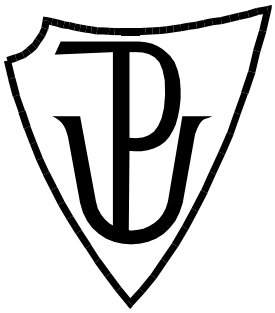
UNIVERZITA PALACKÉHO V OLOMOUCI

Přírodovědecká fakulta

Katedra biochemie



Komparativní analýza subgenomů *Pachycladon exilis*

DIPLOMOVÁ PRÁCE

Autor:	**Bc. Lucie Šimková**
Studijní program:	B1406 Biochemie
Studijní obor:	Bioinformatika
Forma studia:	Prezenční
Vedoucí práce:	**Mgr. Jan Bartoš, Ph.D**
Rok:	2019

Prohlašuji, že jsem diplomovou práci vypracoval/a samostatně s vyznačením všech použitých pramenů a spoluautorství. Souhlasím se zveřejněním diplomové práce podle zákona č. 111/1998 Sb., o vysokých školách, ve znění pozdějších předpisů. Byl/a jsem seznámen/a s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, ve znění pozdějších předpisů.

V Olomouci dne

Poděkování

Nejdříve bych na tomto místě ráda poděkovala Mgr. Janu Bartošovi, PhD. za odborné vedení, trpělivost, podnětné a cenné rady a ochotu, kterou mi v průběhu zpracování diplomové práce věnoval. Ráda bych také poděkovala své rodině a blízkým, kteří mě při vytváření této práce podporovali, a bez jejichž pomoci by nebylo možné práci dokončit.

Bibliografická identifikace

Jméno a příjmení autora	Bc. Lucie Šimková
Název práce	Komparativní analýza subgenomů *Pachycladon exilis*
Typ práce	Diplomová
Pracoviště	Katedra biochemie
Vedoucí práce	Mgr. Jan Bartoš, Ph.D
Rok obhajoby práce	2019
Abstrakt	*Pachycladon exilis* je mesopolyploidní druh endemický na Novém Zélandě, který vznikl z allopolyploidního předka před asi 2 miliony let. V rámci diplomové práce byla provedena komparativní analýza *P. exilis* s *Arabidopsis lyrata*, byly porovnávány subgenomu v rámci *P. exilis* a následně provedena komparativní analýza se sestaveným transkriptomem *Crucihimalaya himalaica*. Podařilo se vytvořit metodiku pro identifikaci kandidátních subgenomů *P. exilis* a jejich přiřazení k rodičovským genomům.
Klíčová slova	Celogenomová duplikace, kolinearita, polyploidie, *Pachycladon exilis*
Počet stran	51
Počet příloh	0
Jazyk	Český

Bibliographical identification

Autor's first name and surname	Bc. Lucie Šimková
Title	Comparative analysis of *Pachycladon exilis* subgenomes
Type of thesis	Diploma
Department	Department of Biochemistry
Supervisor	Mgr. Jan Bartoš, Ph.D
The year of presentation	2019
Abstract	*Pachycladon exilis* is a mesopolyploid species, endemic in New Zealand with an allopolyploid origin. This diploma thesis focuses on a comparative analysis of *P. exilis* and *Arabidopsis lyrata*, a comparison of *P. exilis* subgenomes and comparative analysis with *Crucihimalaya himalaica*. The thesis provide a methodology for identification of candidate subgenomes of *P. exilis* and their assignment to a potential evolutionary parental genomes.
Keywords	Whole genome duplication, polyploidy, collinearity, *Pachycladon exilis*
Number of pages	51
Number of appendices	0
Language	Czech

OBSAH

Cíle práce

1) Vypracování literární rešerše na téma evoluce genomu v kontextu polyploidie, sestavování genomické sekvence a komparativní analýza.

2) Porovnání genových oblastí dvou subgenomů *P. exilis*.

3) Provedení srovnávací analýzy s genomy *Arabidopsis lyrata* a *Crucihimalaya himalaica*.

1 ÚVOD

Tato diplomová práce se zabývá analýzou genomu *Pachycladon exilis*. Jedná se o druh z čeledi *Brassicaceae* nacházející se v Jižních Alpách na Jižním ostrově Nového Zélandu. Dřívější studie rodu *Pachycladon* provedené Mandákovou (2010) nebo Lysákem (2009) naznačují evoluční umístění rodu *Pachycladon* v rámci čeledi *Brassicaceae* a jeho vztah k rodu *Arabidopsis*. Cílem diplomové práce je uvést druh *P. exilis* do souvislosti s druhem *Crucihimalaya himalaica* jakožto kandidátem na evolučního rodiče jednoho ze dvou subgenomů *P. exilis*. V případě, že náhled na určení subgenomů bude úspěšný, otevřou se možnosti na určení původu i druhého subgenomu a osvětlení evolučního původu celého rodu *Pachycladon*.

Během experimentální části bude provedena komparativní analýza s genomem *A. lyrata*, určení unikátních a duplikovaných genů v genomu *P. exilis*, sestavení transkriptomu *C. himalaica*, komparativní analýza s tímto sestaveným transkriptomem a následně pilotní analýza určení subgenomů *P. exilis*. Výsledkem je určení kolineárních oblastí s *A. lyrata*, vztahů mezi dvěma vybranými scaffoldy a definování cesty pro kompletní analýzu s *C. himalaica*.

2 SOUČASNÝ STAV ŘEŠENÉ PROBLEMATIKY

2.1 Celogenomová duplikace

Duplikace je formou mutace, při které je genová oblast replikovaná a v některých případech vložená na fyzicky odlišné místo. K tomuto jevu přispívá několik mechanismů. Při zohlednění dopadu na obsah genů je nejdramatičtější duplikace celého chromosomu nebo dokonce celého genomu (Panchy *et al*, 2016). Vzhledem k frekvenci ancestrálních i novodobých celogenomových duplikací (WGD) u rostlin není překvapením, jak velké množství duplikovaných genů rostlinné genomy obsahují. Například u *Arabidopsis* můžeme k nejméně 60% všech genů přiřadit alespoň jednoho paraloga v odpovídajícím syntenickém bloku, který vznikl během jedné ze tří WGD (Renny-Byfield a Wendel, 2014).

Na rozdíl od ostatních eukaryotických genomů se rostlinné genomy vyvíjely ve výrazně vyšší míře (Kejnovsky *et al*., 2009) a rozdíly ve velikostech genomu mezi rostlinami jsou mnohem větší než mezi jakýmikoliv jinými eukaryotami. Například u dvouděložných rostlin se velikost genomu pohybuje od 63 Mb u masožravých rostlin rodu *Genlisea* (Greilhuber *et al*., 2006), až k přibližně 150 Gb u *Paris japonica* z čeledi kýchavicovité (Pellicer *et al.*, 2010). Rostlinné genomy navíc často obsahují velké množství duplikovaných genů, WGD se v posledních 200 milionech let u krytosemenných rostlin objevila hned několikrát, na rozdíl od člověka, u kterého se poslední WGD vykytla asi před 450 miliony let (Panopoulou *et al*., 2003), a u pučících kvasinek pak asi před 200 miliony let (Wolfe a Shields, 1997). WGD nebo polyploidizace je nejdramatičtější mechanismus genové duplikace, který vede k náhlému zvýšení jak velikosti genomu, tak také zdvojení celé genové sady. Avšak nemusí to být jediným mechanismem vedoucím k duplikovaným genům. Obecně genová duplikace vytváří dvě genové kopie, což teoreticky umožňuje jedné nebo oběma nebýt pod selektivním tlakem a v některých případech získat nové genové funkce, které mohou napomáhat k adaptaci. Můžeme tedy říct, že genová duplikace je jednou ze změn v genomu, která může vést k evolučním novinkám (Panchy *et al.*, 2016). Mezi další patří vybírání mezi již existujícími geny (True a Carrol, 2002), vytváření genů *de novo* z mezigenových sekvencí (Schlötterer, 2015) nebo vytváření nových regulačních míst transkripce, které mohou vést k alternativní expresi genů (Wray *et al.*, 2003).

Vzhledem k tomu, jak časté jsou WGD, a genové duplikace obecně, ve všech evolučních liniích kvetoucích krytosemenných rostlin, je překvapující, jak jsou genomy krytosemenných rostlin často male a dokonce i množství identifikovaných genů je mnohem nižší, než by se očekávalo. Zejména pak při předpokladu, že každá WGD zdvojnásobí jak velikost genomu, tak i počet genů. Přestože je rozsah velikosti genomu u krytosemenných rostlin velký, velikost genomu zvolených modelových rostlin je stále poměrně malá (Leitch *et al.*, 1998). Vysvětlení velikosti rostlinných genomů je možné vyvodit z analýzy pěstovaných užitkových rostlin a jejich příbuzných druhů, kdy po WGD obvykle následuje také proces diploidizace, kdy v polyploidním genomu probíhají ne příliš objasněné procesy způsobující delece genů za vzniku téměř diploidního genomu (Wolfe, 2001). Tento proces je evidentně velmi účinný a maže tak stopy WGD, která až do éry moderní genomiky nebyla vůbec evidovaná. Zejména pak u semenných rostlin, kde pravé diploidie neexistují a téměř všechny suchozemské rostliny jsou paleoploidy (Renny-Byfield a Wendel, 2014). Již dlouho je známo, že rozdíly ve velikosti genomu nemusí být spojeny s rozdílným počtem genů, ale spíše s líšícím se nashromážděním nebo odstraněním repetitivní a negenové DNA. Není tedy překvapením, že ke zmenšení velikosti u allopolyploidních genomů nejvíce přispívá odstraňování repetitivních sekvencí DNA (Renny-Byfield *et al.*, 2011). Tento proces obvykle probíhá rekombinačními mechanismy, kdy některé z nich za sebou nechávají stopy, jako například sólové koncové repetice, které můžeme pozorovat například u rýže nebo ječmene (Shirasu *et al.*, 2000, Vitte *et al.*, 2003), nebo nahromadění malých delecí, které můžeme pozorovat například u bavlníku (Grover *et al.*, 2007). Zajímavostí je, že toto odstraňování repetitivních sekvencí může být subgenomově specifické. Například u tabáku je mateřský genom relativně neporušený, kdežto otcovský projevuje známky genomové eroze (Renny-Byfield *et al.*, 2011). Procesy odstraňování DNA tedy způsobují snížení velikosti genomu, ale také mohou vést k rozdílnosti homeologních chromosomů, kdy se stávají více diploidními. Spekuluje se o tom, že tento proces by mohl být důležitý mimo jiné pro zarovnání chromosomových párů před rekombinací během meiosy, stabilizování meiotických párů a zvyšování fertility vzikajících allopolyploidních linií (Feldman a Levy, 2012). Navzdory skutečnosti, že se homeologní chromosomy mohou zřetelně lišit s ohledem na jejich unikátní části a oblasti obsahující duplikace, zbývající repetitivní obsah allopolyploidních genomů má tendenci být více homogenní. Zejména bývají zachovávány původní genomické odlišnosti způsobené translokacemi, divergencí sekvencí a výměnami repetitivní DNA, ke kterým docházelo během uplynulých více než

milionu let. Jak ukazuje *in situ* hybridizace, subgenomy se tedy z dlouhodobého hlediska stávají téměř nerozeznatelné (Renny-Byfield *et al.,* 2013). Navíc se během více než milionu let počet duplikovaných chromosomů obvykle sníží, takže polyploidé se ve výsledku stejně vrátí k podobnému počtu chromosomů, jako se vyskytovalo u jejich ancestrálního předka (Lim *et al.,* 2007).

Navzdory příspěvku mnoha duplikačních mechanismů a dalších změn na velikost genomu, zůstává u rostlin obsah genů relativně stejný napříč všemi suchozemskými druhy. Vzhledem k tomu, že v rostlinných liniích vedoucích k *Arabidopsis* proběhlo nejméně pět WGD a za předpokladu, že společný předchůdce všech suchozemnských rostlin měl přibližně 10 tisíc genů, u existujících rostlin by počet genů stoupl až na 320 tisích i v případě, že neuvažujeme další duplikační mechanismy (Panchy *et al.,* 2016). Toto očekávané vysoké číslo nám indikuje rozsáhlou ztrátu genů, a i přestože některé duplikované geny přežily přes miliony až stovky milionů let, převládající situací je, že jedna z kopií duplikovaného genu je ztracena (Hanada *et al.,* 2008). I přesto se však v rostlinných genomech vyskytuje převaha duplikovaných genů, což je způsobeno zejména vysokou mírou duplikace během evoluce a také stavem, kdy u některých duplikací je preferováno jejich zachování (Panchy *et al.,* 2016). Proces ztráty genů může zahrnovat delece celé duplikované sekvence a/nebo pseudogenizaci za vzniku mutací způsobujících ztrátu funkce. V případě, že jsou dva duplikované geny kompletně identické, neměla by delece kterékoliv z nich danému organismu přinést žádnou selektivní nevýhodu. Avšak reálně jsou jedinci s duplikovanými geny jen zřídka shodní. Nefunkční duplikace nejsou vždy ztraceny, rostlinné genomy jsou posety tisíci zdánlivě nefunkčních duplikací označovaných jako pseudogeny (Guo *et al.,* 2009). Pseudogeny se identifikují na základě jejich podobnosti s funkčními geny a výskytu mutace zbavující funkce (například předčasný stop kodon nebo posunutí čtecího rámce v genech kódujících proteiny) (Vanin, 1985). Ačkoliv jsou pseudogeny pravděpodobně nefunkční, existuje malá skupina pseudogenů v rýži a *Arabidopsis*, která je i přesto exprimována (Yamada *et al.,* 2003, Thibaud-Nissen *et al.,* 2009).

Společným tématem ohledně modelů zachování duplikovaných genů je otázka, zda se v případě, že jsou obě kopie genů zachovány, vyskytují rozdíly ve funkci, expresi nebo interakci. V některých případech duplikované geny získávají nové funkce a přispívají tak k evolučním novinkám. U některých duplikací je pak možné pozorovat nové funkce díky změně v morfologických znacích, kdy některý ze znaků může být z fenotypu vyřazen nebo naopak posílen (Hanada *et al.,* 2009).

Po duplikaci genů se rychlost evoluce (a substituce sekvencí) obvykle zvyšuje, alespoň zpočátku, díky tomu, že přítomnost dvou kopií uvolňuje selekční tlak proti dříve nepříznivým mutacím (Scannell a Wolfe, 2008). V souladu s tímto předpokladem duplikované geny vykazují volnější purifikační selekci, stejně jako rozdíly ve strukturách sekvencí, jako například v délce kódující oblasti nebo v distribuci indelů (Wang *et al.*, 2013). Tento růst rychlosti evoluce nemusí být nezbytně stejný pro obě kopie – například u kukuřice se rozdíl asymetrického vývoje WGD duplikací pohybuje kolem 21,2% a u *Physcomytrella patens* pak dokonce kolem 68,3% (Carretero-Paulet a Fares, 2012). Vzhledem k neofunkcionalizačnímu modelu tato situace odpovídá předpokladu, kdy jedna z kopií zachovává původní funkce a mutace přispívající k novým funkcím se hromadí ve druhé kopii. Avšak situace může být také popsána jako postupná ztráta funkcí v jedné kopii, kdy ve chvíli, kdy je jedna z kopií zasažena degenerativní mutací, pravděpodobnost výskytu další mutace ve stejné kopii je vyšší. Z toho vyplývá, že pro každý konkrétní případ asymetrie u duplikovaných částí genomu je třeba určit příčinu a neofunkcionalizaci tedy nelze standardně předpokládat (Panchy *et al.*, 2016). Asymetrie v rychlosti evoluce ukazuje rozdíly v evoluci u duplikovaných kopií, avšak existuje souvislost také s příbuznými neduplikovanými oblastmi. Na posouzení, zda duplikace ovlivňují rychlost evoluce, je možné použít rozdílné přístupy. Zaprvé, duplikovaná dvojice může být porovnána s předpokládaným předkem. Tento přístup je však poněkud obtížný, zejména kvůli komplikacím spojeným s určováním předků. V druhém případě může být duplikovaná dvojice porovnána s blízce příbuznm singletonem, který v minulosti mohl být také duplikován, ale jehož paralogy byly ztraceny. V tomto případě pak duplikáty i singletony mají společného ancestrálního předka a tedy by pravděpodobně měly zachovávat i stejné funkce (Panchy *et al.*, 2016).

2.2 Polyploidie

Situace, kdy organismus obsahuje více než dvě sady chromosomů, se nazývá polyploidie. Tento stav je už delší dobu uznávám jako důležitý rys rostlinných genomů a předpokládá se, že vede k rozmanitosti fenotypů (Soltis *et al.*, 2009). Polyploidie se rozděluje do dvou typů. První, allopolyploidie, zahrnuje vznik mezidruhovou hybridizací. Ve druhém případě se mezidruhová hybridizace nevyskytuje a hovoříme tak o autopolyploidii. Allopolyploidové typicky obsahují v době vzniku dva nebo více odlišných subgenomů, mají v mnoha případech fixní heterozygoty a jejich chromosomy při meiose netvoří multivalenty. Autopolyploidové mají alespoň tři kopie stejného, nebo téměř identického, genomu a stejnoměrně u nich probíhá multivalentní nebo náhodné bivalentní párování mezi homologními chromosomy (Weindel a Doyle, 2005). V minulosti byly polyploidní rostliny identifikovány cytologií zjištěním počtu chromosomů a pokrok, který přinesla genomická éra, odhalil desítky dalších ancestrálních a netradičních případů polyploidie. Například rozsáhlá analýza distribuce stáří událostí vedoucích k duplikaci genů odhalila dříve neznámé případy paleopolyploidie u užitkových a modelových rostlin (Blanc a Wolfe, 2004). Pomocí synonymní substituce bylo možné odhadnout čas divergence mezi duplikovanými geny a identifikovat píky odpovídající duplikačním událostem, které vycházejí z dřívější WGD (Renny-Byfield a Wendel, 2014). Další výskyty WGD byly určeny díky tomu, že některé části rostlinných genomů se u nich vyskytují v několika kopiích (u *Arabidopsis* je to až 58%) a tyto sady genů vykazují vysokou míru kolinearity s jejich duplikovanými protějšky (Blanc *et al.*, 2000). Od prvotní identifikace WGD byly u linie *Arabidopsis* dále dokázány další WGD v průběhu evoluce a začaly se objevovat mimo jiné i důkazy o WGD napříč fylogenezí mnoha skupin krytosemenných kvetoucích rostlin (Soltis *et al.*, 2009). Fylogenetická analýza exprimovaných sekvenačních značek (EST) ze semenných a kvetoucích rostlin, za použití plavuní a mechorostů jako outgroup, ukázala ancestrální WGD již na počátku evoluce semenných rostlin a další vnořenou WGD při vzniku krytosemenných rostlin (Jiao *et al.*, 2011).

