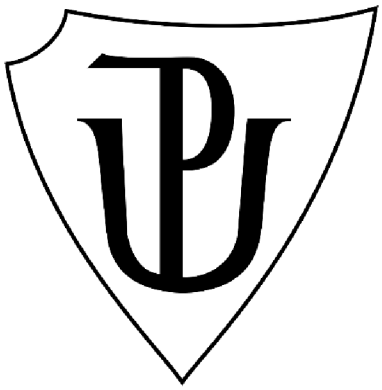
UNIVERZITA PALACKÉHO V OLOMOUCI

Přírodovědecká fakulta

Katedra biochemie



Identifikácia chromozomálnych prestavieb v rodu *Musa* spp. s využitím Oxford Nanopore sekvencií

DIPLOMOVÁ PRÁCA

Autor:	**Bc. Simona Martikánová**
Studijní program:	N1406 Biochemie
Studijní obor:	Bioinformatika
Forma studia:	Prezenčná
Vedoucí práce:	**Mgr. Eva Hřibová, PhD.**
Rok:	2023

Prehlasujem, že som diplomovú prácu vypracovala samostatne s vyznačením všetkých použitých prameňov a spoluautorstva. Súhlasím so zverejnením diplomovej práce podľa zákona č. 111/1998 Sb., o vysokých školách, v znení neskorších predpisov. Bola som zoznámená s tým, že se na moji práci vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorský zákon, v znení neskorších predpisov.

V Olomouci dňa ....

Podpis študenta

Poďakovanie

Chcela by som poďakovať vedúcej svojej diplomovej práce Mgr. Eve Hřibové, Ph.D za odborné vedenie, čas, trpezlivosť a úsilie ktoré venovala konzultáciam. Taktiež ďakujem mojej rodine za pevné nervy a podporu.

Bibliografická identifikace

Meno a priezvisko autora	Bc. Simona Martikánová
Názov práce	Identifikácia chromozomálnych prestavieb v rodu Musa spp. s využitím Oxford Nanopore sekvencií
Typ práce	Diplomová
Pracovisko	Katedra biochémie
Vedúci práce	Mgr. Eva Hřibová, PhD.
Rok obhajoby práce	2023

Abstrakt

Banánovník (*Musa* spp.) je rod bylín z čeľade *Musaceae*. Najdôležitejšie typy jedlých banánov pochádzajú z hybridizácie dvoch druhov: *Musa acuminata* a *Musa balbisiana.* Pozorovanie nepravidelností párovania chromozómov v meióze hybridov medzi týmito poddruhmi naznačovalo prítomnosť veľkých chromozomálnych štrukturálnych prestavieb, hlavne translokácií. Nanopórové sekvenovanie je prístup tretej generácie, ktorý umožňuje získanie DNA alebo RNA sekvencií. V tejto práci sme sa zamerali na použitie čítaní získaných s metódou Oxford Nanopore pre genóm banánovníka pre identifikáciu veľkých chromozomálnych translokácií a ich vizualizáciu.

Kľúčové slová	Banánovník, translokácia, Oxford Nanopore sekvenovanie, chromozómy, genóm
Počet strán	71
Počet príloh	7
Jazyk	Slovenský

Bibliographical identification

Autor's first name and surname	Bc. Simona Martikánová
Title	Identification of chromosomal rearrangements in Musa spp. using Oxford Nanopore sequencing
Type of thesis	Diploma
Department	Department of biochemistry
Supervisor	Mgr. Eva Hřibová, PhD.
The year of presentation	2023

Abstract

Banana (*Musa spp.*) is a genus of herbs from the *Musaceae* family. The most important types of edible bananas come from the hybridization of two species: *Musa acuminata* and *Musa balbisiana*. Observation of chromosome pairing irregularities in meiosis of hybrids between these subspecies indicated the presence of large chromosomal structural rearrangements, mainly translocations. Nanopore sequencing is a third-generation approach that allows obtaining DNA or RNA sequences. In this work, we focused on using the Oxford Nanopore reads for the banana genome for a large number of chromosomal translocations and their visualization.

Keywords	Banana tree, translocation, Oxford Nanopore sequencing, chromosomes, genome
Number of pages	71
Number of appendices	7
Language	Slovak

OBSAH

1 ÚVOD 1

2 SÚČASTNÝ STAV RIEŠENEJ PROBLEMATIKY 2

2.1 Banánovník 2
2.1.1 Taxonómia 3
2.1.2 Morfológia 5
2.1.3 Genóm banánovníka 6

2.2 Sekvenovanie 10
2.2.1 Metódy sekvenovania tretej generácie 11

2.3 Sekvenovanie pomocou Nanopórov 11
2.3.1 Vývoj metódy nanopore 12
2.3.2 Princíp „nanopore“ sekvenačnej technológie 13

2.4 Analýza dát 16
2.4.1 „Base calling“ 17
2.4.2 Detegovanie DNA a RNA modifikácií 17
2.4.3 Oprava chýb 17

3 EXPERIMENTÁLNA ČASŤ 20

3.1 Sekvenačné dáta 20

3.2 Bioinformatická analýza 21
3.2.1 Volanie bázií (base calling) 21
3.2.2 Zostavenie kontigov (assembly) 21
3.2.3 Sekvenčné zarovnanie a identifikácia translokácií 22
3.2.4 Získanie informácii o translokáciach 26

3.3 Vizualizácia translokácií 26

3.3 Lokalizácia translokácií u jednotlivých ONT čítaní 29

4 VÝSLEDKY 30

4.1 Analýza sekvenačných dát 30

4.2 Vizualizácia translokácií s assembly 31
4.2.1 Druhy zo sekcie *Eumusa* 31
4.2.2 Polyploidy zo sekcie *Eumusa* 35
4.2.3 Plané druhy zo sekcie Australimusa 38
4.2.4 Plané druhy zo sekcie Callimusa 40
4.2.4 Druhy sekcie Rhodochlamys 42
4.2.5 Druh z rodu *Ensete* 43

4.3 Vizualizácia translokácií u jednotlivých ONT čítaní 44

5 DISKUSIA 49

6 ZÁVER.....53

7 LITERATÚRA.....55

8 POUŽITÉ SKRATKY.....61

9 PRÍLOHY.....62

CIELE PRÁCE

Ciele zahrňujú:

- Vypracovanie literárnej rešerše na zadané téma
- Analýza Oxford Nanopore sekvenačných dát (ONT) vybraných zástupcov rodu *Musa spp.* a identifikácia chromozomálnych prestavieb dvoma rozdielnymi bioinformatickými prístupmi
    1. Mapovanie jednotlivých ONT čítaní na referenčnú genómovú sekvenciu banánovníka a identifikácia potenciálnych prestavieb (translokácií)
    2. Zostavenie dlhých kontigov (tzv. „draft assemblies“) z ONT dát pomocou dostupných programov a *in silico* identifikácia chromozomálnych prestavieb v zostavených kontigoch
- Porovnanie prístupu pre *de novo* identifikáciu satelitnej DNA (tandemovo opakujúcich sa repetícií) v dlhých ONT čítaniach a v zostavených kontigoch

1 ÚVOD

Banánovník (*Musa spp.*) je jednoklíčnolistová trvalá rastlina patriaca medzi hlavné vývozné komodity viacerých rozvojových krajín. Je predpokladané, že na vzniku kultivarov sa pomocou hybridizácie podieľajú štyri hlavné druhy: *Musa acuminata, Musa Balbisiana, Musa schizocarpa* a druhy *Australimusa.* To viedlo aj k prechodu od divokých k jedlým jedincom a vzniku triploidov z jedlých diploidov. Pôvod jedlých kultivarov sa však vďaka nedávnym výskumom zdá oveľa zložitejší ako bolo predpokladané. Boli charakterizované veľké štrukturálne prestavby vo forme recipročných translokácií, rozšírených v kultivovanej zárodočnej plazme. Zložitosť genómov a identifikácia veľkých chromozomálnych translokácií je základom pre rozlúštenie evolučnej histórie banánovníkov a podporu genetických štúdií.

V posledných rokoch sú sekvenačné technológie s dlhými čítaniami na vzostupe. Majú prísľub získania kompletnejších genómových zostáv jednoduchšou cestou ako predchádzajúce generácie. Jednou z týchto platforiem ponúka firma Oxford Nanopore Technologies. Vyvinula novú metódu DNA/RNA sekvenovania, ktorá používa nanopóry a ako jediná ponúka sekvenovanie v reálnom čase. Dokáže vyprodukovať až 20 Gb sekvenačných dát, ale stále sa ukazuje ako pomerne chybová. Analýza veľkého množstva údajov generovaných sekvenačnými technológiami dlhých čítaní je komplexný, viac stupňový proces, ktorý je výpočtovo náročný a často vyžaduje odbornosť v oblasti bioinformatiky.

Teoretická časť tejto práce obsahuje informácie o druhoch banánovníka a jeho genóm. Následne sa zameriava práve na metódu Oxford Nanopore, jej význam a základy spracovania pomocou bioinformatických nástrojov.

Naväzujúca praktická časť sa sústreďuje na vytvorenie cesty, ktorá by mohla viesť k identifikácií veľkých chromozomálnych translokácií pomocou dát získaných prostredníctvom nanopórov.

2 SÚČASTNÝ STAV RIEŠENEJ PROBLEMATIKY

2.1 Banánovník

Banány sú základnou potravinou pre milióny ľudí v rozvojových zemiach sveta. Pochádzajú z juhovýchodnej Ázie a západu Tichomoria, kde sa ich diploidný predkovia stále vyskytujú vo voľnej prírode. Uvažuje sa, že banány boli prevezené z Indonézie na Madagaskar okolo roku 500 n.l. a následne do Afriky. V dobe príchodu Portugalcov v 14-15. storočí už boli známe na západnom pobreží Afriky. Približne v 16. storočí ich práve Portugalci doviezli do Dominikánskej republiky (Robinson & Saúco, 2010).

Na rozdiel od iných typov ovocia sú banány dostupné počas celého roka, tešia sa veľkej popularite hlavne v Indii a Afrike aj vďaka nízkej cene niektorých typov. Využívajú sa na rôzne účely, najmä v južnej Indii. Takmer každá časť rastliny je použiteľná (Rekha &, 2016). Exportný obchod s dezertnými banánmi začal koncom 19. storočia a jeho rozvoz urýchlilo zavedenie chladenia zásielky (Robinson & Saúco, 2010).

Banány sú pestované v tropických až subtropických oblastiach sveta a počas minulého storočia nabrali na popularite ako ľahko dostupné a výživné ovocie v supermarketoch vyspelých krajín. Sú heterogénnym produktom, odrody banánov pestovaných na miestnu spotrebu sa často líšia od odrôd pre vývoz. Tradičným vývozným druhom pôvodom z Jamajky (polovica 19.storočia) bol Gros Michel (*Musa* genóm AAA). Tieto kultivary začala v 20. storočí decimovať fusarióza (známa aj ako Panamská choroba) a ohrozuje aj kultivary Cavendish, ktorými bola postupne nahradená odroda Gros Michel. Kultivary Cavendish sú v súčasnosti zodpovedné za približne 45% všetkej produkcie (Paggi & Spreen, 2003; Ploetz, 2015). V roku 2019 bola na čele produkcie banánov India s 30,4 miliónmi ton. Po nej nasledovala Čína (11,6 miliónov ton), Indonézia (7,2 miliónov ton), Brazília (6,8 miliónov ton) a Ekvádor (6,5 miliónov ton) (Acevedo & Carrillo *et al.*, 2021).

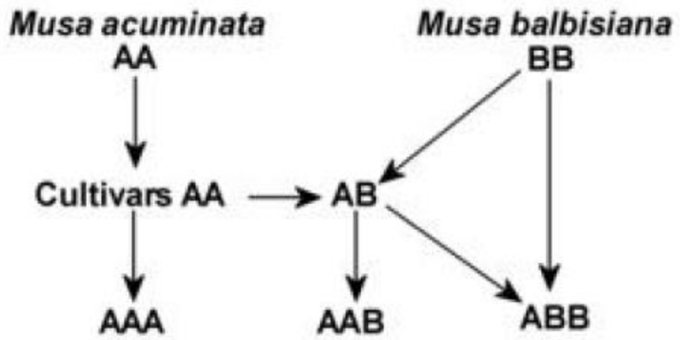
Najdôležitejšie typy jedlých banánov pochádzajú z hybridizácie dvoch druhov (Obr. 1*):* *Musa acuminata* (genóm A) a *Musa balbisiana* (genóm B), ktoré majú korene v juhovýchodnej Ázii. Niektoré kultivary však mohli vzniknúť hybridizáciou s *Musa schizocarpa* (genóm S) a aspoň jeden filipínsky klon môže pochádzať zo starovekej hybridizácie *Musa balbisiana* a *Musa textilis* (genóm T) (Robinson & Saúco, 2010).

