT_Bs_newguppy.fq -r
SpuB_smartdenovo_corr.dmo.cns -P -m -p calls_to_draft.bam -t 32
```

kde mini_align přirovnává čtení vstupnímu assembly (-i), dále je uvedený samotný soubor s hrubou assembly, výstup ve formátu bam (-p) a počet vláken CPU (-t). -P indikuje, že výstup bude popisovat informace o identitě zarovnání, délce, pozici a orientaci. -m udává mód pro zarovnání sekvencí.

2. medaka consensus:

```
Medaka consensus calls_to_draft.bam.bam contigs_out.hdf --model
r941_min_hac_g507 --batch 120 --threads 2
```

Vstupním souborem pro druhý krok byl soubor ve formátu bam z prvního kroku, dále je uveden výstupní soubor ve formátu hdf. Parametr --model určuje použitý model pro tvorbu konsenzuálních sekvencí, v tomto případě byl použit model r941_min_hac_g507, který je optimalizovaný pro použití s nanopórovými sekvenátory typu Oxford Nanopore. Tento krok vyžaduje paralelismus GPU, jehož množství je uvedeno pod parametrem --batch.

3. medaka stitch:

```
Medaka stitch --threads 4 contigs_out.hdf SpuB_smartdenovo_corr.dmo.cns
final.output.fa
```

kde je uveden počet vláken CPU (--t), vstupní a výstupí sobor.

Příkaz pro korekci nástrojem NextPolish byl řízen souborem run.cfg:

```
[General]
job_type = local
job_prefix = nextPolish
task = best
rewrite = yes
rerun = 3
parallel_jobs = 1
multithread_jobs = 4
genome = ./medaka_flye.fa #genome file
genome_size = auto
workdir = ./01_rundir
polish_options = -p {multithread_jobs}

[sgs_option]
sgs_fofn = ./sgs.fofn
sgs_options = -max_depth 100 -bwa
```

který udává parametry pro paralelismus, vstupní soubory referenční sekvence i hrubých dat krátkého čtení (soubor sgs_fofn) a pracovní prostředí.

Pro mapování transkriptomických nástrojem STAR v2.7.10b na referenční sekvenci byl nejdříve vytvořen index příkazem v shellu (unixovém příkazovém řádku).

Příkaz pro vytvoření indexu:

```
STAR --runThreadN 16 --runMode genomeGenerate --genomeDir
index_STAR/ --genomeFastaFiles
SpuB.smartdenovo.Medaka.nextpolish.2.purge_haplotigs_ShortNames.fa
--sjdbGTFfile SpuB_all_genes_merged.gtf
```

kde --runThreadN je počet vláken CPU, --runMode genomeGenerate úloha nástroje generující index genomu, --genomeDir cesta k adresáři pro výsledné soubory indexu, --genomeFastaFiles soubor referenční sekvence *S. purpureosericeum* formátu fasta a --sjdbGTFfile anotační soubor *S. purpureosericeum* formátu GTF.

Po vytvoření STAR indexu proběhlo samotné mapování čtení na referenční sekvenci.

Příkaz pro mapování:

```
STAR --runThreadN 8 --genomeDir index_STAR/ --readFilesIn
raw_sequences/Spu0_RNA_7dap_1_R1 raw_sequences/Spu0_RNA_7dap_1_R2
--quantMode TranscriptomeSAM --outSAMtype BAM Unsorted --
outFileNamePrefix mapped/ Spu0_RNA_7dap_1
```

kde --runThreadN udává počet vláken CPU, --genomeDir složku pro výstupní soubory indexu a --readFilesIn vstupní transkriptomická hrubá data. --quantMode udává mód nástroje Star, v tomto případě byl použit mód pro TranscriptomeSAM s výstupním soubourem BAM, který udává --outSAMtype parametr. Výsledky byly ukládány do složky /mapped pod parametrem --outFileNamePrefix.

Následně byla vytvořena reference pro kvantifikaci nástrojem RSEM:

```
rsem-prepare-reference SpuB.smartdenovo.Medaka.nextpolish.2.fa
--gtf SpuB_all_genes_merged.gtf RSEM_reference
```

kde mód rsem-prepare-reference nástroje RSEM potřebuje jako argumenty referenční sekvenci ve formátu fasta, GTF anotační soubor a jméno výstupního souboru.

Příkaz pro kvantifikaci čtení nástrojem RSEM:

```
rsem-calculate-expression --bam --no-bam-output -p 16 --paired-end
mapped/Spu0_RNA_7dap_1_Aligned.toTranscriptome.out.bam
RSEM_reference Spu0_RNA_7dap_1_counts
```

kde bylo v módu rsem-calculate-expression nástroje RSEM zadány parametry vstupu ve formátu bam, počet vláken CPU (-p), umístění RSEM reference a název výstupního souboru.

Příloha 2: Koncentrace izolované RNA, výstup sekvenování a kvalita mapování

Tabulka 3: Koncentrace izolované RNA ze vzorků (ng/µl), hrubé čtení z NGS a unikátně namapované čtení (%) pro 4 vývojové kategorie embryí.

Stádium	Replika	Koncentrace (ng/µl)		Hrubé čtení z NGS		Unikátně namapované čtení (%)	
		B+	B0	B+	B0	B0	B+
7DAP	1	0,096	0,099	16,821,950	23,576,790	32,85	58,12
	2	0,027	0,075	81,113	22,655,131	53,86	42,25
	3	0,204	0,175	24,084,803	14,676,869	71,86	63,11
	4	0,209	0,215	22,088,045	17,140,051	71,45	68,81
	5	0,145	0,324	22,122,690	16,696,992	72,85	64,73
14DAP	1	0,581	0,145	15,022,526	54,323,449	10,45	70,41
	2	2,232	0,804	71,021,099	64,021,230	70,22	70,76
	3	5,651	0,898	27,192,660	17,727,165	67,57	71,73
	4	1,637	0,463	13,215,397	36,773,719	65,25	72,55
	5	2,097	0,515	85,197,519	34,003,384	73,19	72,39
21DAP	1	1,886	0,800	29,140,715	35,248,072	62,49	62,71
	2	4,610	0,327	20,189,621	50,978,559	62,44	68,01
	3	2,341	3,485	31,302,315	49,204,073	68,68	75,44
	4	5,651	2,060	25,422,232	28,985,985	71,08	72,30
	5	3,152	3,874	27,143,887	17,862,420	73,70	74,07
28DAP	1	3,811	2,251	25,547,255	18,025,387	67,76	66,13
	2	11,029	6,277	13,731,710	39,695,386	63,97	62,90
	3	23	14	17,980,259	63,742,647	72,85	71,06
	4	12	8	12,783,664	29,317,966	74,55	73,69
	5	49	61	29,211,546	21,740,494	71,66	68,98