Navzdory značným nevýhodám polyploidizace, například problematičtější replikace genomu, náchylnost k polyploidní mitóze a meiozé za vzniku aneuploidních buněk, nebo epigenetické nestabilitě, paleogenomická data naznačují, že právě opakované cykly polyploidizace vedly k velké genetické rozmanitosti rostlin (Soltis *et al.*, 2009). Není pravděpodobné, že by polyploidizace sama o sobě vedla k tolika zaznamenaným obměnám napříč fylogenezí krytosemenných rostlin, spíše se to připisuje post-

polyploidní diploidizaci (PPD), která v liniích paleopolyploidie a mesopolyplodie vytvářela značnou genetickou a taxonomickou rozmanitost (Mandáková a Lysák, 2018). Komplexní proces PPD, který zahrnuje velké množství evolučních modifikací transformujících polyploidní genom do kvaziploidního, je často přehlížený a málo studovaný (Dodsworth *et al.*, 2016). PPD je spojeno s širokou řadou procesů, například zmenšování genomu, rozdělení specifických subgenomů (k čemuž patří zachování nebo ztráta určitého genu a genová sub- nebo neo- funkcionalizace), modulace genové exprese, aktivace transponovatelných elementů a epigenetické reprogramování. Na chromosomální úrovni je PPD zprostředkované homeologními rekombinacemi a nelegitimními rekombinacemi mezi transponovatelnými elementy, což vede k strukturním chromosomálním změnám, tedy i redukci počtu chromosomů (Mandáková a Lysák, 2018). Jedná se o jednu ze základních cest diploidizace a snižování počtu chromosomů. Tradičně je tento proces vnímán jako jeden z mechanismů vedoucích k přeměně polyploidů na funkční diploidy. Momentálně dostupná data komparativní genomiky naznačují, že tento proces je v evoluci mnohem častější než například zvyšování počtu chromosomů zlomem v centromeře (Mandáková a Lysák, 2018).

I když polyploidie přináší mnoho nových informací k pochopení rostlinných genomů, stále se jedná o obtížné odvětví, zejména kvůli množství genomických dat. Oproti diploidům mají polyploidy obvykle větší velikost genomu a to i bez ohledu na jeho zmenšování v rámci evoluce (Leitch a Bennett, 2004), což už samo o sobě činí polyploidy méně vhodnými modelovými organismy. Sekvenování větších genomů je finančně náročnější a je potřeba větší výpočetní kapacita při sestavování genomu než u preferovaných inbredních diploidů (Kelly *et al.*, 2012). Například první sekvenovaný genom kvetoucí rostliny patří *Arabidopsis thaliana* a byl vybrán částečně díky jeho malé velikosti. Prvotní sestavování genomu bylo založeno na Sangerově metodě sekvenování a hlavním hlediskem pro výběr potencionálních druhů na sekvenování byly zejména náklady s tímto spojené (Renny-Byfield a Wendel, 2014). Až sekvenování další generace umožnilo velký rozmach projektů zaměřených na sekvenování genomu a rozšíření na sekvenování dalších užitkových a ostatních rostlin v celé fylogenetické šíři. I tak bylo sekvenování omezeno na druhy s relativně malými genomy, částečně kvůli ceně a také kvůli problémům s repetitivními úseky, které tvoří převládající část většiny rostlinných genomů (Kelly *et al.*, 2012). Další výzvou při sestavování allopolyploidních genomů byla jejich genomová složitost a nadměrné množství repetitivních sekvencí. Aktuální pokrok v sekvenačních technikách naštěstí přináší naději ve vyřešení i těchto komplikací.

Nicméně i tak stále zůstávají problémy při sestavování polyploidních genomů a při správném určení a sestavení homeologních genů a alel (Renny-Byfield a Wendel, 2014).

2.3 Polyploidie v čeledi brukvovité

U australských rostlin čeledi *Brassicaceae* (brukvovité) s diploidním počtem chromosomů (n = 4 až 6) byla jako příčina násobného počtu chromosomů odhalena mesopolyploidní WGD. Cytogenetická analýza ukázala, že karyotyp australského tribu *Camelineae* pochází z osmi rodových chromosomů (n = 8) vzniklých WGD, kdy následně došlo ke snížení počtu chromosomů. (Mandáková *et al*., 2010).

Hybridizace a WGD jsou důležitými procesy vedoucími k genetické obměně a specializaci u suchozemských rostlin. Zejména u krytosemenných rostlin bylo zdokumentováno mnoho případů specializace způsobené těmito procesy (Soltis *et al.*, 2007). Komparativní výzkum vycházející z celogenomových a EST sekvenačních dat přináší mnoho přesvědčivých důkazů o ancestrálních WGD během evoluce ve fylogenetické linii krytosemenných. U brukvovitých (případně křížatých, *Brassicaceae*) byla provedena analýza na základě genomové sekvence *Arabidopsis thaliana* a ukázal se výskyt tří paleopolyploidních WGD (alpha, beta, gama) (Bowers *et al.*, 2003). Zatímco fylogenetické zařazení nejstarší události (gama) je stále nejisté (Soltis *et al.*, 2009), u beta WGD bylo potvrzeno, že vedla k rozdělení fylogeneze mezi čeleděmi *Caricaceae* a *Brassicaceae* (Tang *et al*., 2008). Nejmladší (alpha) duplikace zřejmě proběhla pouze v rámci *Brassicaceae* a je podobná k celogenomocé triplikaci v sesterské čeledi *Cleomaceae* (Barker *et al.*, 2009). Alpha duplikaci můžeme zařadit do období asi 23-43 milionů let zpět (mya) (Barker *et al.*, 2009) a beta duplikaci asi 72 mya (Ming *et al*., 2008). Během tří paleoploidních WGD probíhala také diplodizace vedoucí k diploidnímu genomu, zmenšení velikosti genomu a přeskupení chromosomů (Thomas *et al.*, 2006). Například funkčeně diploidní a stabilní genom *Arabidopsis* je výsledkem přeskupení karyotypu již diploidizovaného genomu ancestrálního předka z čeledi křížaté (Lysák *et al.*, 2006). Po WGD byly některé duplikované geny smazány nebo podrobeny sub- nebo neo- funkcionalizaci (Sémon a Wolfe, 2007). V případě genů, u kterých záleží na množství produktu, jako jsou například transkripční faktory, byl obvykle preferenčně zachován pouze gen v jednom z duplikovaných genomů, což je popsáno v teorii rovnováhy genů (Veitia et al., 2008). Kromě tří výše uvedených WGD genomy některých skupin z *Brassicaceae* podlehály take dalším WGD, například celogenomová triplikace asi 8-14 mya pravděpodobně vedla k diversifikaci v rámci čeledi *Brassiceae* (Lysák *et al.*, 2007).

Předpokládalo se, že geny koordinující funkce, které organismu přinášejí nějakou selektivní výhodu, a adaptivní geny by mohly být identifikovatelné studiemi zabývajícími se rozdílnou genovou expresí napříč rostlinými druhy v heterogenním prostředí (Hoffman a Willi, 2008). Stejné studie mohou figurovat v situaci, kdy se porovnávají profily exprese blízce příbuzných druhů, jako například těch, které se vyvinuly během poslední faze třetihor při rýsování Alp na Novém Zélandě. Také by mohly pomoci při určení kandidátních genů, které měly vliv při diversifikaci druhů (Voelckel *et al.*, 2010). Při ověřování tohoto přístupu byly nejprve porovnávány expresní profily na microarrayích přirozených populace dvou blízce příbuzných alpských rostlin z převážně endemického novozélandského druhu *Pachycladon* (brukvovité). Analýzy dlouhodobých vztahů a ontologie ukázaly, že se profily sekundárních metabolitů těchto dvou populací liší. Tato informace podpořila tvrzení, že profilování transkriptomu používané při predikci může být použito jako slibný přístup při identifikování domnělých adaptivních změn mezi blízce příbuznými druhy (Voelckel *et al*., 2010).

2.4 *Pachycladon exilis*

Do australské flóry patří 20 až 21 tisíc druhů krytosemenných rostlin, kdy 90% z nich je endemických (Chapman, 2009). Rozšíření druhů přes oceán hrálo velkou roli v původu nynější australské flóry, kdy bylo zdokumentováno několik rozptýlení na velké vzdálenosti přes celou plochu Austrále až na Nový Zéland. Toto rozšíření je možné dát do souvislosti s hybridizací a polyploidií (Dierschke *et al*., 2009). Tribus *Microlepidieae* zahrnuje 17 rodů a asi 56 druhů endemických v Austrálii a na Novému Zélandě (Heenan *et al*., 2012). Na rozdíl od mnoha euroasijských druhů křížatých, bylo všech 16 australských rodů a *Pachycladon* z Nového Zélandu ve výzkumu zanedbávány a jejich fylogenetická pozice nebyla dlouhou dobu známa. První reprezentativní fylogenetická analýzy rodu *Pachycladon* odhalila jeho vznik skrze intertribální hybridizací (Joly *et al*., 2009). Dále pak byl potvrzen allopolyploidní původ *Pachycladon* pomocí srovnávací analýzy malování chromosomů (Mandáková *et al*., 2010), kdy se ukázalo, že pravděpodobně všech 11 druhů z rodu *Pachycladon* majících 10 chromosomů má stejnou strukturu, která vznikla sloučením dvou indentických nebo velmi podobných rodičovských genomů s osmi chromosomy (označené jako ancestrální karyotyp křížatých, ACK, Schranz *et al*., 2006). Původ deseti chromosomů *Pachycladon* pak byl vysvětlen jako postpolyploidní diploidizace původního mesotetraploidního genomu, dále doprovázenou postupnou diploidizací (Mandáková *et al*., 2017).

Rod *Pachycladon* se vyvinul z allopolyploidního předka před asi 2 miliony let (Joly *et al*., 2009) a během posledního milionu let se diverzifikoval do deseti morfologicky a ekologicky rozdílních druhů. Devět z těchto druhů je endemických v Jižních Alpách na Novém Zélandě a jeden se vyskytuje na Tasmánii (Heenan *et al*., 2012). Devět novozélandských druhů tvoří tři rozdílné skupiny odlišné svou genetickou charakteristikou, morfologickými vlastnostmi a preferovaným místem výskytu (Heenan a Mitchell, 2003). Tyto druhy se od sebe liší výrazně zejména jejich preferencí na určitý druh půdy. Například druhy *P. novaezelandiae* a *P. wallii* jsou omezené na půdu obsahující břidlici, zatímco další skupina skládající se z *P. fastigiatum, P. enysii* a *P. stellatum* se vyskytuje na drobové půdě. Dalšími druhy vyhledávajícími konkrétní půdu jsou *P. fasciarium* specializovaný na vápenec a *P. exilis* vyskytující se v půdách se zásaditými ionty. Zbývající dva druhy, *P. cheesemanii* a *P. latisiliquum* nemají žádná geologická specifika (Voelckel *et al*., 2010). Tyto preference odpovídají hypotéze, že adaptace na rozdílný druh půdy vedla z velké části k diverzifikaci a evoluci odlišných

druhů z rodu *Pachycladon.* Vzhledem k tomu, že druhy se liší také v rámci nadmořské výšky, můžou být i tyto preference hnací silou diverzifikace (Yogeswaran *et al*., 2010).

Pachycladon je monofyletický rod charakterizovaný malými genetickými rozdíly mezi druhy, všechny druhy se samooplodňují. U šesti druhů *Pachydladon* byla provedena analýza karyotypu a u všech druhů byl zjištěn stejný počet chromosomů 2n = 20 a genom srovnatelný s *Arabidopsis* (Joly *et al*., 2009), kdy oba tyto rody patří do stejné polyfyletického tribu *Camelineae.* Pomocí fylogenetického výzkumu byla relativně blízká příbuznost rodu *Pachycladon* a modelové rostliny *Arabidopsis thaliana* potvrzena (Heenan *et al*., 2002). Blízká příbuznost mezi těmito rody je podpořena generací pohlavně odvozené od mezidruhového hybrid mezi *A. thaliana* a *P. cheesemanii* (German *et al*., 2009). Na základě počtu chromosomů a předběžných cytogenetických dat se dále předpokládal polyploidní předek všech druhů *Pachycladon* (Lysák *et al*., 2006). Díky identifikaci dvou paralogních kopií pěti jednokopiových jaderných genů byl potvrzen allopolyplodní původ zařazený do období pleistocénu mezi asi 0,8 až 1,6 mya (Joly *et al*., 2009). Dále byla objevena asi 6 až 9 milionů let stará allopolyploidní WGD zasahující všechny australské druhy, což vedlo až k existujícím diploidním karyotypům (n = 4-6) (Mandáková *et al*., 2010). Díky podobným doposud skrytým WGD událostem, které jsou stále odhalovány pomocí komparativních genetických a cytogenetických analýz, můžeme tyto rody označit jako mesopolyploidy (Mandáková *et al*., 2010). Ačkoliv studie označují tribus *Camelineae* jako allopolyploidního předka rodů z Austrálie a Nového Zélandu, stále nebyl ani pomocí struktury genomů objasněn vztah mezi oběma polyploidními *Camelineae* skupinami (Mandáková *et al*., 2010).

Studium *Arabidopsis* a dalších příbuzných modelových rostlin vede k mnohem většímu porozumění genetických a molekulárních procesů, které se podílejí na vývoji a evoluci rostlin (Hall *et al*., 2002). Bylo zjištěno, že i *Pachycladon* má několik vlastností, které ukazují tento rod jako vhodného kandidáta pro studium molekulárních a genetických procesů, které mohou být základem pro rostlinný vývoj. Zejména jeho druhy vykazují značnou morfologickou rozmanitost. Například listy mohou být pilovité nebo laločnaté, květenství laternální nebo terminální, semena okřídlená i neokřídlená a další morfologické rozdíly mezi druhy tohoto rodu (McBreen a Heenan, 2006). S využitím výhod a informačních zdrojů, které jsou u *Arabidopsis* dostupné, a přiřozené morfologické a ekologické rozličnosti v rámci *Pachycladon,* je možné rozvinout výzkumný program mající za úkol porozumět genetickým procesům, které jsou důležité při rozšiřování rostlinných druhů na Novém Zélandě (McBreen a Heenan, 2006).

Studie čtyř druhů *Pachycladon* provedená Lysákem *et al.* (2009) je první celogenomovou analýzou druhů rozmístěných napříč celým ostrovem. Všechny tyto druhy mají deset chromosomů (Lysák *et al.*, 2009) a kromě určení intergenních chromosomálních vlastností byla provedena i analýza karyotypu (Mandáková *et al*., 2010). Obecně podobné struktury genomu podpořily předpoklad monofyletického původu celého rodu (Heenan *et al*., 2002) a umožnily odvození původního karyotypu, který zůstal zachovaný i v nynějších druzích. Stáze karyotypu u *Pachycladon* jasně ukazuje, že rozšíření druhů na Jižním ostrově na Novém Zélandu během pleistocénu nebylo spojeno s výrazným přeuspořádáním chromosomů (Heenan a Mitchell, 2003). Karyotypy čtyř zkoumaných druhů se liší pouze v počtu heterochromatických uzlů a NOR, tedy míst klíčových pro formování jadérka, kdy žádný z těchto rozdílů nebyl spojený s intragenními fylogenetickými vztahy. Proto je možné říct, že speciace probíhala skrze homoploidní divergence původního allopolyploidního genomu (Mandáková *et al*., 2010). Na základě fylogenetické analýzy australského tribu *Camelineae* a rodu *Pachycladon* (Mandáková *et al.*, 2010) autoři dospěli k závěru, že obě skupiny mohou pocházet z velmi podobného allopolyploidního předka. Toto srovnání naznačuje, že poslední kroky vedoucí ke snížení počtu chromosomů u australských druhů byly zprostředkovány tandemickými koncovými translokacemi následovanými inaktivací nebo ztrátou centromer, aniž by došlo k narušení struktury chromosomů a chromosomových ramen (Mandáková *et al*., 2010). Data napovídají, že původní karyotyp *Pachycladon* (n=10) byl allopolyploidií odvozen z ACK (n=8) (Mandáková *et al*., 2010). Nejen že se očekává, že z ACK je odvozen původní genom *Pachycladon,* také se předpokládá, že z ACK pocházejí i všechny doposud analyzované genomy rodu *Camelineae*. Jako další také karyotypy *Crucihimalaya* a *Transberingia,* dvou rodů často příbuzně spojovaných s *Pachycladon*, vykazují podobnou strukturu jako ACK. Je tedy pravděpodobné, že původní genom *Pachycladon* byl odvozen hybridizací mezi dvěma genomy podobným ACK (Heenan *et al.*, 2002, Joly *et al.*, 2009). Primární allopolyploid měl strukturu podobnou duplikovanému ACK (n=16) nebo byly zúčastněné genomy redukované před hybridizací a daný allopolyploid měl méně než 16 chromosomálních párů. Skutečnost, že paralogní genomické bloky neleží na stejných chromosomech, naznačuje, že nynější chromosomy *Pachycladon* spíše byly přeorientovány během hybridizace, než díky homologní rekombinaci mezi dvěma genomy se stejným allopolyploidním předkem (Mandáková *et al.*, 2010).

3 EXPERIMENTÁLNÍ ČÁST

3.1 Vstupní data

Hlavními datovými soubory byly již anotované genomové a proteinové sekvence *Pachycladon exilis*. Anotace byla vytvořena pomocí TriAnnot upraveného pro čeleď *Brasiccaceae*. Velikost genomu je okolo 430 Mb. Sestavení sekvence proběhlo v kombinaci Illumina WGS sekvenování, Chicago scaffolding (Dovetail genomics) a BioNano hybrid scaffolding. Délka nejdelšího scaffoldu 12 je 18,3 Mb, celková délka všech 5055 scaffoldů je 392,9 Mb a průměrná délka 588,3 Kb. 78 scaffoldů přesahuje délku 1 Mb. Anotace *Pachycladon exilis* byla provedena na 3018 scaffoldech z celkových 5055, kdy zbývající byly z anotace vyloučeny z důvodu, že jejich délka nepřesáhla 2000 bp. Na zmíněných anotovaných 3018 scaffoldech je umístěno 33556 protein kódujících genů. Nejdelší protein má 5081 aminokyselin a průměrná délka anotovaného proteinu je 421 aminokyselin. V nejdelším scaffoldu 12 je 2548 anotovaných genů s 7541 exony, které zabírají 1,5 Mb scaffoldu, a 9342 anotovaných pseudogenů, které zabírají 2,2 Mb.