2.1.1 Taxonómia

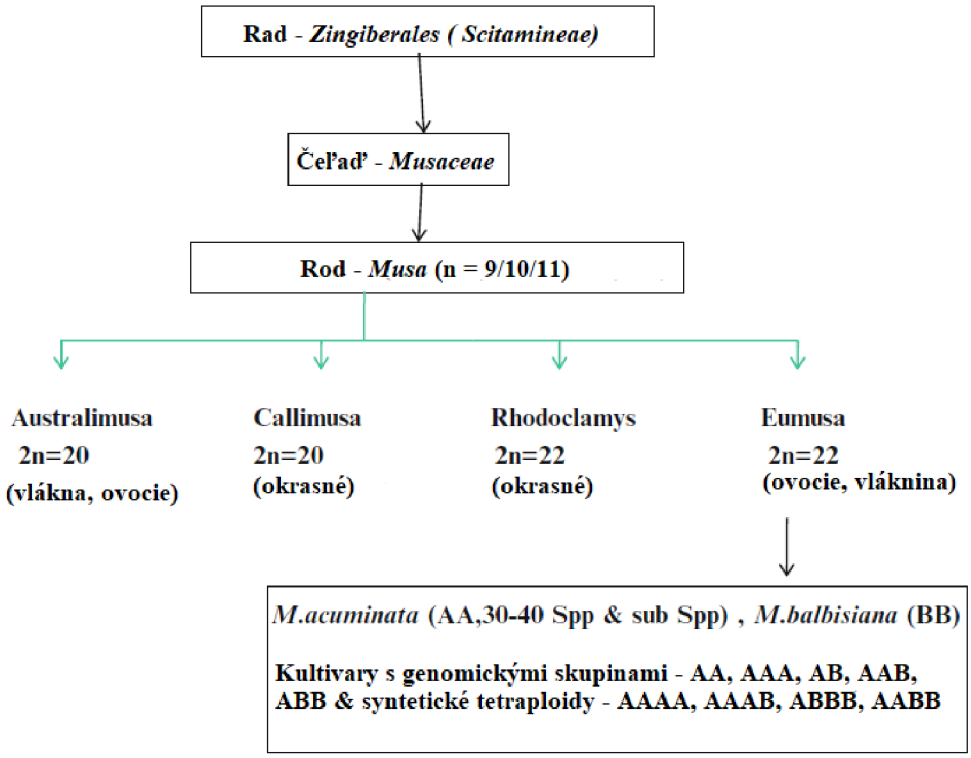
Musaceae (rad *Zingiberales*) je čeľaď kvitnúcich rastlín pozostávajúcich z 3 druhov: *Musa, Musella* a *Ensete*. Banán je pojem zahŕňajúci mnoho hybridov rodu *Musa*, napr. plantainy a dezertné banány (Borrell *et al*., 2019). Rodové meno *Musa* je odvodené z arabského slova *mouz*, v preklade prst (finger). Najstaršiu „vedeckú" klasifikáciu banánov vytvoril Linnaeus v roku 1783, tým že všetkým dezertným banánom dal názov *Musa sapientium* a plantajnom, s vyšším obsahom škrobu, *Musa paradisiaca*. Tieto názvy sa v dnešnej taxonomickej klasifikáciu už nepoužívajú a nemožno ich preto použiť na rozlíšenie medzi banánmi a plantajnami. S modernou metódou klasifikácie prišiel Simmonds a Shepherd (1955) (Robinson & Saúco, 2010).

Rod Musa sa delí do štyroch pomocných taxonomických sekcií (Obr. 2). *Callimusa* a *Australimusa* s počtom chromozómov 2n = 20 alebo 2n = 18, *Eumusa* a *Rhodochlymys* s počtom chromozómov 2n = 22. (Rekha &, 2016).

Sekcia *Eumusa* je geograficky najrozšírenejšia a obsahuje dva hlavné druhy : *Musa acuminata* a *Musa balbisiana*, ktorých hybridizácia viedla k vzniku veľkej väčšiny kultivarov jedlých banánov a plantajnov. *Musa acuminata* sa ďalej delí na osem poddruhov: *banksii, burmannica, burmannicoides, malaccensis, microcarpa, truncata, siamea* a *zebrina*. *Musa balbisiana* je menej diverzifikovaná. Kultivary môžu byť diploidné alebo polyploidné a prevažná časť vznikla kombináciou genómov *M.acuminata* označovaná ako genóm A a *M. balbisiana* (genóm B).Typickými zástupcami jedlých typov banánov sú diploidné a triploidné kultivary s genómovým zložením AA, AB, AAB alebo ABB (Obr.1). Existujú aj tetraploidné banány, ktoré sa avšak v prírode nevyskytujú a pochádzajú zo šľachtiteľských programov (Robinson & Saúco, 2019).



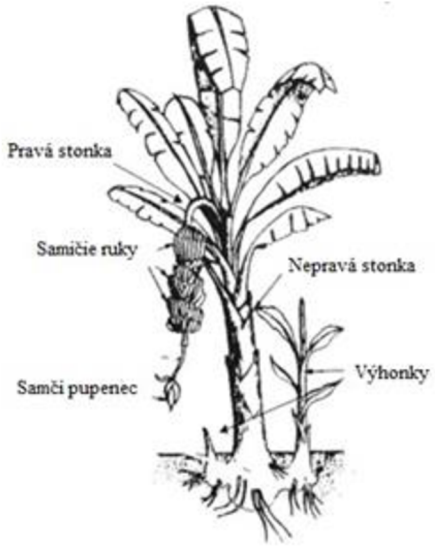
Obr. 1 Vzťahy kríženia jedlých banánov (MacBryde, 2009)



Obr. 2 Taxonomická klasifikácia *Musaceae* (prevzaté a preložené z Simmonds, 1962)

Australimusa má päť až šesť druhov, ale najdôležitejšie sú *M. textilis* a ďalšia skupina jedlých banánov, označovaných ako Fe'i. *M. textilis* je rozšírený zväčša v juhovýchodnej Ázii a vyrábajú sa z neho pevné prírodné vlákna známe ako Manilské konope. Kultivary typu *Fe'i* vznikli nezávisle na jedlých typov zo sekcie Eumusa. Charakterizované sú ich vzpriamenými trsmi, ružovo-červenou až fialovou šupkou a sýto žltou alebo oranžovou dužinou. Vyskytujú sa najmä na ostrovoch v Tichomorí a väčšina z nich má vysoký obsah beta-karoténu (prekurzora vitamínu A) (Ploetz *et al*., 2007).

Sekcia *Rhodochlymus* zahŕňa plané členy okrasného charakteru evolučne príbuzné sekcii Eumusa. Vyznačujú sa štíhlym vzrastom a vzpriameným súkvetím s pestrofarebnými metlinami. Callimusa pozostáva z rastlín s okrasným významov bez jedlého plodu. Tieto dve sekcie rastú najmä na Bornee a okolitých ostrovoch (Mohandas & Ravishankar, 2016).



Obr. 3 Morfológia banánovníka (*Musa spp.*) (upravené a prevzaté z MacBryde, 2009)

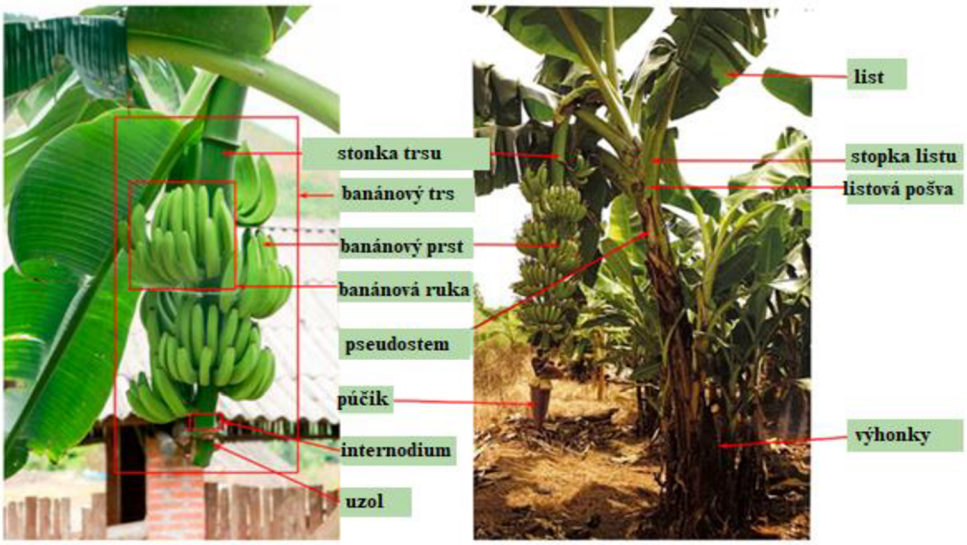
2.1.2 Morfológia

Banánovníky sú veľké bylinné trvalky. Ich vzhľad sa podobá stromu (Obr. 3). Komerčne pestované typy dosahujú výšku medzi 2 až 5 metrami, ale jeden divoký druh (*Musa ingens*) môže dosiahnuť výšku až 16 metrov. „Nepravý kmeň" je zvaný pseudostem (nepravá stonka) a tvoria ho spojené listové pošvy. Koruna je tvorená ružicou z veľmi dlhých, veľkých až eliptických listov (Karamura & Karamura, 1995).

Banán je jednoklíčnolistá rastlina s podzemnou časťou zvanou sympódium, bežne nazývanou korm. Korm podporuje sériu listov, ktorých pošvy tvoria pseudostem. Prvé listy výhonku sú šupinaté, nasledujú kopijovité listy a neskôr vo vývoji vznikajú laminárové listy. Na výhonky môže byť vytvorených 30-50 alebo viac listov, ale v jednom okamihu je prítomných len 10-14 živých listov. Zvyčajne trvá 7 až 14 dní, kým vzíde jeden list (Turner *et al.*, 2007).

Banány majú kompletné súkvetie: samičie kvety, samčí pupenec a neutrálne kvety, ktoré sa nevyvíjajú do plodov a opadávajú počas dozrievania plodov. Vývoj banánovníka zahŕňa 2 hlavné fázy: vegetatívna a reprodukčná (Stover & Simmonds, 1987).

Vegetatívna fáza je charakterizovaná vznikom listov a reprodukčná vznikom kvetov (Stover & Simmonds, 1987). 6 až viac mesiacov po výsadbe sa na vrchole pseudostonky objaví súkvetie a pokračuje postupne v predlžovaní a dozrievaní. Plody sa zvyčajne stáčajú smerom k zemi (Turner *et al.*, 2007).



Obr. 4 Morfologická charakteristika nadzemnej časti banánovníka (prevzaté a preložené z Guo *et al.*, 2021).

2.1.3 Genóm banánovníka

Evolučne zmrazené, geneticky sterilné celosvetovo známe ovocie banán zostalo nepoznačené zelenou revolúciou a aj v dnešnej dobe výskumníci čelia prekážkam pre jeho odrodové zlepšenie. Do éry genomiky vstúpilo v posledných 10tych rokoch vďaka dekódovaniu genómu dvojitého haploidu planého genotypu *Musa acuminata* 'Pahang'. Hoci banánovník patrí k rastlinám s menším genómom (1C ~ 550 – 750 Mb), jeho genóm sa ukázal ako komplexný a jeho dekódovaním s pomocou hybridných sekvenačných stratégií bola odhalená celá rada génov a transkripčných faktorov (Dash & Rai, 2016).

Historicky banánovníky zažili mnoho pandemických bakteriálnych, plesňových a vírusových ochorení a množstvu abiotického stresu, čo zničilo niektoré komerčné plantáže aj malých farmárov. Dekódovanie genómu dodalo impulz k pochopeniu množstvu génov podieľajúcich sa na odolnosti voči chorobám a iným faktorom ovplyvňujúcich život banánovníkov (Dash & Rai, 2016).

Ako bolo zmienené vyššie, predpokladá sa, že 4 genetické skupiny sa podieľajú na vzniku kultivarov, hlavne intra a interšpecifickou hybridizáciou s rôznym množstvom príspevku a to: *Musa acuminata (A), balbisiana(* B), *schizocarpa (S)* a druhy *Australimusa* (T). Počas domestikácie došlom k dvom udalostiam. Prechod z divokých k jedlým diploidom a vznik triploidov z jedlých diploidov (Martin *et al.*, 2020). Procesy

polyploidie a hybridizácie viedli k vzniku mnohých diploidných, triploidných a tetraploidných klonov rôznou kombináciou genómov A a B (AA, AB, AAA, AAB, ABB, AABB, AAAB, ABBB) (Davey *et al.*, 2013). Medzidruhová hybridizácia, respektíve hybridizácia rozdielnych poddruhov viedla k vzniku bezsemenných, partenokarických jedlých typov banánovníkov. Na vzniku väčšiny jedlých triploidných AAA banánov sa podieľali tri rôzne poddruhy *M. acuminata* (*malaccensis, banksii a zebrina*), ktoré sa líšia štruktúrou svojich genómov – prítomnosti veľkých chromozomálnych prestavieb (Martin *et al.*, 2020; Šimoníková *et al*. 2020).

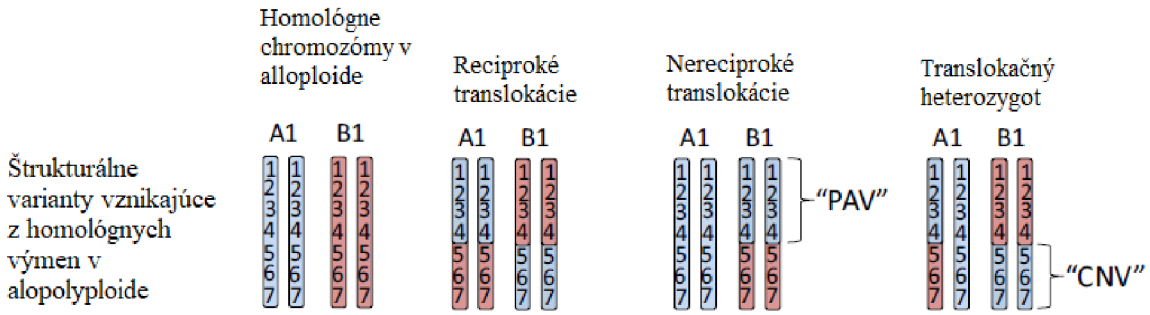
2.1.3.1 Verzie genómu

Hoci sa veľkosť genómu *Musa acuminata* pohybuje okolo 600 megabáz (Mb) (Čižková *et al.*, 2015), a je tak dva/trikrát väčší ako genóm *Arabidopsis* (150 Mb), stále je omnoho menší než obilniny (napríklad pšenica, napríklad 1C ~ 17 000 Mb) a vyznačuje sa komplexným genómom. Preto bol pripravený dvojitý haploid sekvenovaného genotypu *M. acuminata* spp. *malaccensis*, ktorý bol následne použitý pre celogenómové sekvenovanie (D'Hont *et al.*, 2012).