Vzhledem k tomu, že geny a scaffoldy obsahovaly složité názvy na základě anotačního software, bylo provedeno kompletní přejmenování pomocí příkazu v bash shellu, tedy v unixovém příkazovém řádku.

Příkaz pro fasta soubor s proteinovou sekvencí:

```
sed '/./!d;s/\([^ ]*\) *\(.*\)/\\|\1|s||\2|g/' ID_change.txt | sed -f - AllGenes_Proteins.fa > All_Genes_newID_Proteins.fa
```

kde AllGenes_Proteins.fa je původní fasta soubor, All_Genes_newID_Proteins.fa nový soubor a ID_change.txt je vytvořený slovník obsahující původní a nové názvy.

Pro úvodní komparativní analýzu byla využita genomová a proteinová sekvence *Arabidopsis lyrata* získaná z databáze EnsemblPlants, kdy v osmi chromosomech je uspořádáno 32667 genů kódujících proteiny.

Při identifikaci rodičovského subgenomu byla dále využita sekvenační raw data *Crucihimalaya himalaica* získaná ze SRA archivu pod ID SRR3138110.

3.2 Kontrola kvality vstupních dat

Sekvenování a genomické přístupy poskytují stále rostoucí možství genomických a transkriptomických dat, avšak tato data dosahují rozdílné kvality, což vede k nutnosti důkladné kontroly kvality výsledných sekvenačních dat. Tento problém řeší například hodnotící nástroj Benchmarking Universal Single-Copy Ortholog (BUSCO) pomocí intuitivních kvantitavních měření genomických dat a jejich úplnosti za podmínky očekávaného genového obsahu. BUSCO identifikuje kompletní, duplikované, fragmenované a chybějící geny a umožňuje jejich podobnostní srovnání s rozdílnými sadami dat. (Waterhouse et al, 2017).

Kontrola kvality vstupních dat je klíčová pro zhodnocení výsledků práce, tudíž bylo nutné vstupní data zkontrolovat a určit jejich kvalitu (1. fáze celého experimentu zobrazena na obr. 4). Byl využit software BUSCO v3 (http://busco.ezlab.org) implementovaný na linuxovém serveru a analýza byla provedena pro všechny genomové sekvence *Pachycladon exilis* a pro anotované geny *P. exilis* a dále také pro sestavený transkriptom *C. himalaica*. Vzhledem k tomu, že se jedná o druhy z čeledi brukvovité, jako lineage databázi, tedy OrthoDB soubor, byl zvolen embryophyta_odb9 s *Arabidopsis* jako výchozím druhem pro analýzu.

OrthoDB (www.orthodb.org) je hierarchický katalog orthologů. Funkční anotace skupin orthologů je vytvořena pomocí služeb InterPro, GO, OMIM a fenotypů modelových organismů s dalšími referencemi na hlavní zdroje například UniProt, NCBI nebo FlyBase. OrthoDB tedy pomáhá určit vlastnosti orthologů, jako například genové duplikace a ztráty, míru divergence, příbuzné skupiny, exon-intron architekturu, syntenické orthology a fylogenetické stormy a hierarchicky klasifikuje ortology a pomocí klastrování na každé úrovni umožňuje určit druhové fylogenetické zařazení (Waterhouse et al., 2013).

OrthoDB v10 pokrývá dohromady 1271 eukaryot, 5609 bakterií, 404 archeí a 6488 virů. Celkově se jedná o 37 milionu genů, klasifikovaných do více než 8,5 milionů orientačních skupin ortologů na 624 úrovních přesnosti. Tyto úrovně odkazují na posledního společného předka z něhož se existující ortholog vyvinul a jsou definovány na základně taxonomie na NCBI. Překlad genů kódujících proteiny byl získán hlavně z RefSeq a kompletních genomů NCBI, kdy ID genomu odpovídá vyhledatelné taxonomii daného organismu (Kriventseva et al., 2019).

Příkaz pro fna soubor s genomovou sekvenci:

```
python /software/busco/3.0.2b/scripts/run_BUSCO.py -i AllGenes_CDS.fna
-o busco_genes -l embryophyta_odb9/ -m genome -c 1 -sp arabidopsis
```

kde AllGenes_CDS.fna je vstupní a busco_genes je výstupní soubor, embryopthyta_odb9 určuje lineage z OrthoDB, -sp určuje blízce příbuzný druh pro lepší práci softwaru Augustus (použitý při běhu BUSCO) a -m nastavení módu, v tomto případě tedy genomovou sekvenci.

Příkaz pro grafické zobrazení:

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

kde složka BUSCO_summaries obsahuje txt soubory, pro předchozí běh je to například short_summary_busco_genes.txt. Výchozí je pak PNG soubor s grafem. Grafické zobrazení je vytvářeno programovacím jazykem R s balíčkem ggplot2 pro vykreslení dat.

3.3 Určení podobnosti mezi sekvencemi *Pachycladon exilis* a *Arabidopsis lyrata*

Během zpracování dat byl v mnoha případech použit BLAST, jednak u samotného určení podobnosti, dále také pro vytvoření jednoho ze vstupních souborů pro běh software MCScanX. Nejdříve byla vytvořena databáze, pro proteinovou i genomovou sekvenci *A. lyrata*.

Příkaz pro vytvoření databází:

```
makeblastdb -input_type fasta -in lyr_prot.fa -dbtype prot
makeblastdb -input_type fasta -in lyr_CDS.fa -dbtype nucl
```

BLAST byl poté spouštěn s vytvořenými databázemi a nejdříve proteinovými a genomovými sekvencemi *P. exilis,* kdy parametr -b 1 určuje, že se hledal pouze jeden nejlepší hit. Pro běh MCScanX pak bylo důležité, aby výstupní blast soubor byl ve formátu m8, který specificky určuje pořadí sloupců v tabulce. V další analýze soustředěné pouze na scaffoldy 13 a 51 bylo pracováno pouze s jejich sekvencemi, jednak pro opětovný běh BLASTu, tentokrát i s možností 2 nejlepších hitů, a dále pro srovnání těchto scaffoldů, kdy byl jako databáze určen vždy právě jeden z nich.

Příkaz pro samotný běh BLASTu, první případ pro proteiny, druhý pak pro geny:

```
blastall -i All_Genes_newID_Proteins.fa -d lyr_prot.fa -p blastp -e
1e-10 -b 1 -v 1 -m8 -o lyr_pach_prot.blast
blastall -i All_Genes_newID_CDS.fa -d lyr_CDS.fa -p blastn -e 1e-10 -b
1 -v 1 -m8 -o lyr_pach_CDS.blast
```

Dále byly určeny nejlepší reciproční hity mezi proteinovými sekvencemi *A. lyrata* a *P. exilis*. Nejdříve byl spuštěn BLAST se sekvencemi *A. lyrata* jako databází a sekvencemi *P. exilis* jako dotazem, následně s opačně zvolenými sekvenčními soubory. V obou případech byl zvolen parameter -b 2, tedy přípustné vyhledávání dvou nejlepších hitů, a výstupní format souboru jako -m8. Po vytvoření těchto blast souborů byla dale provedena jejich komparativní analýza, kdy byly hledány nejlepší reciproční hity.

3.4 Komparativní analýza syntenických oblastí

Pro zjištění a vyzualizaci kolineárních a syntenických oblastí mezi *A. lyrata* a *P. exilis* v druhé fázi experiment (obr. 4) byl zvolen program MCScanX.

MCScan je algoritmus schopný porovnat genomy, případně chromosomy, za účelem identifikace předpokládané homologní chromosomové oblasti a zarovnání těchto oblastí za použití homologie a vzdálenosti genů. Sada nástrojů MCScanX implementuje a upravuje algoritmus MCScan pro detekci syntenie a kolinearity, což rozšiřuje původní software zařazením 14 užitečných programů pro vizualizaci výsledků a další downstream analýzy. Schéma znázorňující práci s tímto programovým balíkem je zobrazeno na obr. 1. MCScanX může být využit pro efektivní analýzu strukturních změn chromosomů a určení historie expanse genových rodin, která mohla vést k adaptaci druhu (Wang *et al*, 2012).

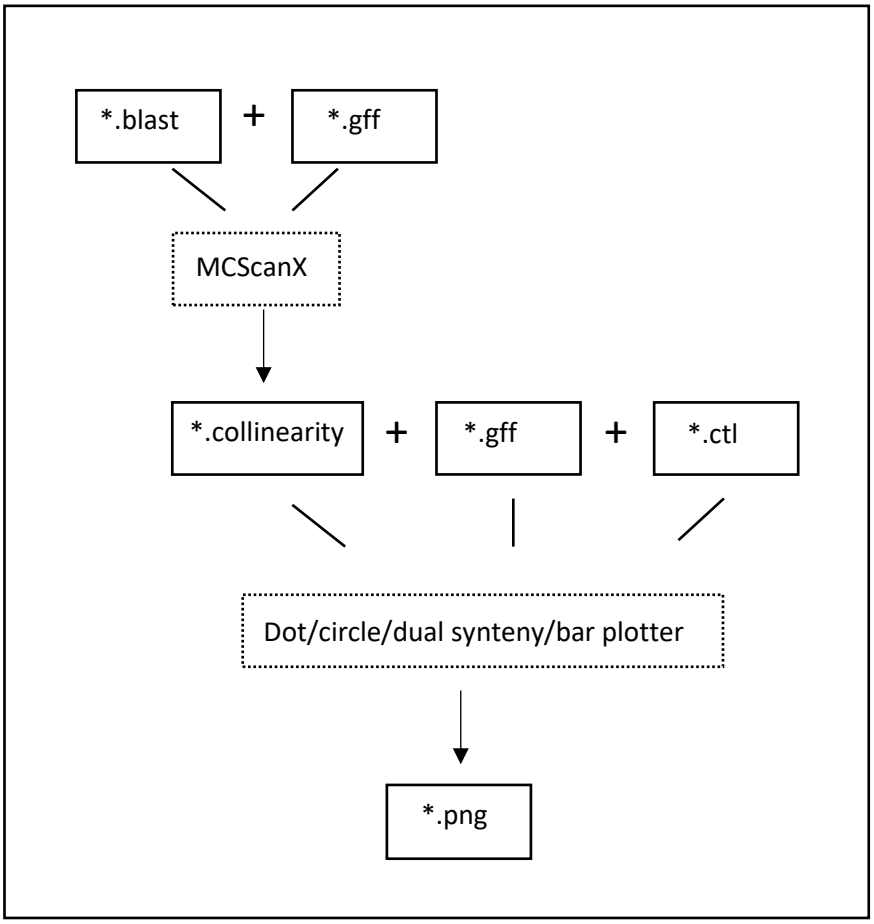
Spuštění tohoto nástroje je velmi intuitivní. V prvním kroku se na základně vstupních BLAST a GFF souborů vytvoří tzv. collinearity soubor, který je využit pro další analýzy – v případě této diplomové práce pro tvorbu dotplotu, circle a dual synteny plotu. V collinearity souboru jsou zobrazeny kolineární bloky získané párovým přiřazením genů za vstupních podmínek, v tomto případě byl minimální počet genů tvořících kolineární blok stanoven na 5.

Příkaz pro spuštění MCScanX pro proteinové sekvence:

```
MCSCanX lyr_pach_prot
```

kde vstupními soubory jsou lyr_pach_prot.blast (BLAST soubor mezi protienovými sekvencemi *A. lyrata* jako databáze a *P. exilis* jako dotaz) a lyr_pach_prot.gff (gff soubor obsahující pozice všech genů na chromosomech/scaffoldech). Výstupním souborem je pak lyr_pach_prot.collinearity.

Celogenomové výsledky z BLASTP jsou dále použity k výpočtu kolineárních bloků pro všechny potenciální páry chromosomů a scaffoldů. Nejprve jsou shody z BLASTP seřazeny podle pozic genů. V případě, že po sobě jdoucí BLASTP záznamy mají shodu ve stejném genu a tyto geny jsou odděleny méně než pěti jinými geny, tyto shody se spojí do jednoho reprezentativního genového páru s nejmenší BLASTP E-value, čímž se zabrání vysokému počtu lokálních kolineárních genových párů způsobených tandemovými oblastmi. Poté je pomocí dynamického programování nalezena cesta s největším skóre (tedy řetězce kolineárních genových párů) (Wang *et al*, 2012).



Obrázek 1: Schéma znázorňující práci s programovým balíkem MCScanX. Plnou čarou jsou ohraničeny soubory, se kterými se pracuje, případně které jsou výsledkem, tečkovanou čarou je znázorněn běh programu. *.ctl soubour je control file obsahující informace pro sestavení grafů.

Analýza pomocí programu MCSCanX byla nejprve provedena pro všech osm chromosomů *A. lyrat*a a 10 nejdelších scaffoldů *P. exilis*. Následně bylo množství nejdelších scaffoldů rozšířeno na 100 z celkových 3018, pomocí vizualizace v dot plotteru (downstream analýza, java skript jako součást balíčku stejně jako další zmíněné vyzualizační skripty) byly přiřazeny ke konkrétním chromosomům *A. lyrata* a pro lepší vyzualizaci pokrytí chromosomů byl vytvořen bar plot. Pro každý chromosom byl dále vytvořen kruhový graf, pomocí kterého je možné přehledně vyzualizovat kolineární oblasti a duplikace scaffoldů. Pro další práci byl jako reprezentativní zvolen chromosom 3 *A. lyrata* a k němu konkrétně scaffoldy 13 a 51 assembly *P. exilis*. Pro lepší vizualizaci byl vytvořen také dual synteny plot.

Příkaz pro vytvoření circle plotu pro pokrytí osmi chromosomů *A. lyrata* nejdelšími 10 scaffoldy *P. exilis*:

```
java circle_plotter -g lyr_pach_prot.gff -s lyr_pach_prot.collinearity -c circle_longest_10.ctl -o <output_PNG_file>
```

kde vstupními soubory jsou lyr_pach_prot.gff použitý pro vytvoření collinearity souboru, lyr_pach_prot.collinearity a circle_longest_10.ctl jako txt soubor obsahující informace pro vyvoření grafu. Výstupním souborem je pak png obrázek s vytvořeným kruhovým grafem.

Circle_longest_10.ctl:

```
800
al1,al2,al3,al4,al5,al6,al7,al8,Scaffold51,Scaffold28,Scaffold5,Scaffold25,Scaffold7,Scaffold21,Scaffold88,Scaffold13,Scaffold30,Scaffold12
```

Příkaz pro vytvoření dot plotu zobrazující pokrytí chromosomu 3 *A. lyrata* a scaffoldy *P. exilis*, které patří mezi 100 nejdelších scaffoldů:

```
java dot_plotter -g lyr_pach_prot.gff -s lyr_pach_prot.collinearity -c dot_al3.ctl -o <output_PNG_file>
```

kde vstupními soubory jsou lyr_pach_prot.gff použitý pro vytvoření collinearity souboru, lyr_pach_prot.collinearity a dot_al3.ctl jako txt soubor obsahující informace pro vyvoření plotu. Výstupním souborem je pak png obrázek s vytvořeným dot plotem.

Dot_al3.ctl:

```
800
800
al3
Scaffold13,Scaffold21,Scaffold5,Scaffold51,Scaffold83,Scaffold104,Scaffold108,Scaffold90
```

3.5 Analýza kolinearity vybraných scaffoldů

Pro podrobnější analýzu byly zvoleny scaffoldy 13 a 51 *Pachycladon exilis* (třetí fáze experiment, obr. 4). Nejdříve byl vytvořen BLAST mezi proteinovými sekvencemi vybraných scaffoldů a proteinovými sekvencemi *Arabidopsis lyrata,* kdy *A. lyrata* byla zvolena jako databáze a *P. exilis* jako dotaz. Databáze s proteinovými sekvencemi byla již vytvořena v předchozím kroku.

Příkaz pro spuštění BLASTu s *A. lyrata* jako databází a *P. exilis* jako dotazem:

```
blastall -i Scaffold13_51_newID.fa -d lyr_prot.fa -p blastp -e 1e-10 -
b 1 -v 1 -m8 -o lyr_13_51_prot.blast
```

Pro lepší vizualizaci situace byl spuštěn MCScanX se vstupními soubory lyr_13_51_prot.blast a upraveným gff souborem, ve kterém se nacházely pozice genů na chromosomu 3 *A. lyrata* a scaffoldů 13 a 51 *P. exilis*. Výstupní collinearity soubor byl vizualizován pomocí circle plotteru.

Následně byl proveden BLAST mezi scaffoldy 13 a 51 navzájem, kdy v prvním případě byl scaffold 13 zvolen jako databáze a scaffold 51 jako dotaz, a ve druhém případě tomu bylo naopak. Pro potřeby analýzy byl hledán pouze jeden nejlepší hit. Z výsledných blast souborů byly nalezeny nejlepší reciproční hity. Dále byla provedena analýza v programu Excel. Hlavním cílem bylo kromě kolineárních sesterských genů najít také duplikované geny, kolinearitu s anotovanými pseudogeny na sesterském scaffoldu a jedinečné geny.

Příkazy pro vytvoření databází pro BLAST:

```
makeblastdb -input_type fasta -in Scaffold13_newID.fa -dbtype prot
makeblastdb -input_type fasta -in Scaffold13_newID.fa -dbtype prot
```

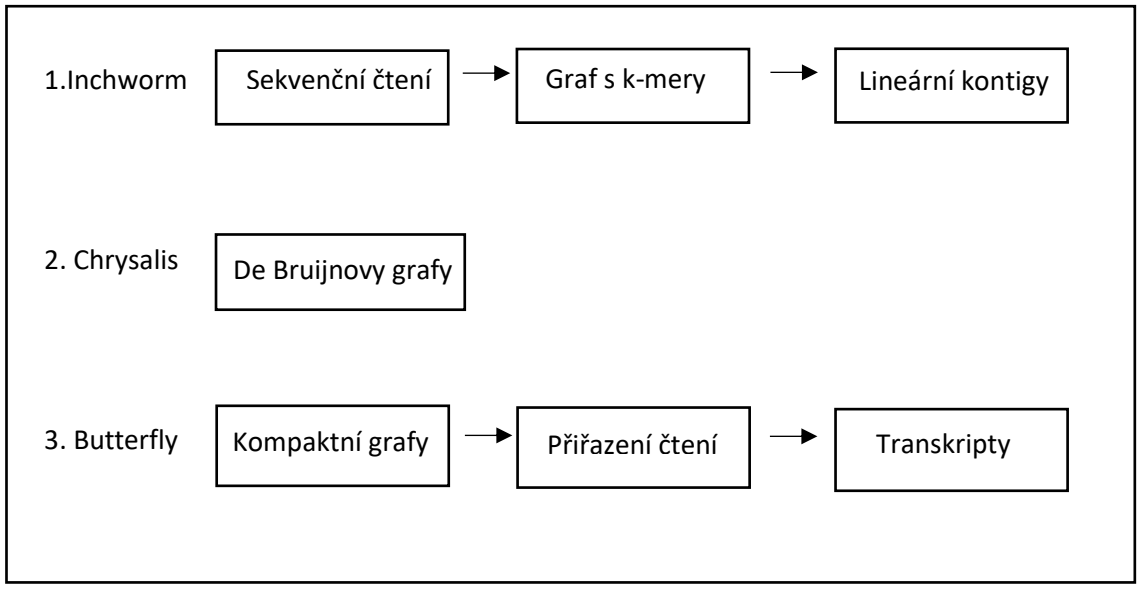
Příkaz pro spuštění BLASTu se scaffoldem 13 jako databází a scaffoldem 51 jako dotazem:

```
blastall -i Scaffold51_newID.fa -d Scaffold13_newID.fa -p blastp -e
1e-10 -b 1 -v 1 -m8 -o 13_51_prot.blast
```

3.6 Sestavení transkriptomu *Crucihimalaya himalaica*

Jako jeden z rodičovských subgenomů pro *P. exilis* se předpokládá *Crucihimalaya himalaica* a tento druh byl z tohoto důvodu take zahrnuta do této diplomové práce. Sekvenační raw data transkriptomu jsou uložena v SRA archivu pod ID SRR3138110 a pro další práci s nimi bylo potřeba vytvořit ze staženého datasetu assembly. K tomuto úkolu bylo použito softwaru Trinity.

Trinity je softwarový balíček kombinující tři moduly: Inchworm, Chrysalis a Butterfly. Na rozdíl od *de novo* assembly genomu, kde několik velkých spojitých sekvenčních grafů reprezentuje spojitosti mezi sekvenačními čteními napříč celými chromosomy, při vytváření assembly transkriptomu očekáváme četné jednotlivé nespojité grafy, kde každý představuje transkripční komplex na nepřekrývajících se místech genomu. Trinity tedy rozděluje data do těchto jednotlivých grafů a následně každý graf nezávisle zpracovává za účelem získání izoforem s celou délkou a transkriptů oddělených od paralogních genů. (Grabherr et al, 2013).



Obrázek 2: Schéma zobrazující průběh výpočtu Trinity. Průběh je rozdělen do tří fází označených jako Inchworm, Chrysalis a Butterfly.

V prvním kroku Inchworm (obr. 2, bod 1.) sestaví čtení do unikátních transkripčních sekvencí, při čemž pracuje na základě greedy přístupu vybírání k-merů, což umožňuje rychlé a účinné sestavení, při kterém se nahrazuje jediný nejlepší zástupce množinou alternativních variant se shodnými k-mery (vzhledem k alternativnímu sestřihu, duplikaci genů nebo variantám alel). Chrysalis (obr. 2, bod 2.) dále shlukuje příbuzené kontigy, které odpovídají částem transktiptům z alternativního sestřihu, případně unikátním částem paralogních genů. Poté Chrysalis sestaví de Bruijnův graf pro každý shluk příbuzných kontigů tak, že každý odpovídá složitosti překrývání mezi variantami. Nakonec Butterfly (obr. 3, bod 3.) analyzuje cestu tvořenou čteními a páry čtení v kontextu odpovídajícím de Bruijnovým grafům a uvádí všechny věrohodné sekvence transkritpu, čímž se získají isoformy vytvořené alternativním sestřihem a transkripty získané z paralogních genů (Grabherr *et al*, 2013).

Příkaz pro běh Trinity-v2.4.0:

```
Trinity.pl --seqType fq --max_memory 10G --left SRR3138110_1.fastq --right SRR3138110_2.fastq --CPU 20 --output trinity_ssr
```

kde vstupními soubory jsou SRR3138110_1.fastq a SRR3138110_2.fastq a výstupním souborem Trinity.fasta obsahující assembly. Z celkové velikosti 1091 Mb pro každý vstupní soubor byl vytvořen fasta soubor o velikosti 220 Mb za 111780 sekund, tedy asi za 31 hodin.

3.7 Přiřazení sekvencí k rodičovským subgenomům

Na závěr byla provedena komparativní analýza se sestaveným transkriptomem *Crucihimalaya himalaica* (konec 4. fáze zobrazené na obr. 4). Tato analýza byla soustředěna zejména na scaffoldy 13 a 51 *Pachycladon exilis* a jejich přiřazení k rodičovským subgenomům. Pro obecný náhled na problematiku byl vytvořen reciproční blast mezi transktiptomem *C. himalaica* a proteinovými sekvencemi *P. exilis*.

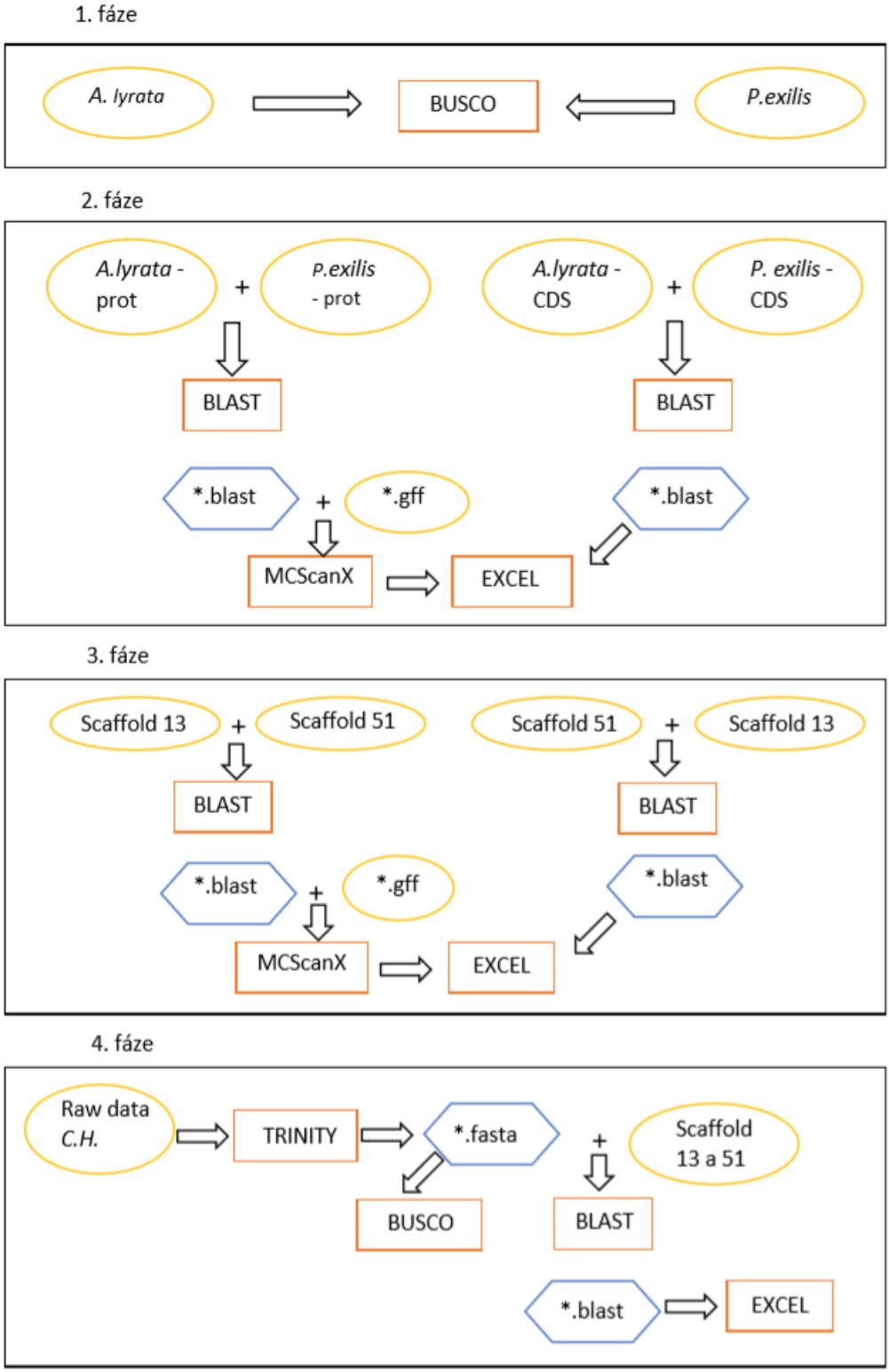
Příkazy pro běh programu BLAST:

```
blastall -i Cru_him.fa -d Scaffold13_51_newID.fa -p blastx -e 1e-10 -b 2 -v 2 -m8 -o 13_51_cru_2.blast
blastall -i Scaffold13_51_newID.fa -d Cru_him.fa -p tblastn -e 1e-10 -b 2 -v 2 -m8 -o cru_13_51_2.blast
```

kde v prvním případě je jako databáze proteinová sekvence scaffoldů 13 a 51 *P. exilis* a jako dotaz sestavený transkriptom *C. himalaica*, ve druhém případě je tomu pak opačně. V obou případech bylo parametrem -b 2 zvoleno vyhledávání dvou nejlepších hitů.

Pro zúžení pohledu na řešenou problematiku, byl BLAST spuštěn znovu pouze se sestaveným transkriptomem *C. himalaica* jako databází, scaffoldy 13 a 51 *P. exilis* jako dotazem a zúžením běhu programu pouze na 1 nejlepší hit. Při následné analýze pomocí programu Excel byly dále vybrány pouze kolineární sesterské geny *P. exilis* z daných scaffoldů a pohled byl soustředěn na průměrnou podobnost daných genů a transkriptů *C. himalaica*.

3.8 Schéma experimentální části

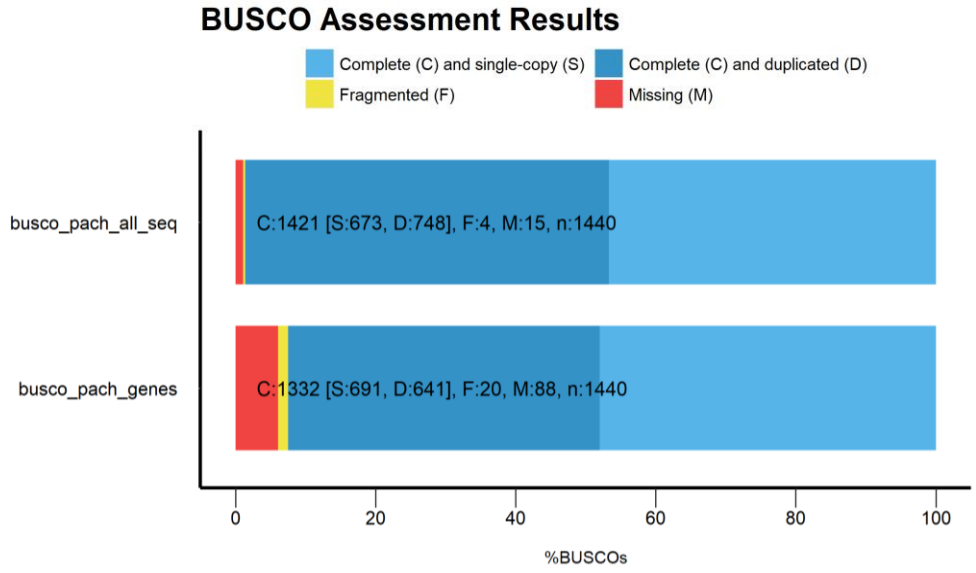


Obrázek 4: Schéma zobrazující čtyři fáze experimentální části této diplomové práce. Oranžovou barvou je označen použitý software, žlutou barvou vstupní soubory a modrou barvou výstupní soubory.

4 VÝSLEDKY A DISKUZE

4.1 Kontrola transkriptomu a anotace

Pomocí BUSCO byla testována kvalita datových souborů. Na základě 1440 sekvencí, které obsahuje datový soubor embryophyta_odb9, pak byly zjištěny hodnoty zobrazené na obr. 5. Pro fasta soubor obsahující všechny anotované geny *Pachycladon exilis* bylo ze zmíněných 1440 sekvencí 1332 kompletních. 641 z nich je duplikovaných a 691 se vyskytuje v jedné kopii. V procentech to znamená 92.5% kompletních BUSCO genů, 48.0% v jedné kopii a 44.5% duplikovaných. 88 sekvencí bylo označeno jako chybějící, avšak není vyloučeno, že by se mohly vyskytovat mezi anotovanými pseudogeny. Tomuto předpokladu odpovídají výsledky analýzy provedené na datový soubor s celým genomem, kdy jako chybějící sekvence jich bylo označeno 15 a počet kompletních BUSCO sekvencí se navýšil na 1421, tedy 98,6%. Vysoká míra duplikovaných BUSCO genů odpovídá předpokladu, že je genom *P. exilis* duplikovaný.



Obrázek 5: Grafické zobrazení vytvořené programem BUSCO, konkrétně rozšířením v programovacím jazyce R s balíčkem ggplot2. Na ose y jsou dva datové soubory, pro které bylo BUSCO spuštěno a které byly zahrnuty do grafického zobrazení –genomová sekvence *Pachycladon exilis* a anotované geny *Pachycladon exilis.* Osa x je rozdělena na procenta. V řádku u každého datového souboru jsou barevně odlišeny výsledky – modrá značí kompletní BUSCO geny (C), kdy světlejší odstín jsou sekvence kompletní a v jedné kopii (S), tmavší odstín pak duplikované (D). Žlutá barva značí fragmentované BUSCO geny (F), červená chybějící BUSCO geny (M). Na každém řádku jsou pak napsány i konkrétní počty sekvencí v každé z kategoriích.

Zajímavým tématem k diskuzi je poměr kompletních BUSCO genů mezi genomem a anotovanými geny *P. exilis*, kdy kompletních BUSCO genů u anotovaných genů je o 6,1% méně než u analýzy genomu. Tento trend je možné pozorovat i u analýz jiných rostlinných druhů. Například při studiu kokosů (*Cocos nucifera,* Xiao *et al.,* 2017) je procento kompletních BUSCO genů u assembly genomu 90,8% a u anotovaných genů potom 81,2%. Komparativní analýzu autoři prováděli společně s dalšími palmami, u kterých byl tento trend totožný. U *Phoenix dactylifera* (verze PDK30) bylo procento kompletních BUSCO genů u assembly genomu 78% a u predikovaných genů jen 56,96%. U *Elaeis guineensis* pak 84,5% kompletních BUSCO genů u assembly genomu a u predikovaných genů 42,2%. Další potvrzení tohoto trendu je možné nalézt u sestavování genomu *Rhodiola crenulate* (Fu *et al.,* 2017), kdy procento kompletních BUSCO genů u assembly genomu je 91,63% a u sady anotovaných genů 86,72%. Podobné výsledky byly získány i v rámci živočišné říše. Například u brouků *Hycleus cichorii* a *Hycleus phaleratus* (Wu *et al.,* 2018) je procento kompletních BUSCO genů u genomu 92,51% pro *H. cichorii* a 92,59% pro *H. phaleratus*. Procento kompletních BUSCO genů u sady anotovaných genů je pak 86,40% pro *H. cichorii* a 84,89% pro *H. phalaratus*. Uvedené příklady ukazují, že rozdíl mezi BUSCO výsledky u genomu a anotovanými geny není ojedinělý a je možné se s touto situací setkat i v rámci jiných studií.

4.2 Komparativní analýza s genomem *Arabidopsis lyrata*

Úvodní komparativní analýzu byla prováděna s genomovou a proteinovou sekvencí *Arabidopsis lyrata*. Sekvenční soubory byly získány z EnsemblPlants. Genom byl sestaven na Stanford Human Genome Center s pokrytím 8x a anotován na U.S. Department of Energy Joint Genome Institute (JGI), tento genom byl publikován v roce 2011 v Nature Genetics (Hu T. T. *et al.*, 2011). V tab. 1 jsou zobrazeny základní informace o velikosti a počtu genů v chromosomech *A. lyrata*.

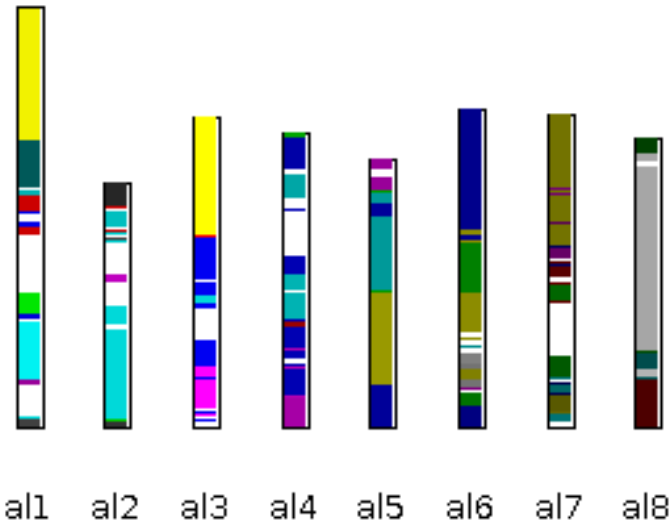
Druh *A. lyrata* byl pro úvodní komparativní analýzu zvolen z toho důvodu, že existuje jeho anotace s informacemi o pořadí genů v chromosomech. Také byla zohledněna příbuznost mezi *P. exilis* a *A. lyrata*.

Na úvod bylo provedeno hledání nejlepších recipročních hitů mezi proteinovými sekvencemi *A. lyrata* a *P. exilis*. Z důvodu duplikace genomu *P. exilis* byly vytvořeny vstupní BLAST soubory s parametrem dvou nejlepších hitů. Při zohledňování nejvyššího bit skóre bylo nalezeno 18236 nejlepších recipročních hitů, při zohledňování nejnižší e-hodnoty pak 18280. Celkem bylo přiřazeno 65484 hitů s *A. lyratou* jako databází a *P. exilis* jako dotazem, v opačném případě pak 56204 hitů.