Prvá verzia (V1) bola zostavená pre dvojitý haploid DH- Pahang *M. acuminata* v roku 2012 tímom výskumníkov v rámci Global Musa Genomics Consortium (D'Hont *et al.*, 2012). Genóm bol sekvenovaný kombináciou technológií Sanger, Illumina a 454. Zostavené kontigy boli následne ukotvené riedkou genetickou mapou do dlhých chromozómovo špecifických scaffoldov. Prvá verzia celogenómovej sekvencie predstavovala asi 63% odhadovanej veľkosti genómu *M. acuminata* spp. *malaccensis* (Wang *et al.*, 2012).

V roku 2016 bola publikovaná druhá verzia, ktorá pridala tzv. mate pair Illumina sekvencie, optickú mapu s nízkou kompletnosťou a hustejšiu genetickú mapu. V nej bola navrhnutá zostava s 11 chromozómami zahŕňajúcimi 76% odhadovanej veľkosti genómu. Tretia verzia nebola zdieľaná s vedeckou komunitou (Martin *et al.*, 2016).

Verzia 4 vznikla pomocou sekvenačnej platformy Oxfor Nanopore (ONT; Oxford Nanopore Technologies). V porovnaní s predchádzajúcimi verziami sa kompletnosť zostavenia výrazne zlepšila. Kontig N50, definujúci kvalitu zostavenia z hľadiska spojitosti, prechádza z desiatok kbp (28 a 43 kbp pre V1 a V2) k 32 Mbp. Kumulatívna veľkosť je bližšie k odhadovanej veľkosti genómu (Belser *et al*., 2021).



Obr. 5: Štrukturálne variácie genómu, ktoré môžu nastať u polyploidov. Dva páry homologických chromozómov A1 a B1 s identickým poradím génov, kde môže nastať reciproká translokácia (chromozómové preskupenie bez straty), nereciproká translokácia (oblasť podobná PAV s absenciou B1 homológu a duplikácia A1) a translokačný heterozygot (môže vzniknúť hybridizáciou medzi jedincom s fixnou reciprokou translokáciou a jedincom bez tejto translokačnej udalosti) (Schiessl *et al.*, 2019)

2.1.3.2 Genomická štrukturálna variácia

Genomická štrukturálna variácia zahŕňa všetky varianty DNA sekvencie, v ktorej sú sekvenčné bloky väčšie ako 1 kb prenesené do iného genómového kontextu. Tieto prenosy môžu mať rôzny následok: sekvenčný blok môže byť presunutý do nového lokusu (translokácia), môže byť zmenená jeho orientácia z 5′-3′ na 3′-5′ v rovnakom mieste (inverzia), stratiť sa (delécia) a skopírovať sa do nového lokusu (duplikácia) (Nussbaum *et al.*, 2004).

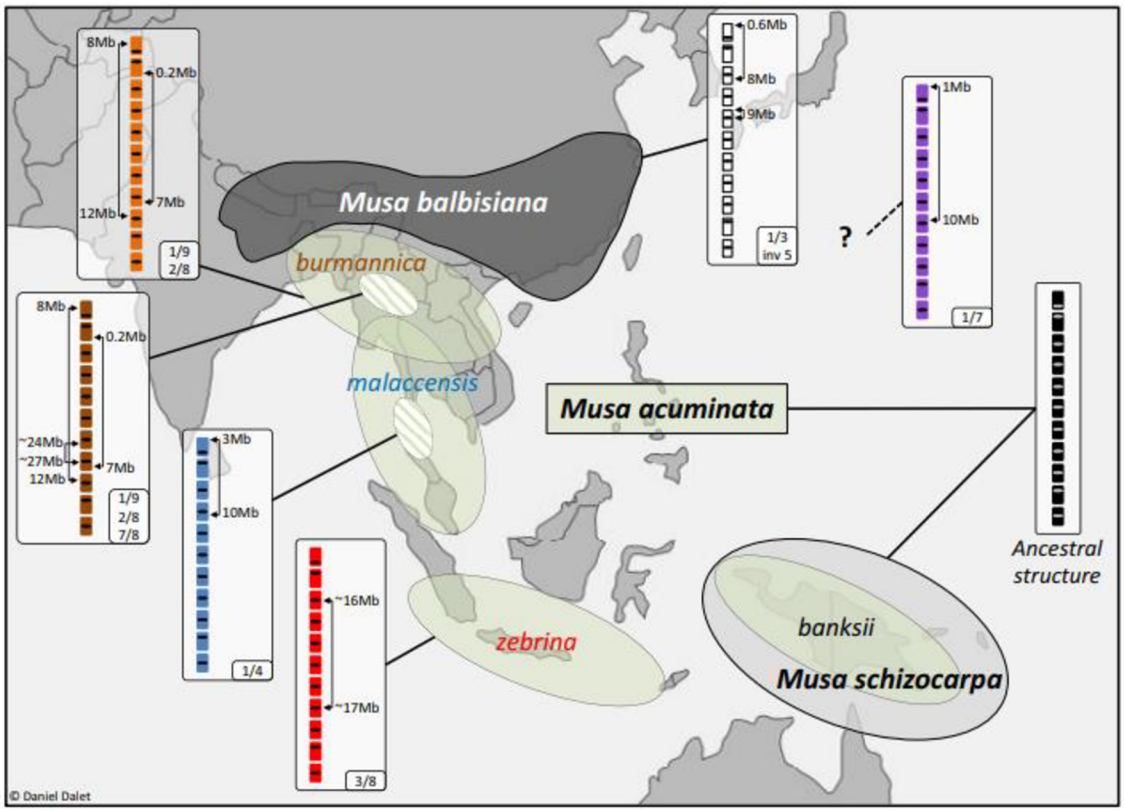
Translokácie a inverzie menia genomický kontext a neovplyvňujú počet kópií prítomnej sekvencie, delécie a duplikácie môžu zmeniť počet kópií génov obsiahnutých v danom sekvenčnom bloku. Tieto zmeny môžu viesť k individuálnej variácii v počte kópií génu – variácia počtu kópií (CNV). Ak gén či jeho oblasť u niektorých jedincov v porovnaní s ostatnými chýba, nazývame to varianty prítomnosti a neprítomnosti (PAV) (Schiessl *et al.*, 2019).

Polyploidy majú vysoký podiel sekvenčnej homológie medzi rôznymi subgenómami, vďaka čomu oveľa pravdepodobnejšie podľahnú chromozómovým prestavbám (Cifuentes *et al.*, 2010). Tieto preskupenia najčastejšie zahŕňajú recipročnú a nerecipročnú translokáciu (Obr 5.), ale aj delécie a duplikácie. Všetky vznikajú výmenou medzi nehomológnymi chromozómami počas meiózy (Xiong *et al.*, 2011). Recipročné translokácie sú charakterizované výmenou materiálu medzi nehomologickými chromozómami, zvyčajne nevedú k zisku alebo strate genetického

materiálu. Nereciprocná translokácia zahŕňa jednosmerný prenos génov z jedného chromozómu na iný nehomologický chromozóm (Schiessl *et al.*, 2019).

2.1.3.3 Dlhé chromozomálne translokácie

Skupiny druhov a poddruhov banánovníkov, ktoré sa líšia v štruktúre svojich genómov boli prvýkrát navrhnuté Shepherdom (1999) vďaka analýze párovanie chromozómov počas meiózy. Na základe konfigurácií párovania chromozómov v intersubšpecifických prírastkov bola navrhnutá prítomnosť deviatich až desiatich translokácií v *M. acuminata* distribuovaných v siedmich translokačných skupinách líšiacich sa jednou až štyrmi translokáciami (Shepherd, 1999). Neskôr boli tieto skupiny potvrdené a bližšie charakterizované pomocou sekvenačných technológií druhej generácie a farbením chromozómov metódou oligo painting FISH (Baurens *et al.*, 2019; Martin *et al.*, 2020; Šimoníková *et al.*, 2019). Takto boli identifikované a popísané chromozomálne prestavby, dlhé translokácie, ktorými sa dané skupiny (Obr. 6) navrhnuté Shepherdom vzájomne líšia.



Obr. 6: Geografické rozloženie hlavných druhov a poddruhov *Musa* súvisiacich s translokovanými chromozómovými štruktúrami (Martin *et al.*, 2020)

Základnou je skupina sa popisuje ako „Štandardná“. Translokácia medzi chromozómami 1 a 4 zodpovedá skupine „Severo-malayská“, prítomnosť translokácií - 1/9 a 2/8 je typická pre skupinu „Severná 1“, translokácia 7/8 v skupine označovanej ako „Severná 2“ a translokácia 3/8 je typická „Jávskej“ skupine. Nové poznatky týkajúce sa chromozómov naznačujú, že ďalšia skupina nazývaná ako „Východná Afrika“ má translokáciu zahŕňajúcu chromozóm 9 a jeden zo štyroch: 5, 6, 10 alebo 11. Posledná skupina „Malayan Highland“ naznačuje, že má recipročnú translokáciu medzi chromozómom 1 a/alebo 4 a ďalším chromozómom. Bola hypotetizovaná ešte jedna skupina s translokáciou 1/7 ale vychádza len z jedného štruktúrne heterozygotného kultivaru (Shepherd, 1999; Martin *et al.*, 2020).

2.2 Sekvenovanie

Prvá kompletná sekvencia nukleových kyselín molekuly, bola získaná Holley *et al.* v roku 1965 purifikovaním tRNA a použitím ribonukleáz (Pevsner, 2015). Veľkým miľníkom v histórii sekvenovania bolo vyvinutie stratégie primerovej extenzie pre sekvenciu nukleotidov DNA v 70-tych rokoch minulého storočia Ray Wuom (Wu, 1970). To sa stalo základom Sangerovho sekvenovania, metódy, ktorá dominovala sekvenovaniu DNA dekády. Projekt sekvenovania ľudského genómu (2001) a následné projekty zamerané na skúmanie rozmanitosti a funkčnosti genómu spustil pokrok v metódach sekvenovania DNA. Nastúpila doba masívnych paralelizovaných vysokovýkonných prístupov, známych ako sekvenovanie novej generácie (NGS) (Glenn, 2011).

Platforma	Sekvenátor	Náklady na sekv.platformu	Počet čítaní za beh	Výkon za beh	Max. dĺžka čítania [1]	Priemerné trvanie behu
Sanger	ABI 3730xl	$100,000	96	100 kbp	1000 bp	2–3 hours
454	GS FLX	$450,000	1,000,000	700 mpb	1000 bp	24 hours
Illumina	HiSeq 3000	$750,000	300,000,000[2]	150 gbp[3]	250 bp	4 days
Illumina	NextSeq500	$250,000	400,000,000	120 gbp[3]	150 bp	30 hours
Illumina	MiSeq	$100,000	25,000,000	15 gbp[3]	300 bp	24 hours
Ion Torrent	Proton II	$224,000	330,000,000	66 gbp	200 bp	4 hours
Ion Torrent	PGM 318	$50,000	5,000,000	2 gbp	400 bp	7 hours
PacBio	RS II	$700,000	50,000	400 mbp	54 kbp	3 hours
Nanopore	MinION	$1,000	80,000[4]	490 mbp[4]	150 kbp	n.a.[4]

[1] Odhaduje sa pre vysokú kvalitu čítania, jednotlivé čítania môžu byť dlhšie.
[2] Jednoduché čítanie v jednom pruhu, schopné pair-end behu
[3] Výstup pre pair-end beh (v prípade HiSeq jeden pruh).
[4] Doba chodu stroja sa zvyčajne prispôsobuje potrebe hĺbky sekvenovania, napríklad 48-hodinový chod.

Obr. 7: Porovnanie výstupu vybraných sekvenčných platforiem (Bleidorn, 2016).

V evolučných štúdiách sa začalo používať 454 pyrosekvenovanie s vysokým počtom čítaní, ktorých dĺžka bola porovnateľná s tými zo Sangera, čiže približne 1 000 bp. S ešte väčším počtom výstupov dosahujúcim až niekoľko miliárd prečítaní na jeden beh poskytuje Illumina, doteraz používaná metóda založená na sekvenovaní reverzibilného ukončenia (Bentley *et al.*, 2008). Dĺžka čítania je však výrazne kratšia, v najnovšej generácii strojov Illumina sa pohybuje okolo 2x300 bp (Obr. 7). Avšak kvalita čítania je vysoká (Hu *et al.*, 2021).

2.2.1 Metódy sekvenovania tretej generácie

Krátko po vyvinutí NGS metód nastúpila Tretia generácia sekvenovania (TGS). Prišli s jednomolekulovým sekvenovaním (single-molecule sequencing, SMS) bez potreby amplifikácie vzorku a poskytujúce sekvenovanie v reálnom čase (Schadt *et al.*, 2010). Prvú SMS technológiu uviedol na trh Helicos Biosciences, podobajúcu sa na Illuminu ale bez mostíkovej amplifikácie. Táto metóda sa ukázala ako pomerne pomalá, drahá a s krátkymi čítaniami (cca 32 bp) (Pushkarev *et al.,* 2009).

Za pravú prvú metódu TGS sa považuje metóda „single-molecule real- time“ (SMRT) od Pacific Biosciences (PacBio) z roku 2011 (Eid *et al.*, 2009). Ďalšia významná technológia bola predstavená v roku 2014 Oxford Nanopore Technologies (ONT) využívajúca sekvenovanie s nanopórmi (Jain *et al.*, 2016). Okrem toho, že TGS nepotrebujú PCR amplifikáciu a sekvenovanie prebieha v reálnom čase, tieto dve metódy produkujú dlhé čítania, čo spôsobilo revolúciu v genomike (Sedlazeck *et al*., 2018).

2.3 Sekvenovanie pomocou Nanopórov

Sekvenovanie prostredníctvom nanopórov sa zakladá na biologickej membráne, kde sa nachádza nanopór, obklopenej elektrolytickým roztokom (Varongchayakul *et al.*, 2018). Cez membránu sa aplikuje napätie indukujúce elektrické pole, ktoré poháňa častice (Chen *et al.*, 2004). Molekula DNA alebo RNA sa zachytí na nanopóre a postupne ním prechádza. Vnútri póru molekula obmedzuje tok iónov, čo je brané ako pokles iónového prúdu. Veľkosť iónového prúdu sa mení na základe rôznych faktorov, ako je napríklad chemické zloženie, veľkosť a geometria. Tieto zmeny možno nasnímať a potenciálne tak identifikovať rôzne molekuly (Si & Aksimentiev, 2017). Presnosť ONT sekvenovania je oproti metódam NGS relatívne nízka, dĺžka čítania poskytovaná elektrickou detekciou je veľká, keďže spolieha na fyzikálny proces translokácie nukleovej kyseliny (Gong *et al.*, 2019). Čítania dosahujú hodnoty presahujúce 150kb (Jain *et al.*, 2016).

Sekvenovanie s dlhým čítaním je veľmi dobre aplikovateľné na vytvorenie dlhých kontigov, či dokonca na zostavenie úplných genómov, s minimálnou prípravou vzorky a pomerne nízkych nákladov. To zapríčinilo stúpajúcu popularitu práve týchto sekvenčných platforiem. Platforma „nanopore“ založená na Oxford Nanopore Technologies sa tak stáva široko použiteľným nástrojom s veľkou škálou aplikácií (de Koning *et al.*, 2020). Tento dopyt viedol k rýchlemu pokroku, čo prinieslo podstatné zlepšenie v dĺžke a presnosti čítania. Taktiež si vyžiadal rozsiahly vývoj experimentálnych a bioinformatických metód pre úplné využitie dlhých čítaní na skúmanie genómov ale aj transkriptómov, epigenómov a epitranskriptómov. Hlavná oblasť využitia sa vzťahuje na zostavovanie genómu, detekcie transkriptov v plnej dĺžke a modifikácii báz. Používa sa aj v špecializovanejších oblastiach, ako je napríklad rýchla klinická diagnostika (Wang *et al.*, 2021).

ONT vyvinuli prvé dostupné zariadenie na sekvenovanie prostredníctvom nanopórov zvané MinION. Tento sekvenátor má veľkosť malého mobilného telefónu a umožňuje zapojenie do počítača pomocou USB (Obr. 8) . Prvé MinIONy boli distribuované do vybraných laborátorií počas roku 2014 pre beta-testovanie (Feng *et al.*, 2015). Okrem tohoto prenosného zariadenia ONT ponúkajú aj flexibilný stolný sekvenátor GridION, na ktorom beží 5 MinION a rôzne verzie modelu PromethION, vhodné pre priame sekvenovanie DNA a RNA vo veľkom meradle (Bleidorn, 2016).

2.3.1 Vývoj metódy nanopore

Sekvenovanie reťazcov DNA prostredníctvom nanopórov bolo veľkou technickou výzvou a vyvíjalo sa viac ako 25 rokov. Pôvodný koncept sa objavil už v 80 rokoch minulého storočia (Bayley, 2015). Prvý nanopór, ktorý dokázal detegovať blokády iónového prúdu homopolymérmi RNA aj DNA bol nanopór a pridružený motorický proteín 1,4 – 8α- hemolyzín, pochádzajúci z membránového kanála *Staphylococcus aureus* s priemerom 1,4 nm až 2,4 nm. (Meller *et al.*, 2000) Úprava divokého typu α-hemolyzínového proteínu umožnila rozlíšenie štyroch DNA bázií (A, T, C, G) na olihonukleotidových molekulách (Stoddart *et al.*, 2009).

V 90 rokoch bola prvýkrát pozorovaná translokácia nukleových kyselín cez nanopóry, taktiež sa vyvinulo stochastické snímanie a došlo k vyriešeniu štruktúry proteínového nanopóru s vysokým rozlíšením (Bayley, 2015). Zlepšenie pomeru signálu k šumu sa dosiahlo začlenením procesných enzýmov pre spomalenie translokácie DNA cez

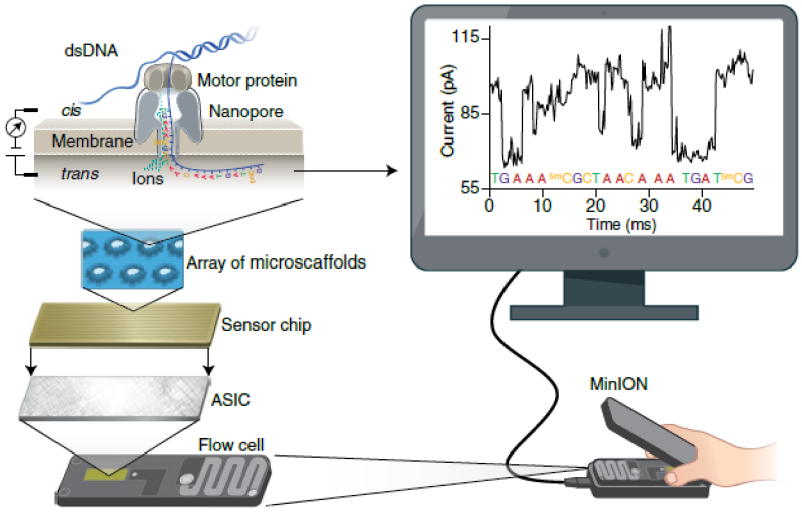
Obr. 8:. MinION zariadenie pripojené do notebooku. (https://nanoporetech.com)

nanopóry (Hornblower *et al.*, 2007). Konkrétne phi29 DNA polymeráza dosahuje vynikajúci výkon v tejto oblasti (Lieberman *et al.*, 2010).

Demonštrácia pokroku vo vývoji bola ukázaná vo februári 2012, kedy dve procesné skupiny prezentovali záznamy iónových prúdov pre jednovláknové molekuly DNA, ktoré mohli byť rozdelené na signály z jednotlivých nukteotidov. A to s kombináciou phi29 DNA polymerázu a nanopór (α- hemolyzín a MspA) (Cherf *et al.*, 2012; Manrao *et al.*, 2012). Ešte v tomto mesiaci ONT oznámila prvý sekvenátor s technológiou používajúcou nanopóry – MinION (Obr. 8) (Mason & Elemento, 2012). Zariadenie vyšlo pre vybraných používateľov v roku 2014 a komercializovalo sa v roku 2015 (Jain *et al.*, 2016).

2.3.2 Princíp „nanopore“ sekvenačnej technológie

Technológia sekvenovania pomocou nanopórov a jej možné aplikácie vo výskume prešli významným rastom od vydania prvého sekvenátora MinION (Jain *et al.*, 2016). Technológia je založená na proteínovom póre v nanorozmere (nanopór) slúžiacom ako biosenzor (Obr. 9). Tento pór je vložený do polymérovej odolnej membrány s elektrolytickým roztokom. V roztoku je aplikované konštantné napätie pre vytvorenie iónového prúdu skrz nanopór tak, aby záporne nabité jednovláknové molekuly DNA alebo RNA boli hnané cez nanopór zo záporne nabitej „cis“ strany na kladne nabitú



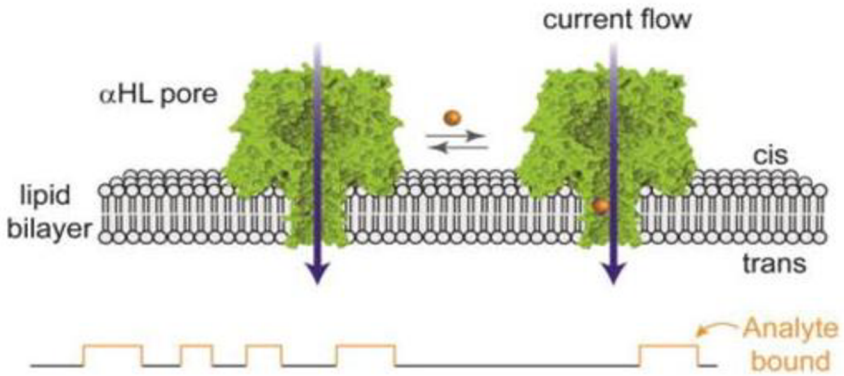
Obr. 9: Princíp technológie nanopór: Flow cell (prietoková bunka) MinION s 512 kanálmi a 4 nanopórmi v každom kanáli – 2 048 nanopórov celkovo. Každý nanopór je v jamke na polymérnej membráne. Každý kanál je spojený so samostatnou elektródou v senzorovom čipe. Ten je riadený a meraný podľa „application-specific integration circuit“ (ASIC) (Wang *et al.*, 2021).

„trans“ stranu. Rýchlosť tohto prenosu je riadená motorickým proteínom, ktorý postupne posúva molekulu nukleovej kyseliny cez nanopór. Nastáva k zmenám v iónovom prúde, čo zodpovedá nukleotidovej sekvencii prítomnej v snímacej oblasti a pomocou výpočtových algoritmov dochádza k dekódovaniu. Tieto algoritmy umožňujú sekvenovanie jednotlivých molekúl v reálnom čase (Wang *et al.*, 2021).

Motorický proteín sa vyznačuje aj helikázovou aktivitou, vďaka ktorej umožňuje rozvinúť dvojvláknové molekuly DNA alebo RNA-DNA do jednovláknových molekúl, pre prechod nanopórom (Wang *et al.*, 2021).

2.3.2.1 Stochaistické snímanie

Pri snímaní je jeden proteínový nanopór umiestnený v lipidovej dvojvstve a monitoruje sa iónový prúd prechádzajúci ním. Molekuly analytu viažuce sa na upravené miesta v póre, sa detegujú prechodnými zmenami prúdu, zvyčajne poklesom. Tieto zmeny sa prejavujú ako blokády štvorcových vĺn a ich frekcencia určuje koncentráciu analytu. Informácie o identite sú obsiahnuté v priemernom trvaní blokád, ich amplitúde a ďalším charakteristikám ako je zvýšenie šumu prúdu (Bayley, 2015).



Obr. 10: Snímanie s α-hemolyzínom

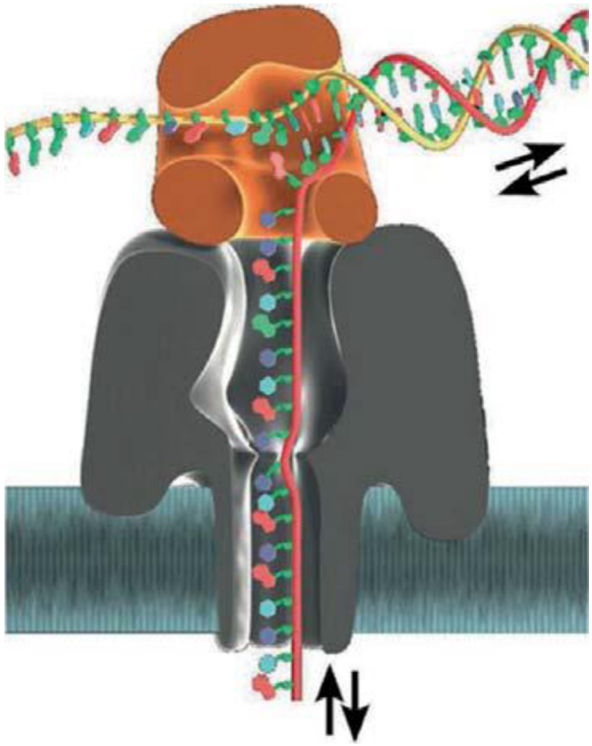
Signály príbuzných analytov sa líšia. Výhodou stochaistického snímania je, že súvisiace analyty možno rozpoznať a zároveň rozlíšiť jediným detekčným prvkom, čo je základom tejto sekvenovacej metódy (Bayley, 2015).

2.3.2.2 Analýza nukleových kyselín s proteínovými nanopórmi

V roku 1996 sa zistilo, že krátke nukleové kyseliny je možno translokovať cez α-hemolyzín (Obr. 10) vplyvom potenciálu. Viacero skupín vedcov skúmalo interakcie DNA a RNA s pórmi a získali dôležité výsledky. Vyplynulo, že jednovláknové (single strand, ss) ale nie dvojvlákové (double strand, ds) nukleové kyseliny môžu byť translokované (Kasianowicz *et al.*, 1996). Zistilo sa aj to, že sa dá určiť smer vstupu (najskôr 5′ alebo 3′) (Wang *et al.*, 2004).

Napriek viacerým zisteniam nedošlo k pokroku v samotnom sekvenovaní takmer 10 rokov. Stimuláciu pokroku značnou mierou zapríčinil 1 000 Genome Project, začatý v roku 2004 spojením zamerania fyzikov na translokáciu a biochemikov na rozpoznávanie molekúl, čo sa prejavilo pri skúmaní stochaistického snímania (Bayley, 2015).

Príprava sekvenčnej knižnice prebieha tak, že sa na jeden koniec molekuly liguje vlásenkový adaptér a na druhý sa liguje motorický proteín. Motorický proteín poháňa molekulu DNA cez nanopór a zároveň ju disociuje dsDNA na ssDNA (Obr. 11). Vlákna molekuly však nie sú úplne oddelené, spája ich vlásenkový adaptér. V ideálnom prípade pórom prechádza jeden reťazec, za ním nasleduje vlásenkový adaptér a potom druhý reťazec. Takto je možné vytvoriť konsenzuálnu sekvenciu pre sekvenovanú dsDNA (Ashton *et al*., 2015).