Tabulka 1: Informace o chromosomech *A. lyrata*. Tabulka zobrazuje počet genů a komulativní délku CDS v daném chromosomu a jeho velikost.

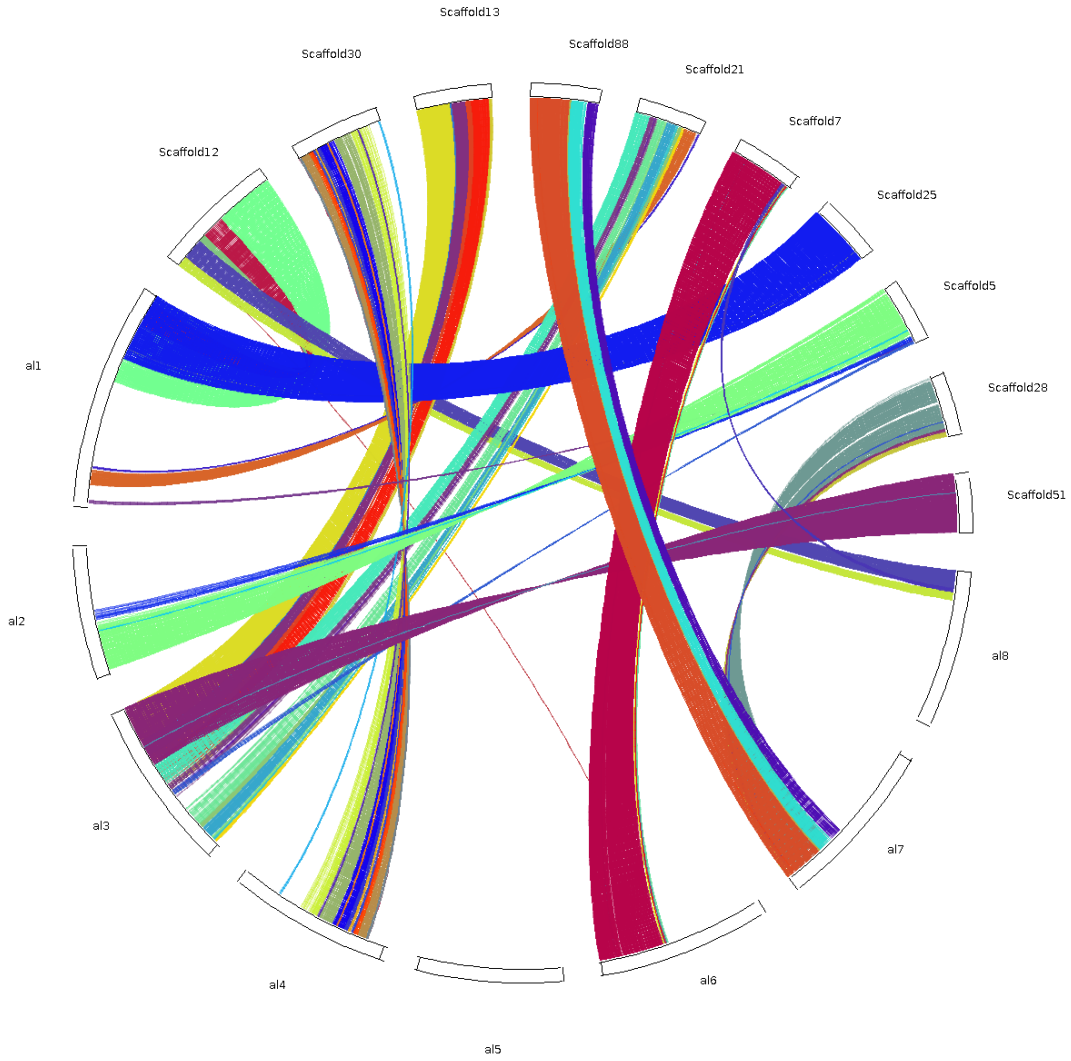
	Velikost	Počet genů	Kumulativní délka CDS
Chromosom 1	34 Mb	5380	6013 kb
Chromosom 2	20 Mb	3006	3257 kb
Chromosom 3	25 Mb	4216	4552 kb
Chromosom 4	24 Mb	3640	3836 kb
Chromosom 5	21 Mb	3470	3686 kb
Chromosom 6	25 Mb	4177	4684 kb
Chromosom 7	25 Mb	4113	4588 kb
Chromosom 8	24 Mb	3476	3768 kb
Celkem	198 Mb	32667	34,4 Mb

K analýze pomocí programu MCScanX byly jako vstupní soubor použity proteinové sekvence, konkrétně vytvořený BLAST soubor ve formátu m8 pomocí blastp pro proteinové sekvence *A. lyrata* a *P. exilis*. *A. lyrata* byla zvolena jako databáze a *P. exilis* jako dotaz. Vyhledávání bylo zúženo pouze na jeden nejlepší hit. Druhým vstupním souborem pak byl gff soubor upravený pro potřeby MCScanX, tedy soubor obsahující pozice všech genů rozdělené podle chromosomů a scaffoldů. Výchozí nastavení MCScanX pracuje s hodnotami match_score 50, gap_penalty -1 (hodnoty využívané při výpočtu hodnoty řetězce párů kolineárních genů, kdy vybrán je ten s nejvyšší celkovou hodnotou) a E-value 1e-05 (pravděpodobnost výskytu genů v kolineárním bloku). Pro nejdelší ze scaffoldů, scaffold 12, byl vytvořeno celkem 7 kolineárních bloků dohromady s 2215 páry genů na chromosomech 1, 6 a 8 *A. lyrata*, jak v plus, tak i v minus směru. Celkově bylo vytvořeno 274 kolineárních bloků s 28153 kolineárními páry. Na obr. 6 je možné pozorovat pokrytí osmi chromosomů *A. lyrata* 100 nejdelšími scaffoldy *P. exilis*, kdy bílá místa zobrazují nepokryté části chromosomů.

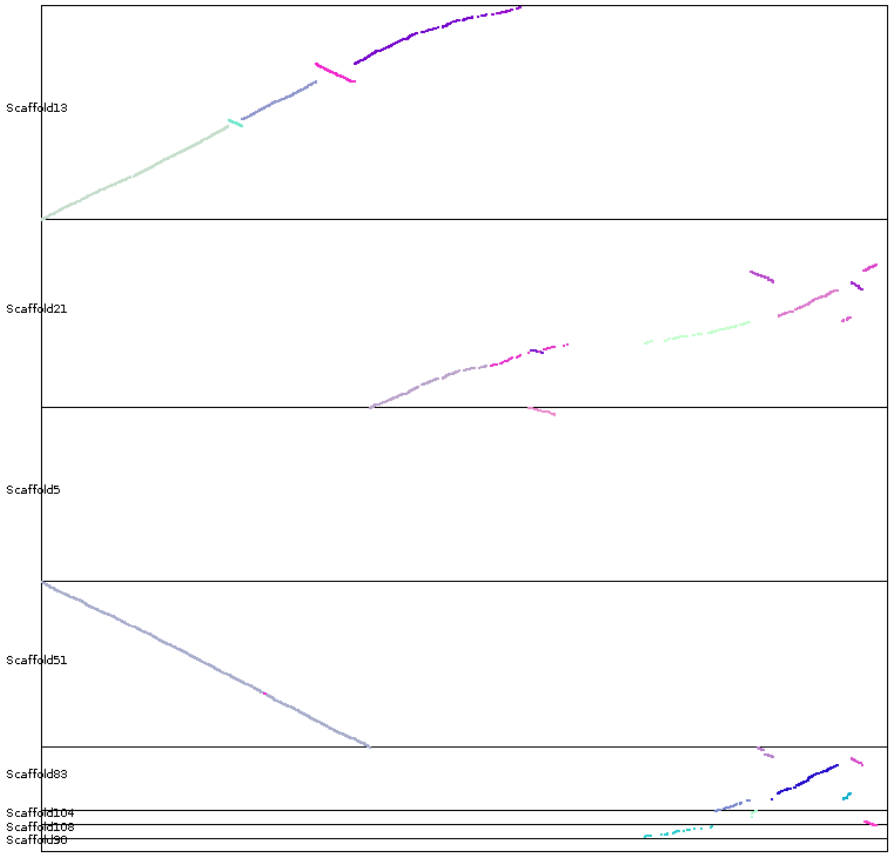


Obrázek 6: Sloupcový graf vytvořený programem MCScanX, konkrétně java skriptem bar plotter jako součást downstream analýzy. Jedná se pouze o část výsledného grafu. Analýza byla provedena pro genomovou *A. lyrata* a *P. exilis* a daný graf zobrazuje rozložení 100 nejdelších scaffoldů *P. exilis* na osmi chromosomech *A. lyrata* (al1-al8).

Pro podrobnější pohled na pokrytí *A. lyrata* byl zvolen circle plot (obr. 7) zobrazující pouze 10 nejdelších scaffoldů *P. exilis* a jen z této části sekvencí *P. exilis* je možné vidět jisté trendy, například že scaffold 25 je celý kolineární s částí chromosomu 1 nebo že chromosom 3 je téměř celý pokryt jen pomocí scaffoldů 13, 21, 5 a 51, kdy scaffold 51 je kolineární opět celý. Zajímavostí je, že chromosom 5 není kolineární s žádným z 10 nejdelších scaffoldů, avšak jak je vidět na obr. 6 mezi dalšími ze 100 nejdelších scaffoldů již kolinearita je. Dále je možné pozorovat duplikace kolineárních bloků, například u chromosomu 1, 3 nebo 7.



Obrázek 7: Kruhový graf vytvořený programem MCScanX, konkrétně java skriptem circle plotterem jakou součást downstream analýzy. Jedná se o zobrazení pokrytí osmi chromosomů *A. lyrata* (al1-al8) deseti nejdelšími scaffoldy *P. exilis* (Scaffold51, Scaffold28, Scaffold5, Scaffold25, Scaffold7, Scaffold21, Scaffold88, Scaffold13, Scaffold30, Scaffold12).

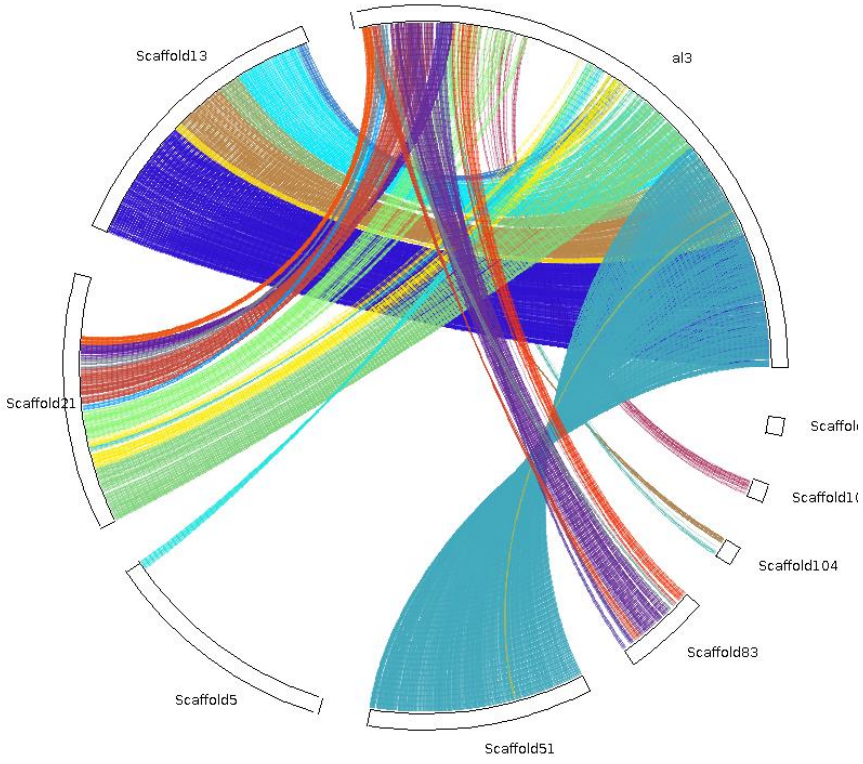


Obrázek 8: Dot plot vytvořený programem MCScanX, konkrétně java skriptem dot plotterem jakou součást downstream analýzy. Jedná se o pokrytí chromosomu 3 *A. lyrata* scaffoldy *P. exilis*. Konkrétně se jedná o scaffoldy číslo 13, 21, 5, 51, 83, 104, 108 a 90, které patří mezi 100 nejdelších scaffoldu *P. exilis*.

Pro podrobnější analýzu byl zvolen chromosom 3 z toho důvodu, že je možné vidět, že jeden úsek chromosomu pokrývají hned dva scaffoldy, konkrétně scaffold 13 a scaffold 51 (obr. 7 a 9). Na dot plotu (obr. 8) je možné pozorovat nejen kolineární bloky u daných scaffoldů, ale take jejich směr. Z celkového počtu 4216 proteinů kódovaných anotovanými geny u *A. lyrata* v chromosomu 3 bylo vytvořeno se všemi proteinovými sekvencemi *P. exilis* 3924 kolineárních párů v 27 kolineárních blocích dohromady v obou směrech. Vzhledem k tomu, že jedním z cílem práce bylo rozlišit rodičovské subgenomy *P. exilis*, byly zvoleny scaffoldy 13 a 51 jako nejpříhodnější pro další analýzu. Na obr. 9 je jasně vidět, že tyto dva scaffoldy tvoří kolineární bloky se stejným úsekem chromosomu 3 *A. lyrata* a tudíž je možné se domnívat, že se jedná o duplikované oblasti genomu *P. exilis*, kdy každý ze scaffoldů patří do jednoho ze subgenomů. Scaffold 51 tvoří dva kolineární bloky, jeden v plusovém směru sekvence chromosomu 3 s 1392 kolineárními páry a druhý v opačném směru sekvence chromosomu 3 s 10 kolineárními páry. Scaffold 13 tvoří 6 kolineárních bloků s chromosomem 3, ve směru sekvence *A.*

lyrata jsou to bloky s 122 a 46 kolineárními páry a v opačném směru pak s 777, 276, 238 a 35 kolineárními páry. Na obr. 9 je možné pozorovat, že část scaffoldu 13 se překrývá take se scaffoldem 21, který plynule navazuje na scaffold 51 spolu s částí scaffoldu 5. Na grafu je také možné pozorovat barevně odlišené kolineární bloky v rámci jednotlivých scaffoldů.

Vzhledem ke kolineárním blokům je možné předpokládat, že některé ze scaffoldů ve skutečnosti tvoří větší oblasti a některé ze scaffoldů je možné rozdělit a jejich části přiřadit k jiným. Například již zmiňovaný scaffold 51 a část scaffoldu 21 spolu s částí scaffoldu 5. Vzhledem k pouze drobné oblasti ve scaffoldu 5, která tvoří kolineární blok s chromosomem 3 *A. lyrata*, by bylo vhodné pro efektivitu komparativní analýzy tento scaffold rozdělit. Další z možností viděných na obr. 9 také může být propojení scaffoldů 104, 108, 83 a také konec kolineárního bloku na scaffoldu 21. Opět je možné pozorovat, že část scaffoldu 21 není kolineární s chromosomem 3 *A. lyrata* a je tedy možné, že by bylo možné tuto oblast propojit s jiným kolineárním blokem u jiného chromosomu.



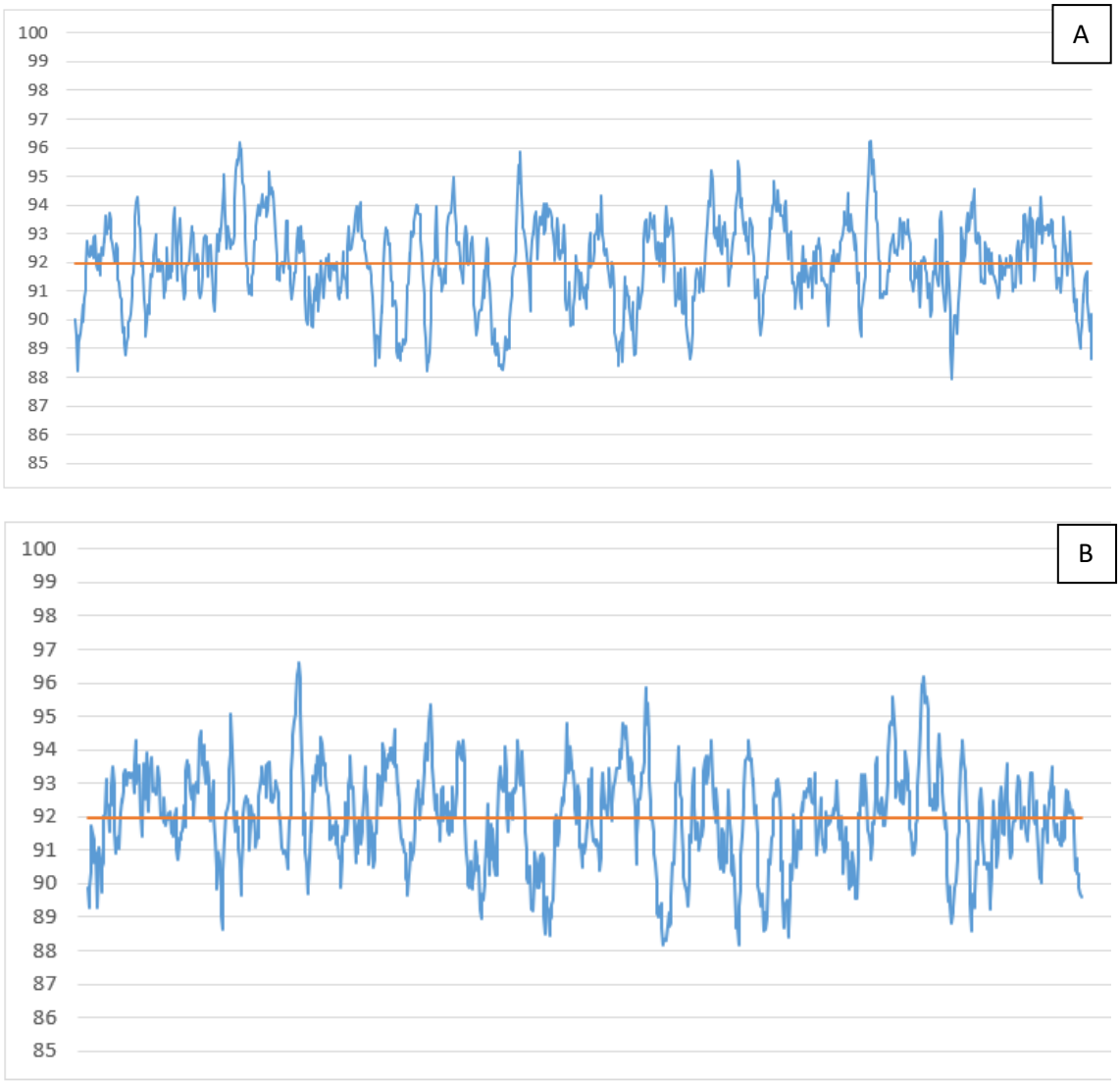
Obrázek 9: Kruhový graf vytvořený programem MCScanX, konkrétně java skriptem circle plotter jakou součást downstream analýzy. Jedná se o přehlednější zpracování obr. 8, tedy dot plotu pokrytí chromosomu 3 *A. lyrata* scaffoldy *P. exilis*. Konkrétně se jedná o scaffoldy číslo 13, 21, 5, 51, 83, 104, 108 a 90, které patří mezi 100 nejdelších scaffoldu *P. exilis*.