Obr. 11: Priechod dsDNA cez nanopór (Schneider & Dekker, 2012)

Dĺžka nanopórového tunela a rýchlosť procesu zapríčiňuje prítomnosť viacerých nukleotidov v póre. Zvyčajne sa zaznamená signál prekrývajúcich sa 5-mérov, takže cloudový softvér MinKNOW potrebuje rozlíšiť 4^5 možné stavy iónového prúdu pre všetky možné 5-méry pre vygenerovanie surovej sekvencie. Preto nie je prekvapením, že pre vytvorené čítania je určitá chybovosť, v roku 2016 pohybujúca sa v rozmedzí 25-40%. ďalej sa výstup publikovaných štúdií pohyboval od 90 do 490 Mbp za 48 hodín. Priemerná dĺžka čítania je okolo 6 kbp, maximálna 150 kbp (Ashton *et al.*, 2015; Laver *et al.*, 2015; Quick *et al.*, 2015).

2.4 Analýza dát

Bioinformatická analýza dát ONT sa neustále zlepšuje. Mnohé analýzy pre ONT sa zameriavajú na efektívnejšie využitie signálu získaného z iónového prúdu pre účely ako je „basecalling“, detekcia modifikácii a „data polishing“. Ďalej sú vytvorené nástroje na opravu chýb, zostavenie (assembly) kontigov (aj scaffoldov) a zarovnanie (alignment) sekvencií. ONT zariadenia nevyžadujú „high-end“ výpočtovú techniku alebo pokročilé zručnosti pre základné spracovanie údajov, takže si mnohé laboratória sami môžu spustiť zber údajov (Wang *et al.*, 2021).

MinKNOW je operačný softvér používaný pre ovládanie ONT zariadení nastavením parametrov sekvencie a sledovanej vzorky. Spravuje získavanie údajov, analýzu v reálnom čase, vykonáva lokálny „base calling“ a výstup ukladá v binárnych súboroch vo formáte fast5 pre uloženie metaúdajov. Formát fast5 organizuje údaje vnoreným spôsobom, čím umožňuje prístup k informáciám bez potreby navigácie cez súbor údajov (Ip, 2015).

2.4.1 „Base calling“

Volanie báz (base calling) dekóduje aktuálny signál na nukleotidovú sekvenciu a je rozhodujúce pre presnosť údajov a detekciu modifikácií. Vývoj tejto metódy prešiel štyrmi fázami:

1. Základné volanie aktuálnych údajov s HMM (hidden Markov model) v počiatočnom štádiu a využitie neurónovej siete koncom roka 2016
2. Základné volanie z nespracovaných aktuálnych údajov v 2017
3. Použitie flip-flop modelu pre identifikáciu jednotlivých nukleotidov v 2018
4. Tréning prispôsobených modelov pre základné volanie v 2019

ONT vyvinula nové volanie báz ako softvér (Nanonet, Scrappie a Flappie). Tie boli následne implementované do dostupných softvérových balíkov, najznámejšie sú Guppy a Albacore. Vývoj Albacore bol ukončený v prospech Guppy, ktorý je rýchlejší (Wang *et al.*, 2021).

2.4.2 Detegovanie DNA a RNA modifikácií

ONT umožňuje detekciu niektorých modifikácií DNA a RNA odlíšením ich súčasných posunov od posunov báz nemodifikovaných. Bolo vyvinutých niekoľko nástrojov pre túto detekciu modifikácií. Prvým nástrojom bol Nanoraw pre identifikáciu modifikácií DNA 5mC, 6mA a N4- metylcytozín. Ďalej nasledoval vývoj viacerých nástrojov pre iné možné modifikácie. Napríklad Nanopolish s 5mC a mCaller pre 5mC a 6mA (Stolber *et al.*, 2016).

2.4.3 Oprava chýb

Priemerná presnosť sekvenovania ONT sa stále zlepšuje, ale určité čítania alebo čítané fragmenty stále vykazujú chybovosť a nízku presnosť. Oprava chýb sa používa pred následnou analýzou dát (napr. zostavenie genómu) pre lepšiu citlivosť a kvalitu

výsledkov. Používajú sa 2 typy algoritmov: autokorekcia a hybridná korekcia (Fu *et al.*,2019; Lima *et al.*, 2020).

Autokorecia využíva prístup založený na grafoch pre vytvorenie konsenzuálnych sekvencii s rôznymi molekulami s rovnakým pôvodom (napríklad Canu). Hybridná korekcia používa vysoko presné krátke čítania pre opravu dlhých čítaní založených na zarovnaní (alignment). Lepšou stratégiou sa vidí hybridná korekcia, ktorá môže znížiť chybovosť dlhého čítania na úroveň 1-4%, zatiaľ čo auto-korekcia znižuje chybovosť na 3-6% (Fu *et al.*,2019; Lima *et al.*, 2020).

2.4.3.1 Zarovnanie (alignment) pre dlhé čítanie

Pre riešenie chybovosti dlhých čítaní boli vyvinuté zarovnávacie nástroje. Veľmi skoré zarovnávače boli vyvinuté pre malý počet dlhých čítaní (napríklad BLAST) (Altschul *et al.*, 1990). Avšak aj v tejto oblasti dochádza k výraznému pokroku. Vývoj bol pôvodne motivovaný hlavne sekvenovaním PacBio a vzniknuté nástroje sa testovali aj na ONT údajoch. Prvý zarovnávač pre ONT bol vyvinutý v roku 2016 – GraphMap (Sović *et al.*, 2016).

Pomocou prístupu „seed-chain-align“ bol vytvorený minimap2, tak aby zodpovedal zvyšujúcej sa dĺžke čítania ONT. Ukázalo sa, že práve minimap2 beží oveľa rýchlejšie ako iné zarovnávače pre dlhé čítania bez zníženia presnosti. Zhromažďuje minimalizátory referenčných sekvencií a indexuje ich do hašovacej tabuľky, kľúčom je hash minimalizátora a hodnotou je zoznam umiestnených kópií minimalizátora. Potom minimap2 pre každú sekvenciu hľadá presné zhody s referenciou (Li, 2016).

2.4.3.1.1 SAM formát

Formát „Sequence Alignment/Map“ je určený pre efektívne uchovanie informácií zo sekvenčného zarovnania. Pozostáva z hlavičky začínajúcou znakom „@“ a časti so zarovnaním. Všetky riadky sú oddelené tabulátorom a každý má 11 povinných polí (Li *et al.,* 2009).

Pole CIGAR je definované ako šifrované zarovnanie dané operáciami: „M“ pre zhodu/nesúlad, „I“ pre insert, „D“ pre vymazanie, „N“ pre vynechané báze, „S“ pre soft-clipping (mäkké orezávanie), „H“ pre hard-clipping (tvrdé orezávanie) a „P“ pre výplň (Li *et al.,* 2009).

2.4.3.2 Štruktúrne variácie (SV) a opakujúce sa oblasti

Ak je dostupný referenčný genóm, údaje z ONT možno použiť pre štúdium genómových údajov špecifických pre vzorku a to vrátane SV a haplotypov, s presnosťou väčšou ako iné techniky. Bolo vyvinutých viacero nástrojov na detekciu SV, napríklad NanoSV a Sniffles. Taktiež sa vyvinuli nástroje pre repetetívne genómové oblasti ako je TLDR. Tieto nástroje bežia dobre hlavne pre dáta z vzoriek ľudskej DNA (Gong *et al.*, 2018).

3 EXPERIMENTÁLNA ČASŤ

3.1 Sekvenačné dáta

Spracované dáta boli získané izoláciou genómovej DNA, prípravou knižnice a sekvenovaním za pomoci platformy Oxford Nanopore v Centre štruktrúrnej a funkčnej genomiky rastlín, Ústave experimentálnej botaniky (UEB) AV ČR, v. v. i. a CR Haná v Olomouci.

V diplomovej práci sa pracovalo s 2 zástupcami zo sekcie *Callimusa* (*coccinea* a *beccarii*), 2 zástupcami *Australimusa* (*textilis* a *maclayi*), 2 zástupcami *Rhodochlamys* (*ornata* a *laterita*) a 1 kultivar *Ensete*. Zo sekcie *Eumusa* bolo zástupcov najviac a to 12. Medzi nimi bolo viacero kultivarou *M. acuminata*, 1 zástupcu *M. balbisiana* a triploidné hybridy: AAA a AAB (Tab. 1).

Tab. 1: Analyzovaní zástupcovia *Musa* spp.

ITC	Rod	Druh / Skupina	Poddruh / Podskupina	Názov položky	Sekcia	2C (pg)	Počet chromo-zómov
0249	*Musa*	*acuminata*	*burmanni-ccoides*	Calcutta 4	*Eumusa*	1,226	2n=2x=22
0728	*Musa*	*acuminata*	*zebrina*	Maia Oa	*Eumusa*	1,325	2n=2x=22
0806	*Musa*	*acuminata*	*banksii*	Banksii	*Eumusa*		2n=2x=22
0248	*Musa*	*balbisiana*		Singapuri	*Eumusa*	1,153	2n=2x=22
0145	*Musa*	AAA	EAHB	Nshika	*Eumusa*	1,935	2n=3x=33
0575	*Musa*	AAA	Red/Green Red	Red Dacca	*Eumusa*	1,836	2n=3x=33
1482	*Musa*	AAA	Cavendish	Poyo	*Eumusa*	1,8	2n=3x=33
0022	*Musa*	AAB	Plantain	Mululou	*Eumusa*	1,7	2n=3x=33
0642	*Musa*	AAB	Plantain	Hartón Tigre	*Eumusa*	1,7	2n=3x=33
1132	*Musa*	AAB	Plantain	3 Hands Planty	*Eumusa*	1,786	2n=3x=33
0370	*Musa*	*ornata*		Musa ornata	*Rhodochlamys*	1,9	2n=2x=22
1575	*Musa*	*laterita*		Musa laterita	*Rhodochlamys*	1,315	2n=2x=22
0539	*Musa*	*textilis*		Musa textilis	*Australimusa*	1,435	2n=2x=22
0614	*Musa*	*maclayi*		Musa maclayi type Hung Si	*Australimusa*	1,476	2n=2x=22
1716	*Musa*			Fe'i	*Australimusa*	2,2	2n=3x=33
1070	*Musa*	*beccarii*	*beccarii*	Musa beccarii	*Callimusa*	1,561	2n=2x=18
0287	*Musa*	*coccinea*		Musa coccinea	*Callimusa*	1,442	2n=2x=20
1387	*Ensete*	*ventricosum*		Ensete ventricosum	*ventricosum*	1,210	2n=2x=18

Ako referenčná genómová sekvencia bola použitá celogenómová sekvencia dihaploidného druhu s prístupovým názvom DH- Pahang (verzia 4; Belser *et al.*,2021). Je to divoká rastlina patriaca k poddruhu *Musa acuminata* spp. *malaccensis*.

3.2 Bioinformatická analýza

Použité bioinformatické programy boli spustené pomocou serveru samsom Centra štruktúrnej a funkčnej genomiky rastlín, patriace k virtuálnej organizácií MetaCentrum (MetaVO). Príkazy boli zadané cez príkazový riadok Ubuntu 22.04.2 LTS.

3.2.1 Volanie bázií (base calling)

Použité dáta boli komprimované pomocou programu Gzip, vo forme '.tar.gz' a dekompresnuté príkazom:

```
tar -xvzf fast5_pass.tar.gz
```

čo tento súbor rozbalilo v podrobnom režime (`-v`), takže sa pri rozbaľovaní súbory vytlačili do príkazového riadku a uložili do nového priečinku. Tento priečinok bol plný len FAST5 súborov s označením „pass" a aj po hlbšej analýze sa neukázalo, že by obsahovala súbory s označením „fail".

Na takto rozbalených jednotlivých behoch ONT pre rôzne kultivary banánovníkov sa spustil Guppy basecaller, pomocou skriptu poskytnutého UEB a upraveného pre jednotlivé behy (príklad viď Príloha 1). Výpočet prebehol na serveroch MetaCentra a behy sa líšili v potrebnom čase na spracovanie, pohybujúceho sa od 30 minút k 40 hodinám. Výsledné súbory boli vo formáte FQ a komprimované s Gzip. Jednotlivé behy boli spojené do jedného súbor s príkazom `cat`.

Získané dáta boli analyzované s programom NanoPlot- súbor nástrojov špeciálne vytvorený na vizualizáciu a spracovanie sekvenačných údajov dlhých čítaní od Oxford Nanopore a Pacific Biosciences. Tento nástroj poskytuje komplexný štatistický súhrn údajov o celkovom počte čítaní, rozložení dĺžok čítania a podobne (De Coster *et al.*, 2018).

3.2.2 Zostavenie kontigov (assembly)

Okrem práce s prvotnými dátami prebehlo zostavenie dlhých kontigov z ONT dát. Pre túto prácu bol použitý Flye *de novo* assembler (verzia 2.8), vhodný pre čítania zo sekvenovania metódami ONT a PacBio. Balík predstavuje kompletnú pipeline, berie

surové ONT čítania ako vstup a výstupom sú dlhšie zostavené kontigy, s vyššou kvalitou sekvencie (Kolmogorov *et al.*, 2019).

Balík bol spustený príkazom:

```
flye --nano-raw vstup.fq.gz --out-dir výstupný_adresár --genome-size XYZm --min-overlap 3000 --threads 42
```

kde:

`--nano-raw` určuje vstupné údaje, ktoré má Flye použiť na zostavenie genómu. V tomto prípade ide o dáta generované ONT sekvenátorom a neboli nijako spracované,

`--out-dir` udáva cestu do adresára, kde sa uložia výstupné údaje,

`--genome-size` zadáva odhadovanú veľkosť zostavovaného genómu,

`--min-overlap` nastavuje minimálnu dĺžku prekrytia medzi dvoma čítaniami, ktoré Flye použije na zostavenie genómu. Používalo sa nastavenie pre minimálne prekrytie 3 000 párov báz (defaultné nastavenie je 5 000 párov bázií),

`--threads` určuje počet vlákien (alebo jadier CPU), nastavilo sa na 42 vlákien.

Pre zostavenie pomocou Flye bol napísaný skript, prispôsobený pre skúmaný kultivar, ktoré sa následne spustili na serveroch MetaCentra (príklad v Prílohe 2).

3.2.3 Sekvenčné zarovnanie a identifikácia translokácií

Jedným z hlavných cieľov práce je nájsť cestu, ktorá by umožnila nájdenie veľkých chromozomálnych translokácií v genómoch banánovníka. Sprvu bolo zvažované rozbehnúť cestu používajúcu nástroj NanoSV. Po viacerých pokusoch a úpravách sa ukázala ako neúčinná. Jedným z dôvodov neúspechu je zrejme skutočnosť, že tento program už nie je aktualizovaný a prioritne vyhľadáva jednonukleotidové polymorfizmy (SNP). Preto sa navrhla nová cesta pracujúca s formátom SAM a nástroj SAMtools (Danecek *et al.*, 2021).

Pre sekvenčné zarovnanie bol použitý všestranný program Minimap2 verzia 2.24, ktorý slúži na zarovnanie pre mapovanie DNA oproti veľkej referenčnej databáze (Li, 2016). Spustil sa príkazom:

```
minimap2 -ax map-ont M_acuminata_pahang_v4.fasta assembly.fasta > assembly.sam
```

pre ONT genomické čítania (`-ax map-ont`) zostavené pomocou Flye (`assembly.fasta`) s referenčným genómom DH-Pahang verzia 4 (`M_acuminata_pahang_v4.fasta`). Ako výstupný formát sa zvolil SAM (`assembly.sam`) – mapa zarovnania sekvencií.

SAM je formát ponúkajúci viacero možností skúmania. Okrem prvotného mapovania zahŕňa aj informácie o sekundárne mapovaných sekvenciách, ich bodov zlomu (breakpoints) a reprezentácií v referencii. Po skúmaní pomocou rôznych skriptov sa došlo k záveru, že by bolo dobré zamerať sa na tie mapované sekvencie, ktoré na začiatku ich popisu s CIGAR obsahujú „soft-clipping“ (mäkké orezávanie). To môže pomôcť k identifikácií štruktúrnych variácií. Bol spustený SAMtools príkaz:

```
samtools view assembly.sam | awk 'BEGIN {FS="\t";OFS="\t"} {soft=0;for(i=12;i<=NF;i++){if(substr($i,2)=="S"){soft+=substr($i,1,length($i)-1)}}; if($10!="*" && soft/length($10)<=0.1){print $0}}' | cat header.sam - | samtools view -bS > softclipped.bam
```

kde:

`view` zobrazuje a konvertuje súbory SAM/BAM/CRAM,

`BEGIN {FS="\t";OFS="\t"}` nastavuje oddeľovače vstupných a výstupných polí na tabulátory,

`{soft = 0; for (i=12;i <= NF; i++) {if(substr ($i,2) == "S") { soft += substr ($i, 1, length ($i) -1)}}` vypočíta celkovú dĺžku „soft-clipped“ bázií v zarovnanom kontigu,

`if($10!="*" && soft/length($10)<=0.1)` odfiltruje kontigy s vysokým podielom „soft-clipped“ bázií – viac ako 10% z celkovej dĺžky čítania,

`{print $0}}` vytlačí celý riadok súboru SAM pre kvalifikované kontigy,

`| cat header.sam - |` prenesie výstup predchádzajúceho príkazu do cat, čo zreťazí súbor hlavičky SAM (`header.sam`) s filtrovanými údajmi

`samtools view -h` prenesie zreťazené údaje SAM do zobrazenia SAM Tools, ktoré ich skonvertuje do formátu SAM a zoradí výstup podľa genómovej pozície.

Ďalším krokom je vyfiltrovanie len tých kontigov (prípadne scaffoldov), ktoré presahujú určitú dĺžku. Mi sme si zvolili minimálnu dĺžku 1 000 kb s príkazom:

```
samtools view -h softclipped.sam | awk 'length($10) > 1000000 || $0 ~ /^@/' | samtools view -h - > softclipped1Mb.sam
```

Na dvoch skúšobných poddruhoch sa testovala dĺžka, ktorá bude vhodnejšie – 100 kb a 1 000 kb. Ako vhodnejšia sa ukázala dĺžka 1 000 kb, s ktorou sa pracovalo ďalej.

Vyfiltrované dáta sa pomocou `samtools fasta` konvertujú na FASTA formát a znovu sa použije Minimap2, tentokrát však s novo vzniknutým FASTA súborom. Referenčný genóm zostáva rovnaký:

```
minimap2 -ax map-ont M_acuminata_pahang_v4.fasta softclipped1Mb.fasta > softclipped_Musa.sam
```

Vzniknutý SAM je zaujímavý tým, že mená kontigov a scaffoldov sa môžu opakovať ale líšiť primárne namapovaným chromozómom. Zo SAM sekvencí sme vybrali tie, ktoré dosahovali požadovanú dĺžku. V našej práci sme vybrali minimálnu dĺžku 30 kb (`samtools view -h softclipped_Musa.sam | awk 'length($10) > 30000 || $0 ~ /^@/' | samtools view -h - > softclipped_Musa_30kb.sam`) a táto vlastnosť súboru bola využitá a pomocou príkazu:

```
Samtools view -h softclipped_Musa_30kb.sam |awk 'BEGIN {FS="\t";OFS="\t"} ($1=="@SQ" && ! seen [$2] ++) {print} ($1!="@SQ") {if( $1 == last_contig && $3! = last_chrom) print last_line $0; last_contig = $1; last_chrom = $3; last_line = $0"\n"}' > translokovane_kontigy.sam
```

sa vybrali práve tie riadky zo súboru, ktoré mali rovnaké meno ale líšia sa namapovaným chromozómom.

- Zabezpečuje, že nový súbor bude mať rovnakú hlavičku so `samtools view -h`, `awk` manipuluje vstupnými textovými údajmi na základe špecifický podmienok, ktoré sa definujú v jednoduchých úvodzovkách,
- `'BEGIN{FS="\t";OFS="\t"}:` definuje oddeľovač vstupného (FS) a výstupného (OFS) poľa ako tabulátor,
- ($1=="@SQ" && !seen[$2]++){print}: kontroluje, či prvé pole vstupných údajov je „@SQ“ a či druhé pole nebolo videné predtým. Ak je podmienka pravdivá, vytlačí vstupný riadok. Tento príkaz extrahuje riadky hlavičky s informáciami o referenčnej sekvencii,
- ($1!="@SQ"): kontroluje či prvé pole vstupných údajov nie je „@SQ“,

- `{if($1==last_contig && $3!=last_chrom) print last_line $0; last_contig=$1; last_chrom=$3; last_line=$0"\n"}`: zabezpečuje, či sa prvé pole vstupných údajov rovná last_contig a tretie pole sa nerovná last_chrom. Ak je podmienka pravdivá, vypíše sa posledný riadok nasledovaný aktuálnym riadkom (`$0`). Premenná last_line sa aktualizuje tak, aby obsahovala aktuálny riadok pre ďalšie porovnanie a premenné last_contig a last_chrom tak, aby obsahovali aktuálne hodnoty pre ďalšie porovnanie. Tento príkaz zmení poradie zarovnaní v súbore SAM tak, aby všetky zarovnania pre rovnakú referenčnú sekvenciu boli zoskupené.

Zo vzniknutého súboru boli ďalej vybrané mená kontigov, prípadne scaffoldov, do textového súboru. Vykonal tak príkaz:

```
awk '{print $1}' translokovane_kontigy.sam| sort -u > names.txt
```

Pomocou mien kotigov sa z pôvodného FASTA súboru (assembly.fasta) vytiahnu celé sekvencie zostavených kontigov so skriptom napísaným v programovacom jazyku python (Obr. 12). Použitá verzia pythonu bola python/3.6.2.

```python
import re
# načítanie názvov kontigov, ktoré chceme vybrať
with open('names.txt', 'r') as f:
    names = [line.strip() for line in f]
# Otvorenie vstupného súboru FASTA a výstupného súboru
with open('assembly.fasta', 'r') as f, open('translocations.fasta', 'w') as out_file:
    current_name = None
    current_seq = []
    # Iterácia každého riadku v súbore
    for line in f:
        # Skontroluj, či je riadok riadkom hlavičky
        if line.startswith('>'):
            # Extrahuj názov sekvencie z riadku hlavičky
            seq_name = re.findall(r'^>(\S+)', line)[0]

            # Skontroluj, či sa má aktuálna sekvencia zapísať do výstupného súboru
            if current_name in names:
                out_file.write(f'>{current_name}\n')
                out_file.write(''.join(current_seq) + '\n')

            # Obnov informácie o aktuálnej sekvencii
            current_name = seq_name
            current_seq = []

        else:
            # Pripoj aktuálny riadok k aktuálnej sekvencii
            current_seq.append(line.strip())

    # Skontroluj, či sa má konečná sekvencia zapísať do výstupného súboru
    if current_name in names:
        out_file.write(f'>{current_name}\n')
        out_file.write(''.join(current_seq) + '\n')
```

Obr. 12: Kód pre získanie DNA sekvencií

```python
import re
# Otvor SAM a výstupný súbor
with open("trasnloc.sam", "r") as sam_file, open("output_table.xlxs", "w") as output_file:
    # Hlavička tabulky
    output_file.write("Contig\tChromosome\tStart Position\tBreakpoint\tLength\tIdentity\n")
    # Zreťazenie
    for line in sam_file:
        # Preskoč hlavičku "@" (header)
        if line.startswith("@"):
            continue
        # Rozdeľ riadok
        fields = line.strip().split("\t")
        # Extrakcia: contig name, chromosome name, starting position, a identity
        contig_name = fields[0]
        chromosome_name = fields[2]
        start_pos = int(fields[3])
        # Výpočet dĺžky namapovaného kontigu
        cigar = fields[5]
        contig_length = 0
        for count, op in re.findall("(\d+)([MIDNSHP=X])", cigar):
            if op in "MDN=X":
                contig_length += int(count)
        # Nájdi breakpoint (bod zlomu)
        breakpoint = start_pos + contig_length - 1
        # Extrakcia identity
        identity = None
        for tag in fields[11:]:
            if tag.startswith("NM:i:"):
                identity = 1 - (int(tag[5:]) / contig_length)
        # Formátuj výstup ako tabuľku
        row = "{}\t{}\t{}\t{}\t{}\t{}\n".format(contig_name, chromosome_name, start_pos, breakpoint, contig_length, identity)
        output_file.write(row)
```

Obr. 13: Kód pre extrakciu informácií zo súboru SAM

Kód sa napokon spustil v príkazovom riadku: `python script.py`

3.2.4 Získanie informácii o translokáciach

Ďalším krokom práce bolo zistiť informácie ohľadom kontigov a scaffoldov, ktoré boli nájdené pomocou opísanej cesty. Pre túto úlohu bol napísaný skript v jazyku Python (Obr. 13), ktorý zo súboru SAM vytiahne dáta týkajúce sa daných úsekov. Zacielilo sa na meno úseku, jeho dĺžku, chromozóm, kde sa úsek namapoval, bod zlomu, kde translokácia mohla nastať a identitu.

3.3 Vizualizácia translokácií

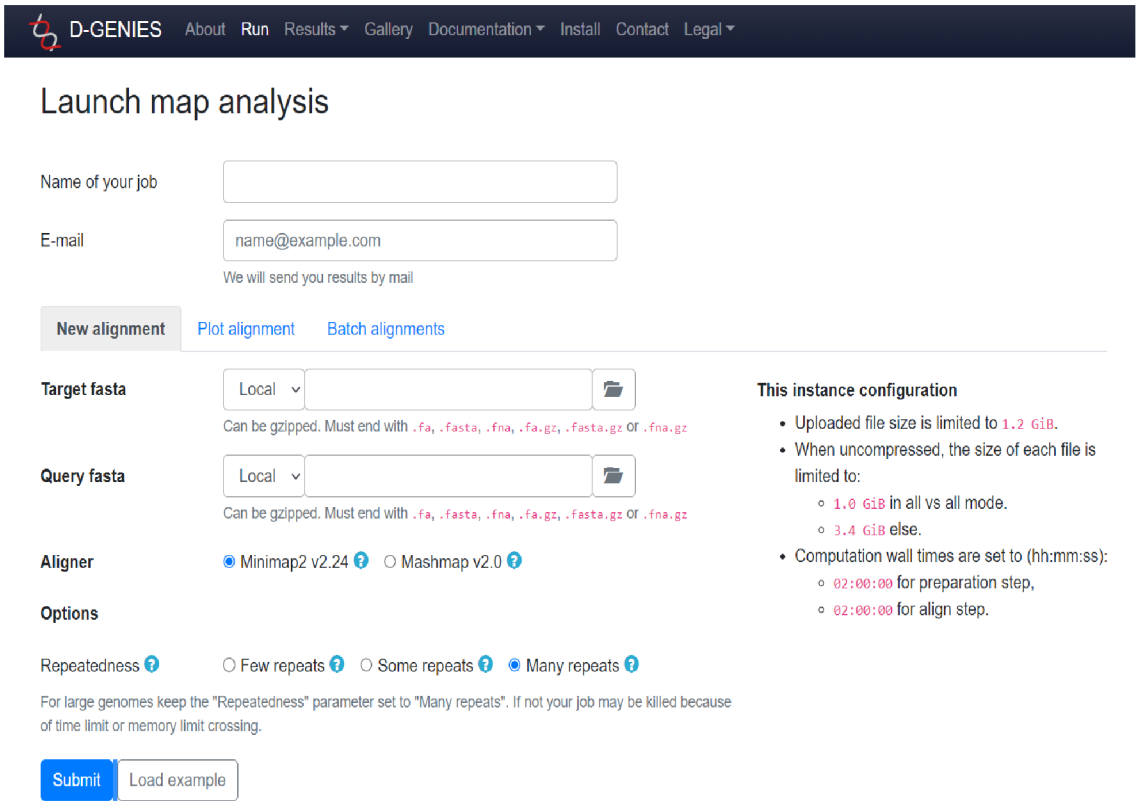
Výsledky získané opísanou cestou boli vizualizované s pomocou webových aplikácií D-Genies. Tento nástroj vytvára Dot plot pre veľké genómy. Bodové grafy (Dot Plot) sa bežne používajú na vizuálne porovnanie dvoch súborov sekvencií. Prezentujú štrukturálne variácie, akými sú translokácie, inverzie a delécie ľahko zrozumiteľným spôsobom. Rozdiely v podobnosti predstavujú pomocou variabilnej hrúbky, tvaru alebo farieb čiary (Cabbanetes & Klopp, 2018).