4.3 Srovnání scaffoldů 13 a 51 *P. exilis*

Pro podrobnější srovnání podobnosti a kolinearity byl nejdříve pro oba scaffoldy spuštěn BLAST, konkrétně blastn s vyhledáváním dvou nejlepších hitů. V prvním případě byl jako databáze scaffold 13 a scaffold 51 jako dotaz, ve druhém případě tomu bylo naopak. V prvním případě bylo nalezeno 1050 hitů s podobností nad 80%, v druhém případě 1063. V případě omezení se na hity s podobností vyšší než 80%, průměrná podobnost pro první případ je 91,97% a pro druhý případ je 91,98%. Na obr. 10 je možné pozorovat kolísání průměru podobnosti po sobě jdoucích deseti hitů a také křivku průměrné podobnosti. Tab. 2 popisuje základní informace o obou scaffoldech. Při bližším porovnání hitů bylo určeno u 1144 genů nejlepší reciproční hit mezi scaffoldy 13 a 51, tedy tyto geny je možné označit jako sesterské. Průměrná podobnost mezi těmito sesterskými geny je 89,9%, kdy pokles byl způsoben zohledněním nejlepších recipročních hitů s podobností menší než 80%. U 41 genů ve scaffoldu 13 a 42 genů ve scaffoldu 51 byla nalezena další významná podobnost s geny, které byly již zařazeny mezi nejlepší reciproční hity, tudíž je možné předpokládat, že se jedná o duplikace existujících genů v daném scaffoldu.

Tabulka 2: Srovnání scaffoldů 13 a 51 *P. exilis*. První informací je celková délka scaffoldů, dale jsou zohledněny do kategorií exony, geny a pseudogeny.

	13	51
Celková délka	11,3 Mb	8,7 Mb
Počet genů	1716	1573
Celková délka genů	3,9 Mb	3,7 Mb
Průměrná délka genu	2279,4 b	2353 b
Medián	1868 b	1876 b
Nejdelší gen	17,63 kb	21,5 kb
Nejkratší gen	89 b	113 b
Počet exonů	5617	5081
Celková délka exonů	1095,7 kb	984,1 kb
Počet pseudogenů	328	277
Celková délka pseudogenů	1328 kb	521,5 kb
Průměrná délka pseudogenu	2424,8 b	1883 b
Nejdelší pseudogen	21 kb	20 kb
Nejkratší pseudogen	146 b	143 b



Obrázek 10: Grafy zobrazující průměr podobnosti deseti po sobě jdoucími geny mezi scaffoldy 13 a 51 *P. exilis*. Horní graf (A) zobrazuje případ, kdy pro BLAST byl jako databáze zvolen scaffold 13 a dotaz scaffold 51, spodní graf (B) ukazuje opačný případ. Konstantní oranžová řada značí průměr podobnosti napříč celými scaffoldy.

Jako vstupní soubory pro MCScanX byl použit již vytvořený BLAST soubor mezi genovými sekvencemi scaffoldů 13 a 51, kdy scaffold 13 byl nadefinován jako databáze a scaffold 51 jako dotaz, a GFF soubor obsahující pozice všech anotovaných genů ve scaffoldech 13 a 51. MCScanX při výchozím nastavení vytvořil kolineární páry s 2276 geny (tedy dohromady 1138 kolineárních párů) z celkových 3289 hitů ve vstupním BLAST souboru. Kolineárních bloků bylo vytvořeno 6, s délkami 111, 41 a 9 ve shodném směru se scaffoldem 13 a o délkách 692, 238 a 48 v opačném směru.

Při použití vstupního BLAST souboru s opačným nastavením vstupních souborů, tedy se scaffoldem 51 jako databází a scaffoldem 13 jako dotazem, byl vytvořen collinearity soubor s 2306 kolineárními geny (tedy celkově 1153 kolineárních párů) z celkových 3289 hitů ve vstupním BLAST souboru. Celkem bylo vytvořeno 6 kolineárních bloků o délkách 114, 41 a 9 ve shodném směru se scaffoldem 51 a o délkách 698, 243 a 49 v opačném směru. Rozdíl je způsoben pravděpodobně duplikovanými oblastmi v rámci scaffoldů. Obecně kolineární páry odpovídají předpokladům z dříve zmíněné BLAST analýzy.

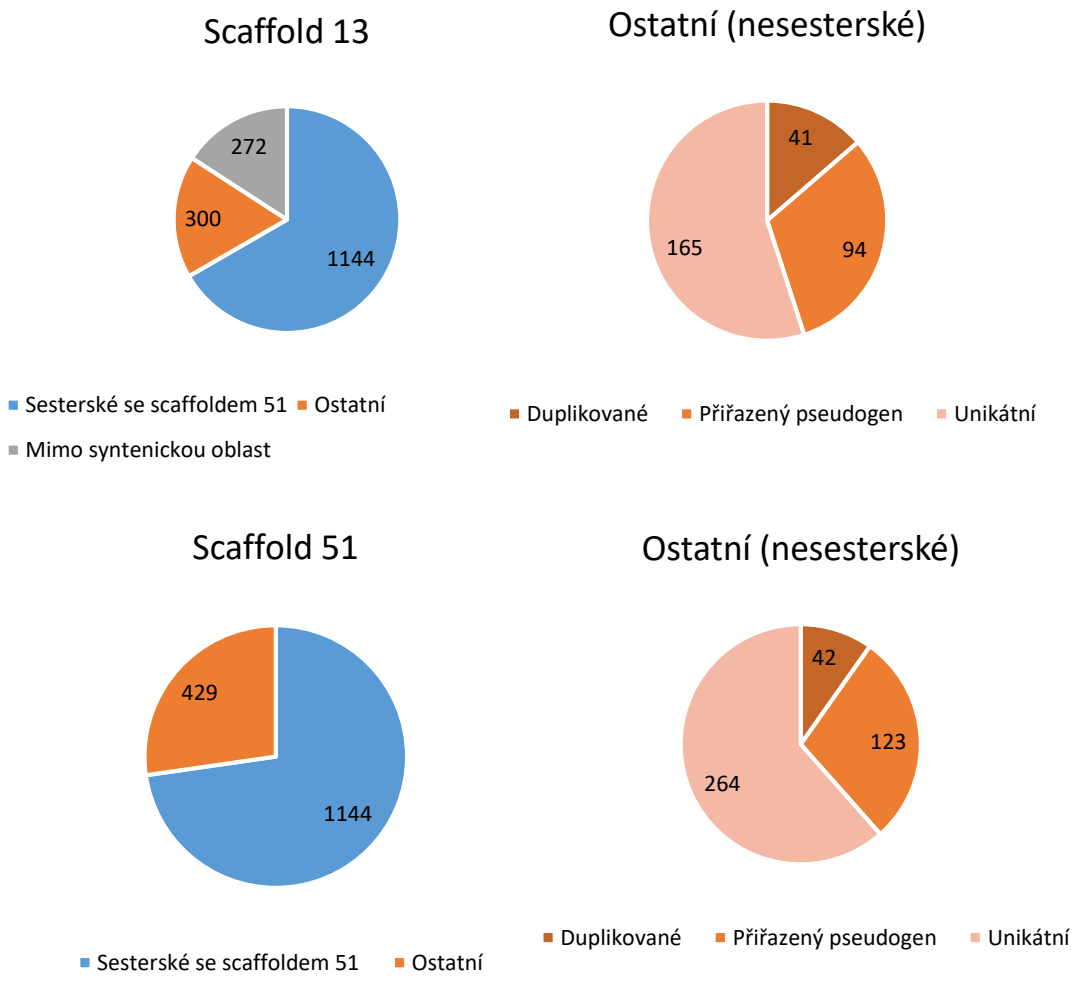
Vzhledem k tomu, že bylo určeno 1144 sesterských dvojic mezi scaffoldy 13 a 51 (tedy 1144 dvojic se vzájemným nejlepším recipročním hitem), hodnoty 2306 a 2276 kolineárních párů souhlasí se skutečností, že některé geny jsou duplikované a tedy se tvoří u některých genů násobné dvojice. Na obr. 11 je graficky zobrazena kolinearita mezi scaffoldy 13 a 51. Je možné pozorovat barevně odlišené kolineární bloky i genový úsek na začátku scaffoldu 13, který není se scaffoldem 51 kolineární. Na některých místech jsou také viditelné inverze celých kolineárních bloků.

Obrázek 11: Graf vytvořený programem MCScanX, konkrétně java skriptem dual synteny plotter jakou součást downstream analýzy. Jedná se o zobrazení kolineárních bloků mezi Scaffoldem 13 (vlevo) a scaffoldem 51 (vpravo) *P. exilis*.

4.4 Pseudogeny, duplikované a jedinečné geny *P. exilis*

Po vyloučení jednoznačně sesterských genů potvrzených nejlepším recipročním hitem se ve scaffoldech 13 a 51 vyskytují další geny, které můžeme rozdělit do několika kategorií, konkrétně duplikované geny, jedinečné geny a geny, jejichž homolog se stal pseudogenem, jak je možné vidět na koláčových grafech na obr. 12.

Scaffold 13 obsahuje celkem 1716 genů, kdy prvních 272 je z analýzy vyloučeno vzhledem k tomu, že se vyskytují mimo syntenickou oblast se scaffoldem 51. Ze zbývajících 1444 genů je 1144 sesterských, tedy kolineárních, s geny ve scaffoldu 51. Na 300 nesesterských genů jsem provedla další analýzu, kdy bylo zjištěno, že 41 genů je duplikovaných k některému z genů, který má sesterskou dvojici ve scaffoldu 51. K dalším 94 genům byla nalezena sesterská shoda mezi anotovanými pseudogeny ve scaffoldu 51. Zbylých 165 genů je tedy možno označit jako unikátní ve scaffoldu 13.



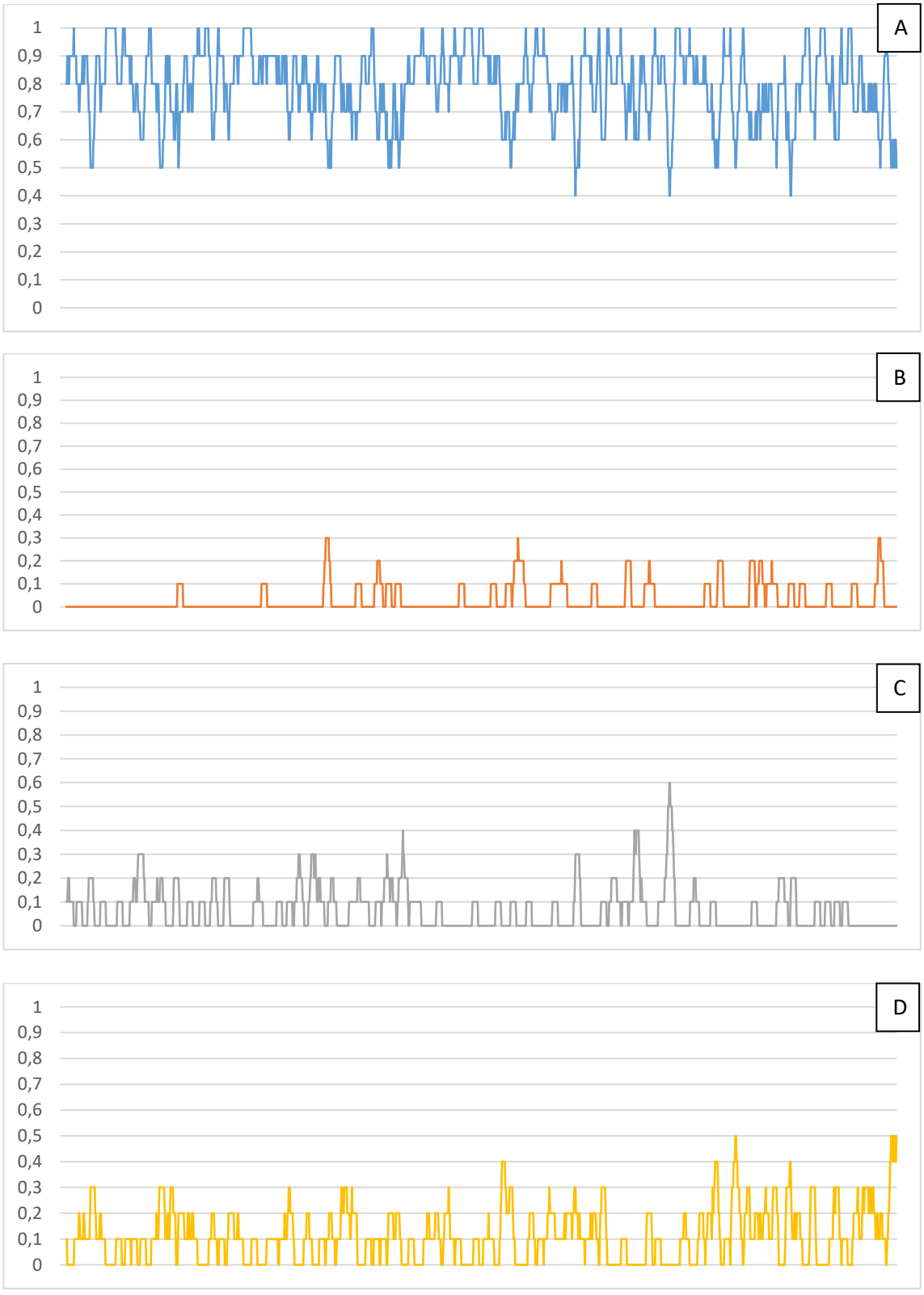
Obrázek 12: Koláčové grafy zobrazují rozložení genů ve scaffoldu 13 (horní dvojice) a scaffoldu 51 (spodní dvojice).

Scaffold 51 obsahuje celkem 1573 genů, kdy 1144 genů bylo označeno jako sesterské, tedy kolineární, s geny ve scaffoldu 13. Ze zbývajících 429 genů je jich 42 duplikovaných k některému z genů, který má sesterskou dvojici ve scaffoldu 13. K dalším 123 genům byla nalezena shoda mezi anotováními pseudogeny ve scaffoldu 13. Zbylých 264 můžeme označit jako unikátní.

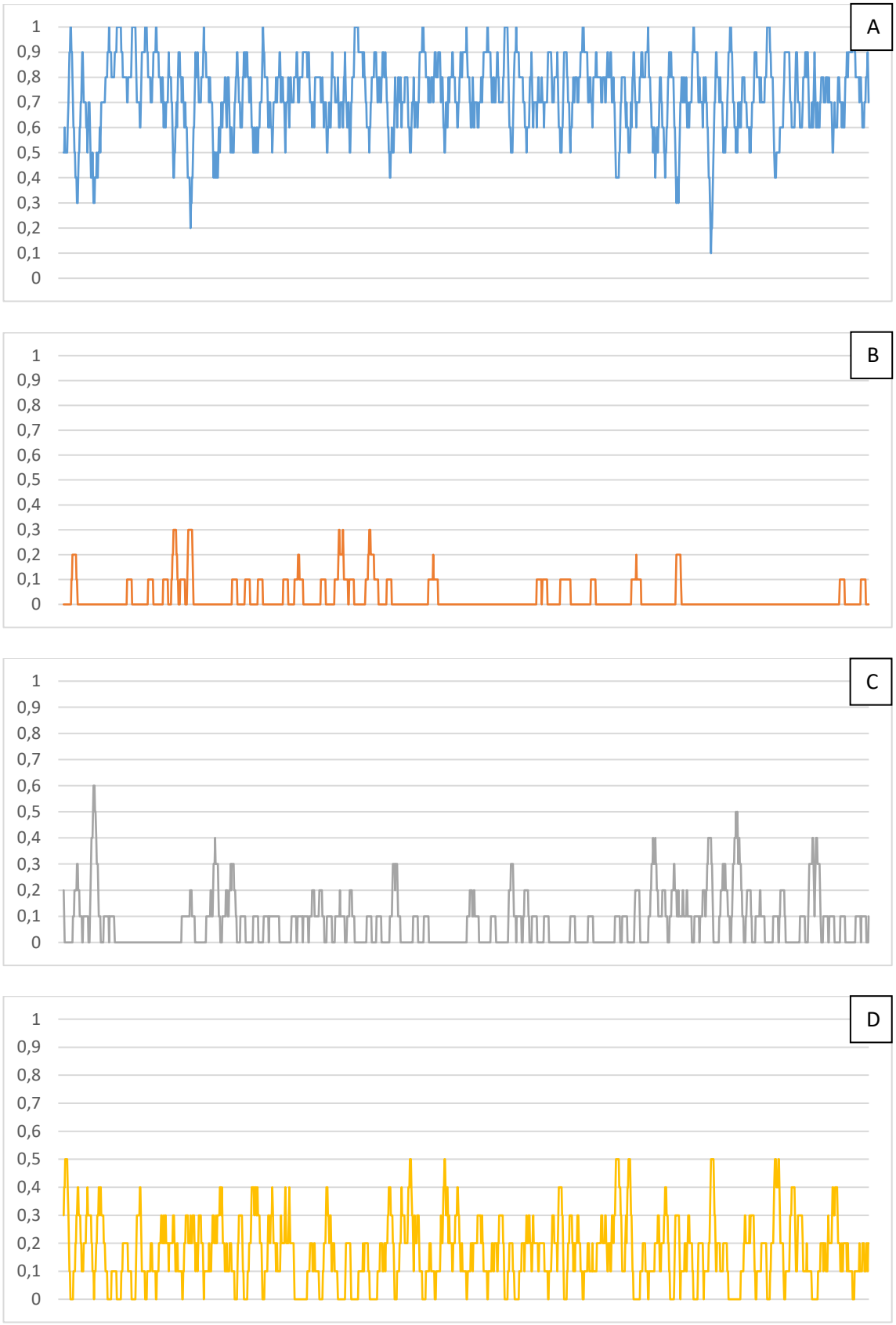
Je vidět, že scaffold 51 obsahuje větší množství unikátních genů než scaffold 13. Vzhledem k tomu, že genom *P. exilis* je duplikovaný, je možné předpokládat, že v rámci evoluce u subgenomu obsahujícím scaffold 13 došlo ke genetické erozi. Tento termín vysvětluje ztrátu genetické variace, například ztrátu alel určujících hodnotu specifického znaku nebo celého souboru znaků (Fasola *et al*., 2015). Genetická eroze pouze u jednoho ze subgenomů byla pozorována i u dalších polyploidních rostlin, například u *Arabidopsis thaliana* (Thomas *et al.,* 2006), *Brassica rapa* (Wang *et al*., 2011) nebo kukuřice (*Zea mays,* Schnable *et al.,* 2011). U kukuřice se pak dokonce dvě třetiny genů, které mají u rýže nebo čiroku známého ortologa, vyskytují samostatně, tedy byl zachován pouze jeden ze dvou duplikovaných genů (Schnable *et al.,* 2011). Mimo genetické eroze je pak další možností transpozice genů, tedy ztráta genu na konkrétní pozici a naopak přesun na pozici jinou (Zhao *et al*., 2017). Tento případ byl popsán jak u rýže (*Oryza sativa,* Tarchini *et al.,* 2000), tak i v rámci čeledi brukvovitých například u *A. thaliana* (Freeling *et al.,* 2008).

Průměrná hustota rozložení unikátních, kolineárních, duplikovaných genů a pseudogenů je zobrazena na obr. 13 pro scaffold 13 a na obr. 14 pro scaffold 51. U obou scaffoldů je možné pozorovat trend, kdy unikátní i kolineární geny jsou pravidelně rozprostřené napříč celými scaffoldy, i když hustoty jejich výskytu jsou na opačných stranách spektra. Průměrný výskyt kolineárních genů ve scaffoldu 13 je 79,3 % a ve scaffoldu 51 je 72,7 %. Průměrný výskyt unikátních genů ve scaffoldu 13 je 11,4 % a ve scaffoldu 51 je 16,8 %. Naopak duplikované oblasti se vyskytují pouze v některých částech, sice obecně s malou hustotou výskytu. Je možné pozorovat, že u scaffoldu 13 se vyskytují spíše v druhé půlce syntenické oblasti, u scaffoldu 51 pak zejména na začátku. Nejvyšší hustoty duplikované oblasti ve scaffoldu 13 se vyskytuje zejména v úseku kolem genů s označením Scaffold13_gene_07350 a Scaffold13_gene_16780 s hodnotami 0,3, tedy tři duplikované geny na deset po sobě jdoucích genů. U scaffoldu 51 pak v úseku kolem genů s označením Scaffold51_gene_02260, Scaffold51_gene_02560 a Scaffold13_gene_06050 opět s hodnotami 0,3. Scaffoldy doposud nejsou uspořádány do chromosomu a není tedy známo konkrétní pořadí a orientace genů v chromosomech. Je tedy pravděpodobné, že jeden ze scaffoldů ve výsledném uspořádání bude mít inverzní

pořadí genů oproti tomu, jak je nyní anotováno. V tomto případě by se pak duplikované geny vyskytovaly ve stejných oblastech chromosomů. Úseky bez zaznamenaných duplikací jsou u scafflodu 13 i 51 napříč celými scaffoldy a jejich výskyt je spíše ojedinělý. Obecně hustota výskytu duplikovaných genů u scaffoldu 13 je 2,8 %, u scaffoldu 51 pak 2,7 %. Geny, k nimž byl nalezen reciproční hit s pseudogenem v opačném scaffoldu, tedy poslední skupina genů, na které byla zaměřena podrobná analýza, se vyskytují s průměrnou hustotou 6,5 % ve scaffoldu 13 a 7,8 % ve scaffoldu 51. Obecně byly úseky s těmito geny identifikovány napříč celými scaffoldy, avšak v některých konkrétních oblastech se nevyskytují vůbec. Například u scaffoldu 13 je to zejména konec scaffoldu. U scaffoldu 51 je možné již na grafu pozorovat hned dvě signifikantní oblasti bez výskytu takových genů, mezi geny Scaffold51_gene_ 01000 a Scaffold51_gene_ 02380 a mezi geny Scaffold51_gene_ 07100 a Scaffold51_gene_ 07930. Naopak velmi signifikantní výskyt u scaffoldu 13 je v úseku kolem genu s označením Scaffold13_gene_13260, kde hustota kolísá kolem hodnoty 0,5 a 0,6. Vzhledem k dříve zmíněnému předpokladu, že existuje možnost reálné inverze orientace scaffoldů, zmíněné oblasti bez existujících pseudogenů by se opět vyskytovaly v podobných oblastech obou scaffoldů. V případě, že by některé z genů byly chybně anotovány jako pseudogeny, je možné, že by se oblasti scaffoldů bez pseudogenů vyskytovaly ve větší míře, než je tomu nyní.



Obrázek 13: Graf zobrazuje výskyt genů napříč scaffoldem 13. Existence byla označena jako 1, absence jako 0, graf pak zobrazuje průměr po sobě deseti jdoucích genů. Křivky zohledňují rozložení kolineárních genů se scaffoldem 51 (A), duplikovaných genů (B), pseudogenů (C) a unikátních genů (D) na scaffoldu 13.

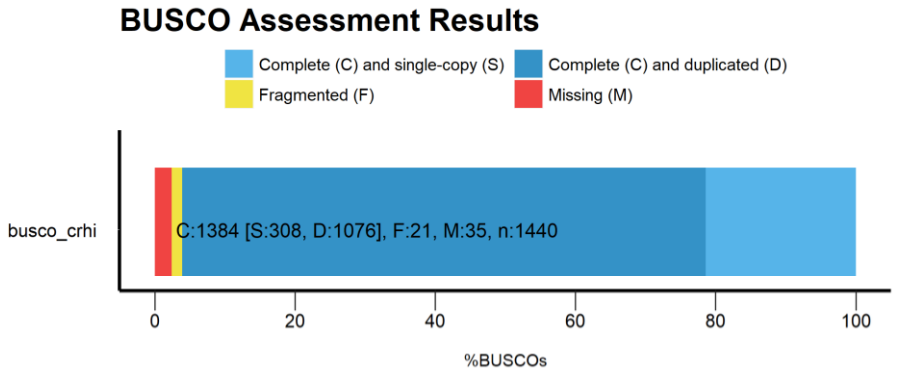


Obrázek 14: Graf zobrazuje výskyt genů napříč scaffoldem 51. Existence byla označena jako 1, absence jako 0, graf pak zobrazuje průměr po sobě deseti jdoucích genů. Křivky zohledňují rozložení kolineárních genů se scaffoldem 13 (A), duplikovaných genů (B), pseudogenů (C) a unikátních genů (D) na scaffoldu 51.

4.5 Komparativní analýza s transkriptomem *Crucihimalaya himalaica*

Vzhledem k předpokladu, že jeden ze subgenomů *Pachycladon exilis* evolučně pochází z genomu *Crucihimalaya himalaica*, byla provedena podrobější analýza vztahu mezi genomy těchto rostlin.

Kontrola kvality byla také byla provedena na sestaveném transkriptomu *Crucihimalaya himalaica*. Grafický výstup je na obr. 15. Počet kompletních BUSCO sekvencí je 1384 (96,1%), z toho 308 (21,4%) v jedné kopii a 1076 (74,7%) duplikovaných, a 35 chybějících BUSCO sekvencí. Je předpokládáno, že tak vysoké číslo u duplikovaných sekvencí bylo způsobeno sestavením transkriptomu, tedy zpracováním dat ze SRA archivu pomocí software Trinity. Autoři sekvenačních dat v SRA archivu také prováděli sestavení transkriptomu pomocí programu Trinity (Qiao Q. *et al.*, 2016) pouze s tím rozdílem, že nadefinovali minimální pokrytí k-meru na 2, v případě této diplomové práce bylo pracováno s výchozím nastavením, tedy hodnotou 1. V článku uvádějí jako výsledek 66084 transkriptů, 49438 unigenes (nejdelší transcript v jednom genu), z nichž 39189 anotovaných se signifikantní shodou. Transkriptom, který byl vytvořen pomocí Trinity v případě této diplomové práce a dále s ním bylo pracováno, obsahuje 117606 transkriptů. Nově vytvořené genomové assembly publikované v článku (Zhang T., *et al.*, 2019) z roku 2019 pak obsahuje 29420 transkriptů sestavených pomocí Trinity s 96% kompletními BUSCO geny, z toho 16% duplikovanými, kdy 26806 genů bylo funkčně anotovaných. Zmíněné výsledky odpovídají tomu, že podle hodnot BUSCO soubor transkriptomu zpracovávaný v této diplomové práci obsahuje 74,7% duplikovaných sekvencí. Tyto duplikace pak pravděpodobně jsou variantami transkriptů téhož genu.



Obrázek 15: Grafické zobrazení vytvořené programem BUSCO, konkrétně rozšířením v programovacím jazyce R s balíčkem ggplot2. Na ose y se nachází transkriptom *Crucihimalaya himalaica*. Osa x je rozdělena na procenta. Barevně jsou odlišeny přiřazené BUSCO geny, kdy každá kategorii je popsána i konkrétní hodnotou.

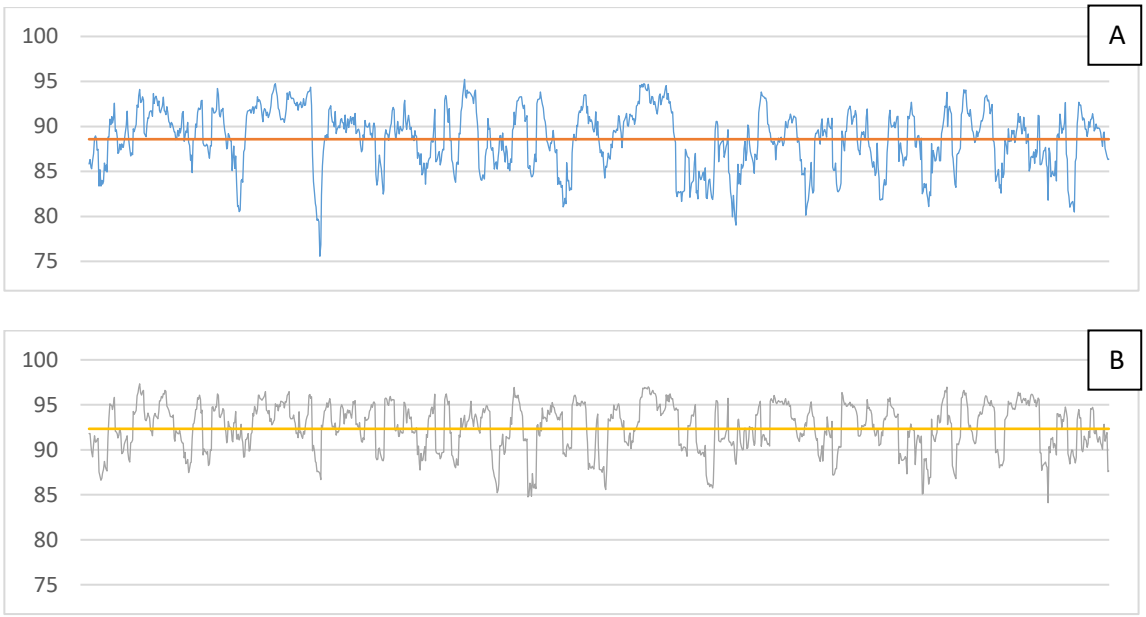
Genom *P. exilis* je duplikovaný, tudíž je možné předpokládat, že jedna z každé dvojice sesterských genů mezi dvěma scaffoldy pochází z *C. himalaica* a z širšího pohledu lze tento předpoklad zobecnit i na celé scaffoldy. Jako ukázkový příklad lze zvolit již zmiňované scaffoldy 13 a 51.

Z recipročního blastu sestaveného transkriptomu *C. himalaica* a proteinových sekvencí scaffoldů 13 a 51 *P. exilis* byly získány následující hodnoty. Z celkově 117606 transkriptů byl přiřazen nejlepší reciproční hit dohromady k 1815 transkriptům při zohledňování co nejvyššího bit skóre (633 s geny ve scaffoldu 13 a 1181 s geny ve scaffoldu 51) a k 1823 transkriptům při zohlednění co nejnižší hodnoty E-value (639 s geny ve scaffoldu 13 a 1183 s geny ze scaffoldu 51). Porovnání bylo provedeno pomocí tblastn v případě, že *C. himalaica* byla databází a *P. exilis* jako dotaz a pomocí blastx v opačném případě. V obou bězích programu byly povoleny dva nejlepší hity.

Další analýza byla zaměřena na průměrnou podobnost genů ve scaffoldech 13 a 51 *P. exilis* a transkriptomem *C. himalaica*, kdy byl spuštěn tblastn s transkriptomem *C. himalaica* jako databází a nejdříve s celým scaffoldem 13 a následně s celým scaffoldem 51 jako dotazem, s omezením běhu programu na jeden nejlepší hit. S celým scaffoldem 13 bylo nalezeno 1429 hitů s podobností větší než 80%, u scaffoldu 51 pak 1377. V případě, kdy daná čísla vztáhneme na určené sesterské geny ve scaffoldech 13 a 51, u scaffoldu 13 bylo určeno 917 genů se sesterskou dvojicí ve scaffoldu 51 a s recipročním hitem s *C. himalaica*, což odpovídá 80,16% všech sesterských genů ve scaffoldu 13. U scaffoldu 51 se pak jedná o 968 genů se sesterskou dvojící ve scaffoldu 13 a s recipročním hitem s *C. himalaica*, což odpovídá 84,62% všech sesterských genů ve scaffoldu 51. Na obr. 16 je možné porovnat křivky průměrné podobnosti deseti po sobě jdoucích genů napříč scaffoldem 13 a 51 s transkriptomem *C. himalaica*, kdy byly v analýze zohledňovány pouze sesterské geny scaffoldů 13 a 51. U scaffoldu 13 je průměrná podobnost 88,58 %, u scaffoldu 51 pak 92,33 %. V případě, kdy jsou křivky podobnosti zobrazeny na jednom grafu, průměrná podobnost scaffoldu 51 a transkriptomu *C. himalaica* se nachází ve většině úseků nad křivkou průměrné podobnosti scaffoldu 13 a transkriptomu *C. himalaica*. Tato čísla odpovídají předpokladům z výsledků hledání nejlepších recipročních hitů, kdy větší počet genů byl u genů ze scaffoldu 51, dokonce prakticky dvojnásobný.

Z výše uvedených skutečností můžeme usoudit, že scaffold 51 bude pravděpodobně patřit do subgenomu evolučně spojeného s *C. himalaica* a scaffold 13 pak do druhého

subgenomu pocházejícího od druhého evolučního rodiče, kdy jejich zkřížením během evoluce došlo k celogenomové duplikaci a vyvinutí *P. exilis*. Stejný předpoklad pak lze mimo zmíněné scaffoldy aplikovat také na celý genom *P. exilis*.



Obrázek 16: Graf zobrazuje hodnoty podobností mezi scaffoldem 13 *P. exilis* (A) a genomovou sekvencí transkriptomu *C. himalaica* a scaffoldem 51 *P. exilis* (B) a genomovou sekvencí *C. himalaica*. Tato srovnávací analýza zahrnuje pouze sesterské geny mezi scaffoldy 13 a 51. Graf zobrazuje průměrné podobnosti 10 po sobě jdoucích genů.

5 ZÁVĚR

Pachycladon je rod z čeledi brukvovité, endemický na Novém Zélandu a Tasmánii. Jedná se o rostlinný mesopolyploidní rod s duplikovaným genomem, který se z allopolyploidního předka vyvinul před asi 2 miliony let (Joly *et al.*, 2009). Konkrétně druh *Pachycladon exilis* se vyskytuje na Novém Zélandu v půdách se zásaditými ionty a jeho genomem se zabývá tato diplomová práce.

Pomocí experimentálních metod bylo při vypracovávání této diplomové práce dosaženo požadovaných výsledků. Nejprve byla zkontrolována kvalita sestavené anotace *P. exilis* pomocí programu BUSCO s výsledkem 98,6% kompletních BUSCO genů pro genom *P. exilis*. Následně byla provedena komparativní analýza s genomem *Arabidopsis lyrata*, kdy byly určeny nejlepší reciproční hity mezi genomem *P. exilis* a *A. lyrata*, celkové pokrytí a kolineární oblasti pomocí programu MCScanX. Data ukazují vysokou míru podobnosti genomů *A. lyrata* a *P. exilis*, konkrétně bylo nalezeno 28153 kolineárních párů, což odpovídá asi 84% genomu *P. exilis*. Pro podrobnější analýzu byl zvolem chromosom 3 *A. lyrata* a scaffoldy 13 a 51 assembly *P. exilis*, u kterých byly zjištěny vzájemné duplikované oblasti. Při zaměření se pouze na scaffoldy 13 a 51 byly určeny jedinečné a duplikované geny a pseudogeny a jejich umístěný ve scaffoldech. Průměrná podobnost mezi scaffoldy 13 a 51 s omezením na BLAST hity s podobností větší než 80% je asi 92%. Po vyřazení prvních 272 genů scaffoldu 13 z komparativní analýzy kvůli tomu, že neležely v kolineární oblasti těchto scaffoldů, bylo nalezeno 1144 sesterských genů, tedy nejlepších recipročních hitů, mezi scaffoldy 13 a 51. Dále bylo ve scaffoldu 13 určeno 165 unikátních genů, 41 genů duplikovaných k některému z recipročních hitů a 94 s recipročním hitem k některému pseudogenu scaffoldu 51. Počet unikátních genů ve scaffoldu 51 je 264, 42 jich je duplikovaných k některému z recipročních hitů a k 123 byl nalezen reciproční hit s některým z pseudogenů scaffoldu 13. Následně byl pomocí programu Trinity sestaven transkriptom *Crucihimalaya himalaica* a provedena komparativní analýza tohoto transkriptomu se scaffoldy 13 a 51 *P. exilis*. Průměrná podobnost mezi transkriptomem *C. himalaica* a scaffoldem 13 je 88,6% a mezi transkriptomem *C. himalaica* a scaffoldem 51 92,3%. Z těchto výsledků je možné usuzovat, že scaffold 51 patří do subgenomu evolučně pocházejícího z *C. himalaica*.