D-Genies využíva minimap2 a dokáže vytvoriť len bodové grafy pre nukleové sekvencie. Obmedzuje spotrebu pamäte a skracuje čas rozdelením sekvenčných dotazov

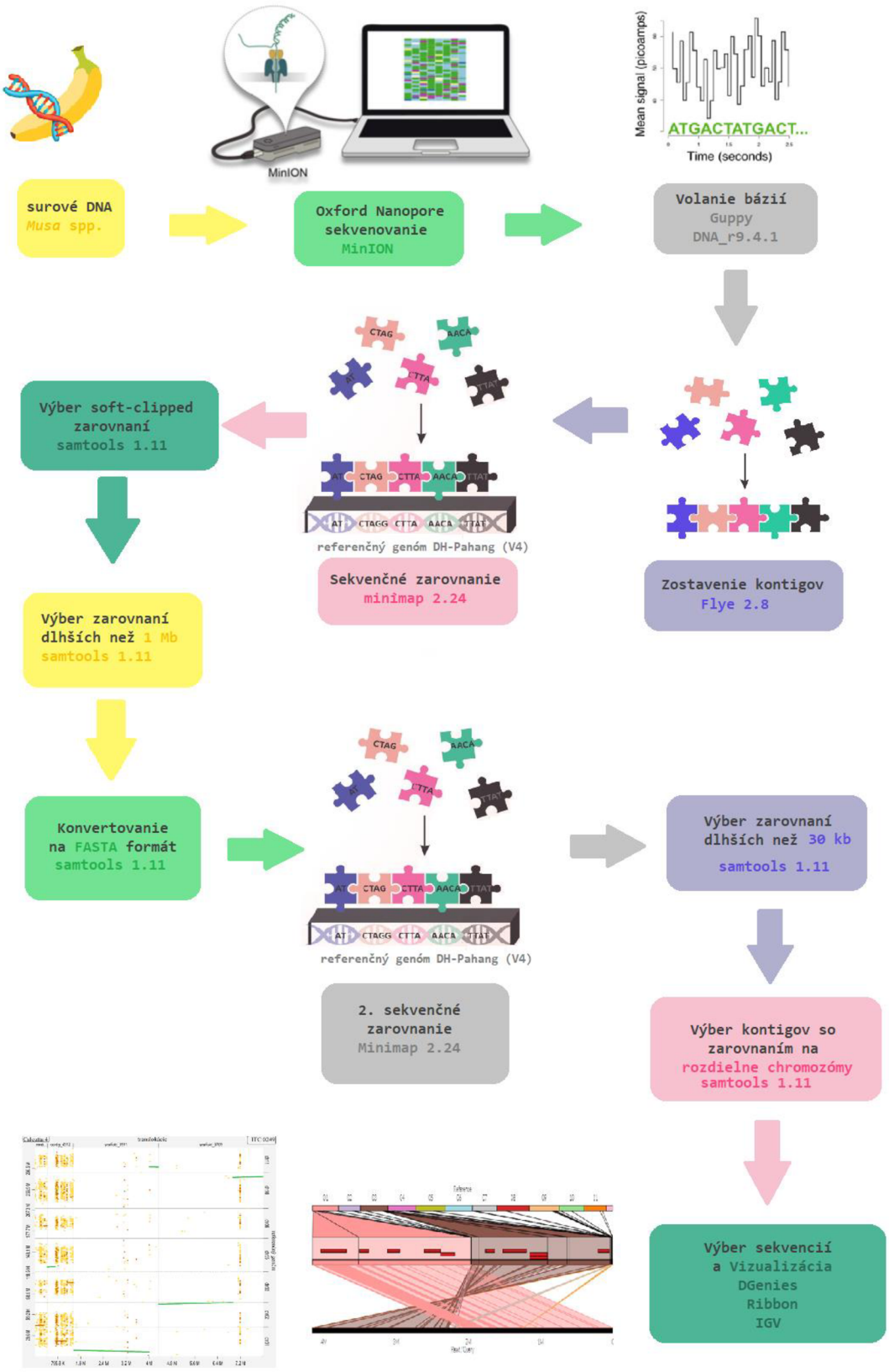
– chromozómov, na 10 častí (V našom prípade 11 chromozómov – 11 častí). Webová aplikácia je prístupná cez webový prehliadač (Cabbanetes & Klopp, 2018).

Použili sme režim „New alignment“ (Obr. 14) môžu byť súbory dotazu aj cieľového súboru FASTA nahrané z lokálneho počítača a vo formáte gzip (Cabbanetes & Klopp, 2018). Ako dotaz sa zobral FASTA výsledný súbor a ako referencia bola použitá genómová sekvencia DH-Pahang.

Vytvorenú cestu (pipeline) v zjednodušenej forme sme znázornili na Obrázku 15.



Obr.14: D-Genies webový nástroj (obrázok webovej stránky https://dgenies.toulouse.inra.fr/run)



Obr. 15: Vytvorená cesta (pipeline) pre vizualizácie translokácií získaných z assembly ONT dát

3.3 Lokalizácia translokácií u jednotlivých ONT čítaní

Pre jednotlivé ONT čítania sa skúšala rovnaká cesta na dátach získaných z volania bázií. Výsledky nedosahovali rovnakých kvalít ako to bolo u dát, ktoré boli sekvenčne zarovnané s nástrojom Flye.

Pre všetky kultivary sa preto najskôr vyskúšala vytvorená cesta s assembly a jej výsledky sa aplikovali na čítania. Prvým krokom bolo sekvenčné zarovanie s minimap2, kde ako referencia boli použité výsledné kontigy, na ktoré sa mapoval súbor s jednotlivými ONT čítaniami a to s príkazom:

```
minimap2 -ax map-ont translokovane_kontigy.fasta ONT_citania.fq.gz > alignment.sam
```

Zo vzniknutého súboru sa vybrali namapované čítania, ktoré presahovali priemernú dĺžku čítania skúmaného kultivaru (napr. 10 kb) pomocou vyššie popísaného príkazu : `samtools view -h alignment.sam | awk 'length($10) > 10000 || $0 ~ /^@/' | samtools view -h - > alignment_10kb.sam`

Tento súbor sa ďalej preformátoval na FASTA : `samtools fasta alignment_10kb.sam > alignment_10kb.fasta`

Takto sa získali len tie čítania, ktoré súvisia so skúmanými kontigmi a majú určitú dĺžku. Po úprave sa namapovali na referenčnú genomickú sekvenciu s minimapom: `minimap2 -ax map-ont referencny_genom.fasta alignment_10kb.fasta > translokovane_citania.sam`

a znova sa vybrali len namapované čítania, ktoré dosahovali aspoň priemernú dĺžku čítania daného kultivaru. SAM sa konvertoval na BAM, triedil so `samtools sort` a indexoval so `samtools index`. Oba súbory boli stiahnuté, a spolu so SAM súborom translokovaných kontigov získaných cestou so sekvenčným zarovnaním (Obr. 15) a referenčným genómom FASTA boli načítané do integračného genomického prehliadaču IGV.

IGV je vizualizačný nástroj umožňujúci intuitívne skúmanie rôznorodých rozsiahlych súborov genómových údajov na počítači v reálnom čase. Efektívne využíva formáty súborov, v reálnom čase ich skúma a to pri minimálnej spotrebe zdrojov počítača (Robinson *et al.*, 2011).

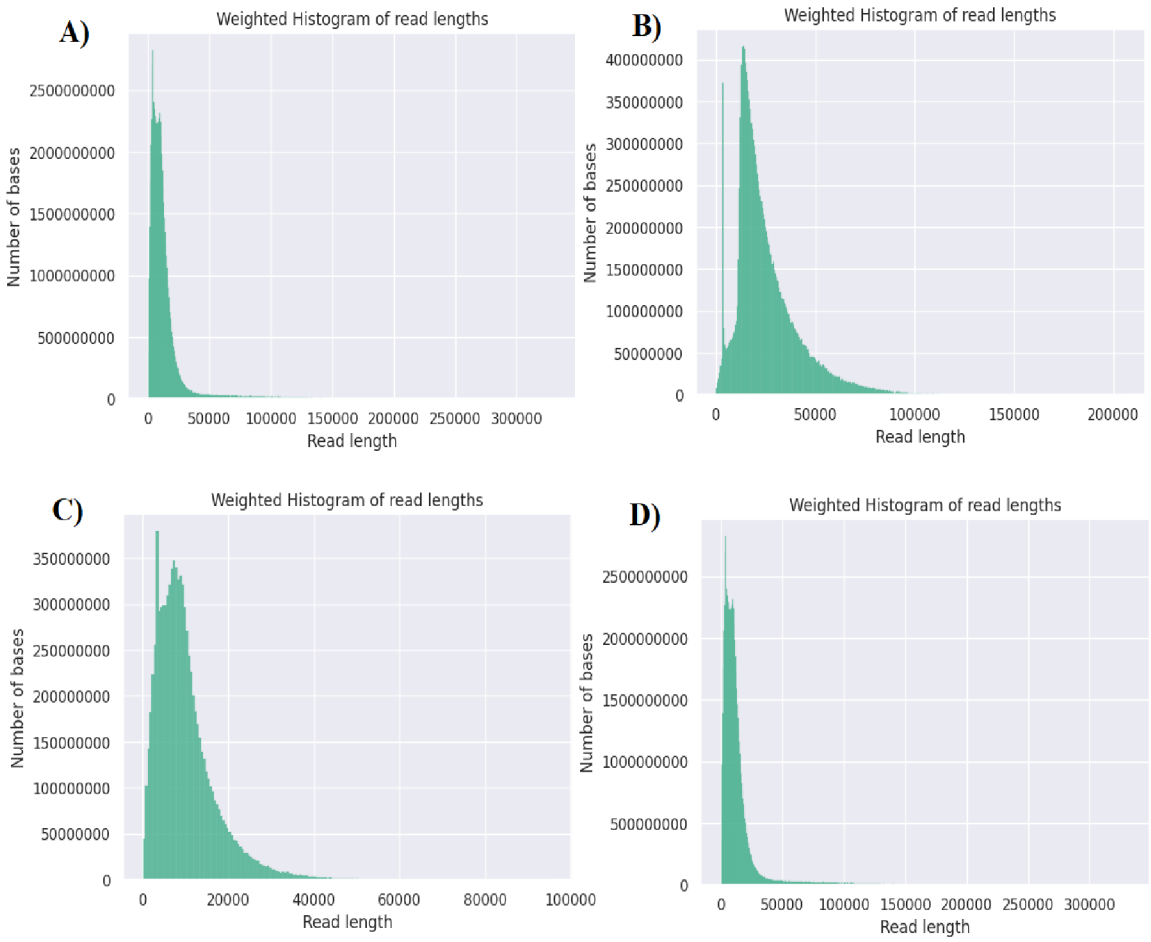
4 VÝSLEDKY

4.1 Analýza sekvenačných dát

Pre analýzu sekvenčných údajov bol použitý nástroj NanoPlot, ktorý poskytuje komplexný prehľad o ich kvalite a dĺžke (Tab. 2). Priemerná dĺžka čítania bola stanovená na cca 10 199 bázií a kvalitu 12,96. Zaujímavé bolo, že u viacerých kultivarov, ktoré mali priemernú dĺžku čítania okolo 5 500 bázií, sa zvyšovala kvalita a celkový počet čítaní. Nedalo by sa však uvážiť, že so znižujúcou sa priemernou dĺžkou čítania by stúpala kvalita a naopak. Minimálnu priemernú dĺžku mal kultivar *M. maclayi* s 5 255,4 (Obr. 16A) a naopak, kultivar *M. beccarii* mal priemer okolo 15 830,7 (Obr. 16B). Najmenej bázií bolo sekvenovaných pre *M. ornata* (Obr. 16C) zo sekcie *Rhodochlamys* s ôsmymi miliardami. Najviac bázii sa sekvenovalo pre triploid (AAB) 3 Hands Planty (Obr. 16D) s vyše 75 miliardami.