Vzhledem k tomu, že tento projekt není u konce a dále se v něm pokračuje, následujícím krokem je identifikace celých subgenomů *P. exilis* a jejich evoluční

přiřazení ke *C. himalaica* a k druhému kandidátnímu rodiči. Díky existujcímu aktuálnímu projektu se sestavením assembly *C. himalaica* (Zhang *et al.*, 2019) bude další komparativní analýza usnadněna. Dalším pokračováním projektu je sestavení chromosomů *P. exilis* a určení, zda dané chromosomy jsou subgenomově specifické nebo jestli během hybridizace a následné evoluce došlo k jejich namixování. Zajímavým tématem také může být určení času hybridizace a dalších událostí v rodu *Pachycladon*.

6 LITERATURA

Barker M. S., Vogel H., Schranz M. E. (2009): Paleopolyploidy in the Brassicales: Analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales. *Genome Biol. Evol.* **1**, 1–9.

Blanc G, Barakat A., Guyot R., Cooke R., Delsney I. (2000): Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell* **12**, 1093-1101.

Blanc G. a Wolfe K. H. (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667-1678.

Bowers J. E., Chapman B. A., Rong J., Paterson A. H. (2003): Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438.

Carretero-Paulet L. a Fares M. A. (2012): Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. *Mol Biol Evol* **29,** 3541–3551.

Dierschke T., Mandáková T., Lysák M. A., Mummenhoff K. (2009): A bicontinental origin of polyploid Australian/New Zealand Lepidium species (Brassicaceae)? Evidence from genomic in situ hybridization. *Annals of Botany* **104**, 681–688.

Dodsworth S., Chase M. W., Leitch A. R. (2016): Is post-polyploidization diploidization the key to evolutionary success of angiosperms? *Bot J Linn Soc* **180**, 1-5.

Fasola E., Ribeiro R., Lopes I. (2015): Microevolution due to pollution in amphibians: A review on the genetic erosion hypothesis. *Environ Pollut.* **204**, 181–190.

Feldman M. a Levy A. A. (2012): Genome evolution due to allopolyploidization in wheat. *Genetics* **192**, 763-774.

Freeling M., Lyons E., Pedersen B., Alam M., Ming R., Lisch D. (2008): Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome Res*. **18,** 1924–1937.

Fu Y., Li L., Hao S., Guan R., Fan G., Shi C., Wan H., Chen W., Zhang H., Liu G., Wang J., Ma L., You J., Ni X., Yue Z., Xu X., Sun X., Liu X., Lee S. M. (2017): Draft genome sequence of the Tibetan medicinal herb Rhodiola crenulata. *Gigascience.* **6(6),** 1-5.

Grabherr M. G., Haas B. J., Yassour M., Levin J. Z., Thompson D. A., 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. W., Nusbaum Ch., Lindblad-Toh K., Friedman N., Rege A. (2013): Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol.* **29(7)**, 644–652.

Greilhuber J., Borsch T., Müller K., Worberg A., Porembski S., Barthlott W. (2006): Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. *Plant Biol (Stuttg)* **8**, 770–777.

Grover C. E., Kim H., Wing R. A., Paterson A. H., Wendel J. F. (2007): Microcolinearity and genome evolution in the AdhA region of diploid and polyploid cotton (Gossypium). *Plant Journal* **50**, 995-1006.

Guo X., Zhang Z., Gerstein M. B., Zheng D. (2009): Small RNAs originated from pseudogenes: cis- or trans-acting? *PLOS Comput Biol* **5**, e1000449.

Haas B. J., Papanicolaou A., Yassour M., Grabherr M., Blood P. D., Bowden J., Couger M. B., Eccles D., Li B., Lieber M., MacManes M. D., Ott M., Orvis J., Pochet N., Strozzi F., Weeks N., Westerman R., William T., Dewey C. N., Henschel R., LeDuc R. D., Friedman N., Regev A. (2013): De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nat Protoc.* **8(8)**.

Hall A. E., Fiebig A., Preuss D. (2002): Beyond the Arabidopsis genome: opportunities for comparative genomics. *Plant Physiology* **129**, 1439-1447.

Hanada K., Kuromori T., Myouga F., Toyoda T., Shinozaki K. (2009): Increased expression and protein divergence in duplicate genes is associated with morphological diversification. *PLoS Genet* **5**, e1000781.

Hanada K., Zou C., Lehti-Shiu M. D., Shinozaki K., Shiu S. H. (2008): Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol* **148**, 993–1003.

Heenan P. B. a Mitchell A. D. (2003): Phylogeny, biogeography and adaptive radiation of Pachycladon (Brassicaceae) in the mountains of South Island, New Zealand. *J Biogeogr* **30(11)**, 1737-1749.

Heenan P. B., Dawson M. I., Smissen R. D., Bicknell R. A. (2008): An artificial intergeneric hybrid derived from sexual hybridization between the distantly related Arabidopsis thaliana and Pachycladon cheesemanii (Brassicaceae). *Bot J Linn Soc* **157**, 533-544.

Heenan P. B., Goeke D. F., Houliston G. J., Lysák M. A. (2012): Phylogenetic analyses of ITS and rbcL DNA sequences for sixteen genera of Australian and New Zealand Brassicaceae result in the expansion of the tribe Microlepidieae. *Taxon* **61**, 970–979.

Heenan P. B., Mitchell A. D., Koch M. (2002): Molecular systematics of the New Zealand Pachycladon (Brassicaceae) complex: generic circumscription and relationships to Arabidopsis sens. lat. and Arabis sens. lat. *New Zealand Journal of Botany* **40**, 543-562.

Hoffmann A. A. a Willi Y. (2008): Detecting genetic responses to environmental change. *Nat Rev Genet* **9(6)**, 421-432.

Hu T. T., Pattyn P., Bakker E. G., Cau J., Cheng J-F., Clark R. M., Fahlgren N., Fawcett J. A., Grimwood J., Gundlach H., Haberer G., Hollister J. D., Ossowski S., Ottilar R. P., Salamov A. A., Schneeberger K., Spannagl M., Wang X., Yang L., Nasrallah M. E., Bergelson J., carrington J. C., Gaut B. S., Schmutz J., Mayer K. F. X., Van de Peer Y., Grigoriev I. V., Nordborg M., Weigel D., Guo Y-L. (2011): The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet.* **43(5)**, 476-481.

Chapman A. D. (2009): Numbers of living species in Australia and the world. *Toowoomba, Qld: Australian Biodiversity Information Services.*

Jiao Y., Wickett N. J., Ayyampalayam S., Chanderbali A. S., Landherr L., Ralph P. E., Tomsho L. P. (2011): Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97-100.

Joly S., Heenan P. B., Lockhart P. J. (2009): A Pleistocene inter-tribal allopolyploidization event precedes the species radiation of Pachycladon (Brassicaceae) in New Zealand. *Molecular Phylogenetics and Evolution* **51**, 365–372.

Kejnovsky E., Leitch I. J., Leitch A. R. (2009): Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends Ecol Evol* **24**, 572–582.

Kellis M., Birren B. W., Lander E. S. (2004): Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**, 617–624.

Kelly L. J., Leitch A. R., Fay M. F., Renny-Byfield S., Pellicer J., Macas J., Leitch I. J. (2012): Why size really matters when sequencing plant genomes. *Plant Ecology & Diversity* **5**, 415-425.

Kriventseva E.V., Kuznetsov D., Tegenfeldt F., Manni M., Dias R., Simao F.A., Zdobnov E.M. (2019): OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, **Vol. 47**, Database issue D807–D811.

Leitch I. J. a Bennett M. D. (2004): Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society* **82**, 651-663.

Leitch I. J., Chase M. W., Bennett M. D. (1998): Phylogenetic analysis of DNA C-values provides evidence for a small ancestral genome size in flowering plants. *Annals of Botany* **82**, 85-94.

Lim K. Y., Kovařík A., Matyasek R., Chase M. W., Clarkson J. J., Grandbastien M. A., Leitch A. R. (2007): Sequence of events leading to near-complete genome turnover in allopolyploid Nicotiana within five million years. *New Phytologist* **175**, 756-763.

Lysák M. A., Berr A., Pecinka A., Schmidt R., McBreen K., Schubert I. (2006): Mechanisms of chromosome number reduction in Arabidopsis thaliana and related Brassicaceae species. *Proc. Natl. Acad. Sci. USA* **103**, 5224–5229.

Lysák M. A., Koch M. A., Beaulieu J. M., Meister A., Leitch I. J. (2009): The dynamic ups and downs of genome size evolution in Brassicaceae. *Mol Biol Evol* **26**, 85-98.

Lysák M. A., Cheung K., Kitschke M., and Bureš P. (2007): Ancestral chromosomal blocks are triplicated in Brassiceae species with varying chromosome number and genome size. *Plant Physiol.* **145**, 402–410.

Mandáková T. a Lysák M.A. (2018): Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology* **42**, 55-65.

Mandáková T., Heenan P. B., Lysák M. A. (2010): Island species radiation and karyotypic stasis in Pachycladon allopolyploids. *BMC Evolutionary Biology* **10**, 367.

Mandáková T., Joly S., Krzywinski M., Mummenhoff K., Lysák M. A. (2010): Fast diploidization in close mesopolyploid relatives of Arabidopsis. *The Plant Cell*, **Vol 22**, 2277-2290.

Mandáková T., Pouch M., Harmanová K., Zhan S. H., Mayrose I., Lysák M. A. (2017): Multispeed genome diploidization and diversification after an ancient allopolyploidization. *Molecular Ecology*, 1-18.

McBreen K. a Heenan P. B. (2006): Phylogenetic relationships of Pachycladon (Brassicaceae) species based on three nuclear and two chloroplast DNA markers. *New Zealand Journal of Botany*, **Vol 44**, 377-386.

Ming R., et al. (2008): The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* **452**, 991–996.

Panchy N., Lehty-Shiu M, Shiu S-H. (2016): Evolution of Gene Duplication in Plants. *Plant Physiology*, **Vol. 171**, 2294-2316.

Panopoulou G., Hennig S., Groth D., Krause A., Poustka A. J., Herwig R., Vingron M., Lehrach H. (2003): New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res* **13**, 1056–1066.

Pellicer J., Fay M., Leitch I. (2010): The largest eukaryotic genome of them all. *Bot J Linn Soc* **164**, 10–15.

Qiao Q., Wang Q., Han X., Guan Y., Sun H., Zhong Y., Huang J., Zhang T. (2016): Transcriptome sequencing of Crucihimalaya himalaica (Brassicaceae) reveals how Arabidopsis close relative adapt to the Qinghai-Tibet Plateau. *Sci. Rep.* **6**, 21729.

Renny-Byfield S. a Wendel J. F. (2014): Doublind down on genomes: Polyploidy and crop plants. *American Journal of Botany* **101(10)**, 1711-1725.

Renny-Byfield S., Chester M., Kovařík A., Le Comber S. C., Grandbastien M.-A., Deloger M., Nichols R. A., Macas J., Novák P., Chase M. W., Leitch A. R. (2011): Next generation sequencing reveals genome downsizing in allotetraploid Nicotiana tabacum, predominantly through the elimination of paternally derived repetitive DNAs. *Molecular Biology and Evolution* **28**, 2843-2854.

Scannell D. R. a Wolfe K. H. (2008): A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res* **18**, 137–147.

Sémon, M. a Wolfe K. H. (2007): Rearrangement rate following the whole-genome duplication in teleosts. *Mol. Biol. Evol.* **24**, 860–867.

Shirasu K., Schulman A. H., Lahaye T., Schulze-Lefert P. (2000): A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Research* **10**, 908-915.

Schlötterer C. (2015): Genes from scratch: the evolutionary fate of de novo genes. *Trends Genet* **31**, 215–219.

Schnable J. C., Springer N. M., Freeling M. (2011): Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. USA* **108**, 4069–4074.

Schranz M. E., Lysák M. A., Mitchell-Olds T. (2006): The ABC's of comparative genomics in the Brassicaceae: Building blocks of crucifer genomes. *Trends in Plant Science* **11**, 535–542.

Soltis D. E., Albert V. A., Leebens-Mack J., Bell C. D., Paterson A. H., Zheng C. F., Sankoff D. (2009): Polyploidy and angiosperm diversification. *American Journal of Botany* **96**, 336-348.

Soltis D. E., Soltis P. S., Schemske D. W., Hancock J. F., Thompson J. N., Husband B. C., Judd W. S. (2007): Autopolyploidy in angiosperms: Have we grossly underestimated the number of species? *Taxon* **56**, 13–30.

Tang H., Bowers J. E., Wang X., Ming X., Alam M., Paterson A. H. (2008): Synteny and collinearity in plant genomes. *Science* **320**, 486–488.

Tarchini R., Biddle P., Wineland R., Tingey S., Rafalski A. (2000): The complete sequence of 340 kb of DNA around the rice Adh1-adh2 region reveals interrupted colinearity with maize chromosome 4. *Plant Cell* **12**, 381–391.

Thibaud-Nissen F., Ouyang S., Buell C. R. (2009): Identification and characterization of pseudogenes in the rice gene complement. *BMC Genomics* **10**, 317.

Thomas B. C., Pedersen B., Freeling M. (2006): Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* **16**, 934–946.

Tiley G. P., Ané C., Burleigh J. G. (2016): Evaluating and characterizing ancient whole-genome duplications in plants with gene count data. *Genome Biol. Evol* **8(4)**, 1023-1037.

True J. R. a Carroll S. B. (2002): Gene co-option in physiological and morphological evolution. *Annu Rev Cell Dev Biol* **18**, 53–80.

Vanin E. F. (1985): Processed pseudogenes: characteristics and evolution. *Annu Rev Genet* **19**, 253–272.

Veitia R. A., Bottani S., Birchler J. A. (2008): Cellular reactions to gene dosage imbalance: Genomic, transcriptomic and proteomic effects. *Trends Genet.* **24**, 390–397.

Vitte C. a Panaud O. (2003): Formation of solo-LTRs through unequal homologous recombination counterbalances amplifi cations of LTR retrotransposons in rice Oryza sativa L. *Molecular Biology and Evolution* **20**, 528-540.

Voelckel C., Mirzaei M., Reichelt M., Luo Z., Pascovici D., Heenan P. B., Schmidt S., Janssen B., Haynes P. A., Lockhart P. J. (2010): Transcript and protein profiling identify candidate gene sets of potential adaptive significance in New Zealand Pachycladon. *BMC Evolutionary Biology*, **10**:151.

Wang, X., et al.; Brassica rapa Genome Sequencing Project Consortium (2011): The genome of the mesopolyploid crop species Brassica rapa. *Nat. Genet*. **43**, 1035–1039.

Wang Y., Tan X., Paterson A. H. (2013): Different patterns of gene structure divergence following gene duplication in Arabidopsis. *BMC Genomics* **14**, 652.

Wang Y., Tang H., DeBarry J. D., Tan X., Li J., Wang X., Lee T., Jin H., Marler B., Guo H., Kissinger J. C., Paterson A. H. (2012): MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **Vol. 40**, No. 7.

Waterhouse R.M, Seppey M., Simao F. A., Manni M., Ioannidis P., Klioutchnikov G., Kriventseva E. V., Zdobnov E. M. (2017): BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution* **35(3)**, 543-548.

Waterhouse R.M., Tegenfeldt F., Li J., Zdobnov E.M., Kriventseva E.V. (2013): OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.* **41**, 358-365.

Wendel J. a Doyle J. (2005): Polyploidy and evolution in plants. CABI publishing, Wallingford, UK.

Wolfe K. H. (2001): Yesterday's polyploids and the mystery of diploidization. *Nature Reviews. Genetics* **2**, 333-341.

Wray G. A., Hahn M. W., Abouheif E., Balhoff J. P., Pizer M., Rockman M. V., Romano L. A. (2003): The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* **20**, 1377–1419.

Wu Y. M., Li J., Chen X. S. (2018): Draft genomes of two blister beetles Hycleus cichorii and Hycleus phaleratus. *GigaScience*, **7(3)**, 1–7.

Xiao Y., Xu P., Fan H., Baudouin L., Xia W., Bocs S., Xu J., Li Q., Guo A., Zhou L., Li J., Wu Y., Ma Z., Armero A., Issali A. E., Liu N., Peng M., Yang Y. (2017): The genome draft of coconut (Cocos nucifera). *Gigascience.* **6(11)**, 1-11.

Yamada K., Lim J., Dale J. M., Chen H., Shinn P., Palm C. J., Southwick A. M., Wu H. C., Kim C., Nguyen M. (2003): Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* **302**, 842–846.

Yogeswaran K., Heenan P. B., Voelckel C., Joly S. (2010): Pachycladon. In Wild Crop Relatives: Genomic and Breeding Resources Wild Relatives of Oilseeds. *Springer*.

Zhang T., Qiao Q., Novikova P. Y., Wang Q., Yue J., Guan Y., Ming S., Liu T., De J., Liu Y., Al-Shehbaz I. A., Sun H., Van Montagu M., Huang J., Van de Peer Y., Qiong L. (2019): Genome of Crucihimalaya himalaica, a close relative of Arabidopsis, shows ecological adaptation to high altitude. *Proceedings of the National Academy of Sciences*, **116 (14)**, 7137-7146.

Zhao M., Zhang B., Lish D., Ma J. (2017): Patterns and Consequences of Subgenome Differentiation Provide Insights into the Nature of Paleopolyploidy in Plants. *The Plant Cell*, Vol. **29**, 2974–2994.

7 SEZNAM POUŽITÝCH ZKRATEK

WGD	celogenomová duplikace
PPD	post-polyploidní diploidizace
ACK	ancestrální karyotyp křížatých
mya	milionů let zpět
EST	exprimovaná sekvenační značka
BUSCO	Benchmarking Universal Single-Copy Ortholog
BLAST	Basic Local Alignment Search Tool
Mb	milion bazí
kb	tisíc bazí
b	báze