Tab. 2: Štatistika údajov získaných s NanoPlot

Priemerná dĺžka čítania	10 198,73	**minimum**	5 255,4
		maximum	15 830,7
Priemerná kvalita čítania	12,965	**minimum**	12,5
		maximum	13,7
Stredná dĺžka čítania	8 871,7	**minimum**	3 103
		maximum	14 232
Stredná kvalita čítania	12,83	**minimum**	12,4
		maximum	13,6
Priemerný počet čítaní	2 964 112,85	**minimum**	800 218
		maximum	13 709 551
Priemerná N50 dĺžka čítania	14 083,95	**minimum**	7 653
		maximum	21 496
STDEV dĺžka čítania	7 989,85	**minimum**	3 943
		maximum	12 875
Priemerný počet bázií	24 671 265 549	**minimum**	8 832 549 817
		maximum	75 941 883 372

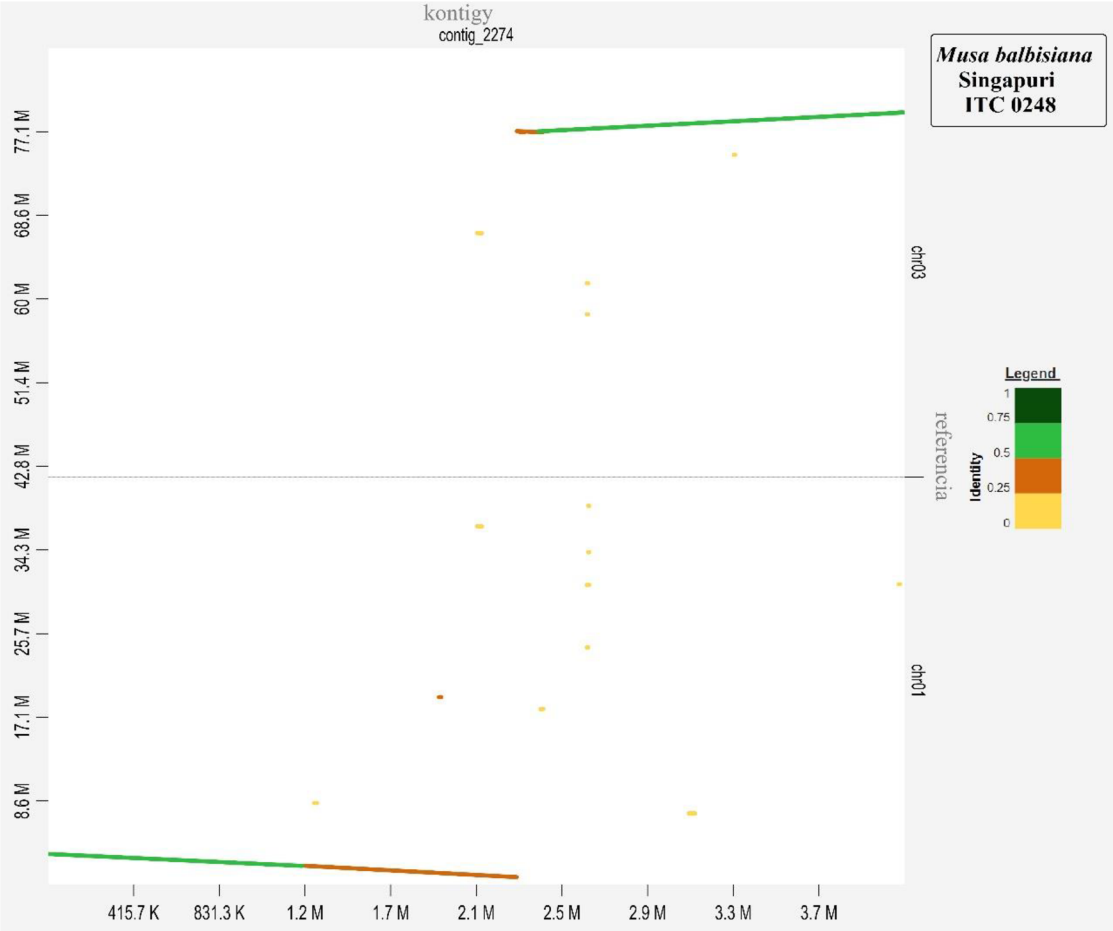


Obr. 16: Histogramy znázorňujúce rozloženie ONT čítaní pre: A) *Musa maclay (*ITC 0614) B) *Musa beccarii* (ITC 1070) C) *Musa ornata* (ITC 0370) D) *Musa* 3 Hands Plenty (ITC 1132)

4.2 Vizualizácia translokácií s assembly

4.2.1 Druhy zo sekcie *Eumusa*

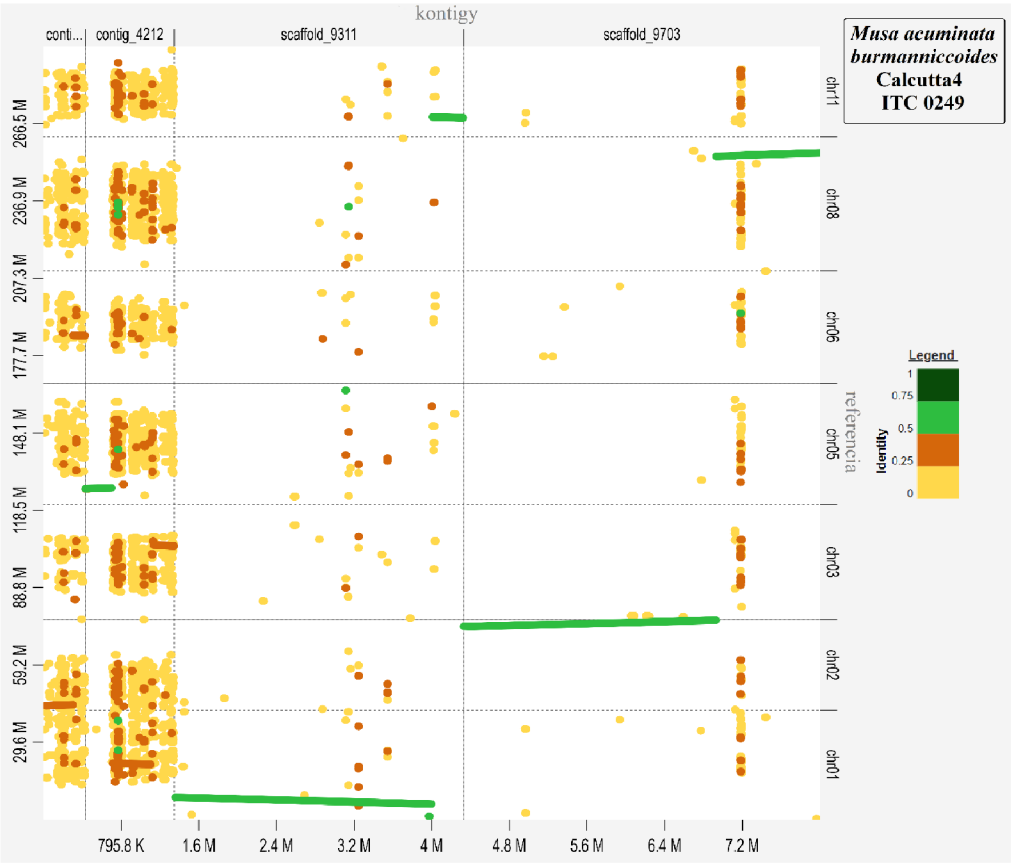
Translokácie boli vizualizované pomocou webovej stránky D-Genies. Ako vstup si berie FASTA súbor s kontigmi, ktoré boli výsledkom opísanej cesty a FASTA referenčného genómu (Musa acuminata DH-Pahang V4). Pre zlepšenie prehľadnosti boli v niektorých prípadoch z genómu vybrané chromozómy, na ktorých sa lokalizovala pravdepodobná translokácia alebo sa použil celý referenčný genóm s 11 chromozómami. Cesta sa skúšala na dvoch poddruhoch zo sekcie *Eumusa*: Calcutta 4 (ITC 0249) a Singapuri (ITC 0248). Nasledujúce strany zobrazujú výsledky pre vybrané druhy z analyzovaných zástupcov vo webovej aplikácií D-Genies. Výsledky pre všetky skúmané druhy sú dostupné v prílohách k diplomovej práci, v podobe Dot plot-u pre všetky skúmané kultivary (Príloha 7).



Obr. 17: Vizualizácia translokácie identifikovanej u *Musa balbisiana* Singapuri (ITC 0248) s D-Genies

U *M. balbisiana* sa podľa dostupných informácií (Baurens *et al.*, 2019) nachádza jedna veľká reciproká translokácia medzi chromozómom 1 a 3. Z tohto druhu (skupiny) sa skúmal kultivar Singapuri a podarilo sa nájsť kontig s číslom 2274, ktorý tejto translokácií odpovedá (Obr. 17). Úseky mali veľmi dobrú kvalitu a presahovali dĺžky 100 kb.

V článku z roku 2020 Šimoníková *et al.* uvádza pre poddruh Calcutta 4 (ITC 0249) dve veľké translokácie. Translokácie medzi chromozómom 1 a 9 a medzi chromozómom 2 a 8. Boli identifikované 2 scaffoldy a 2 kontigy (Obr. 19), z čoho práve 2 scaffoldy vizualizáciou zodpovedali translokáciam. Potvrdila sa výmena úsekov medzi chromozómami 2 a 8 ale nepodarilo sa identifikovať veľkú translokáciu medzi chromozómami 1 a 9. To mohlo nastať vďaka veľkom rozdiele vo veľkosti genómu referencie (500 Mb) a skúmaného poddruhu (600 Mb).



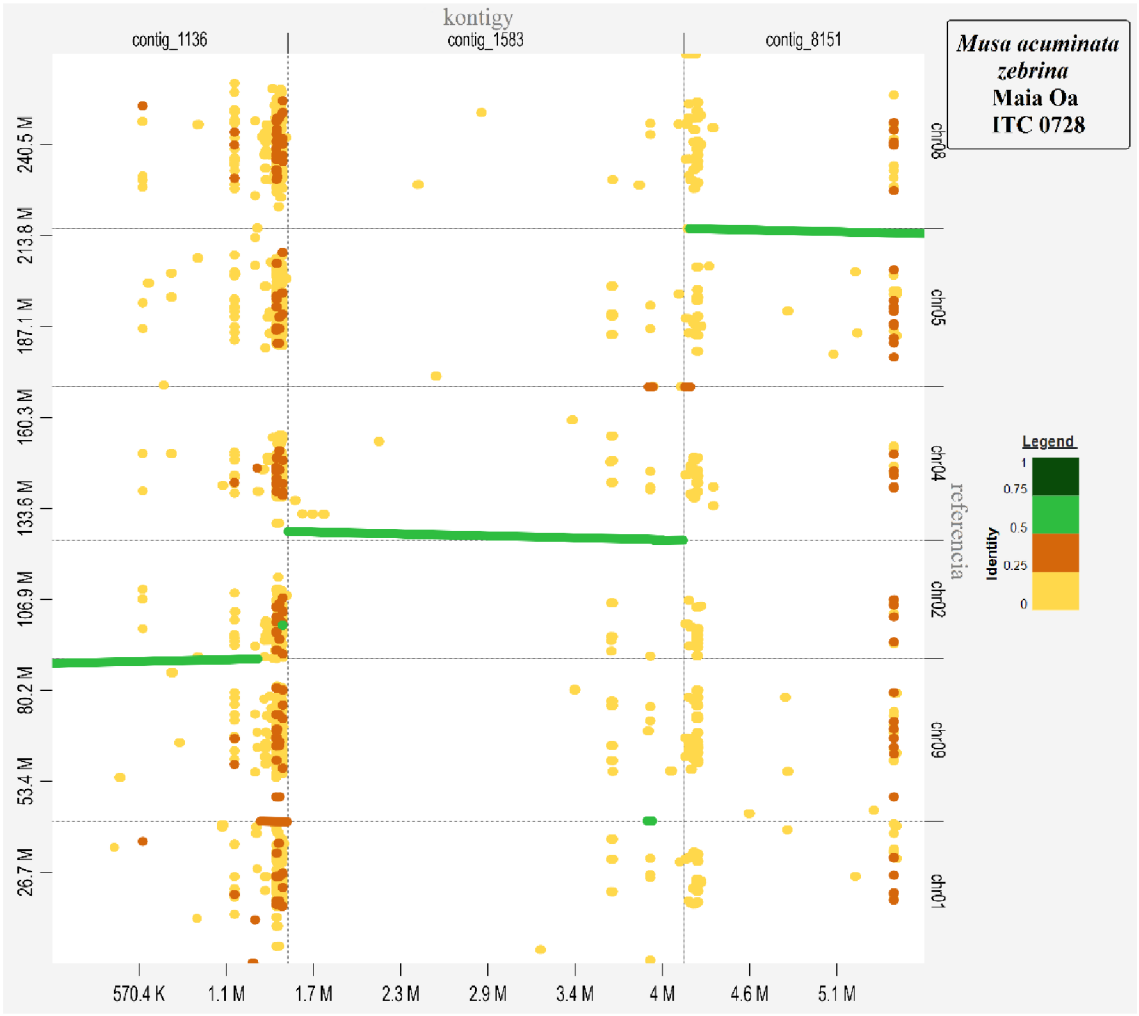
Obr. 18: Vizualizácia identifikovaných úsekoch v *M. acuminata burmanniccoides* Calcutta 4 (ITC 0249) pomocou D-Genies

Podarilo sa však identifikovať translokáciu chromozómu 1 s chromozómom 11 a vizualizáciou aj s chromozómom 5 (Obr. 18). Keďže táto prestavba nebola dosiaľ v genóme *M. acuminata* Calcutta 4 objavená ani s farbením chromozómov (Šimoníková *et al.*, 2020) ani s využitím mate-pairs Illumina sekvenačných dát, je otázkou, či vysoko chybové ONT čítania neviedli k zostaveniu hybridného scaffoldu.

Ďalej sa dentifikovali translokácie medzi chromozómom 2 a 6 (contig_169) a chromozómami 1, 3 a 5 (contig_4212). Tieto dve prestavby nevykazujú veľkú identitu.

Pre *M. acuminata* spp. *zebrina* kultivar Maia Oa je poddruhovo špecifická recipročný centromerická translokácia medzi chromozómami 3 a 8 a bola identifikovaná aj Šimoníková *et al.* (2020). V tejto práci sa však nepodarilo ju vizualizovať. Zaujímavé je, že nebola identifikovaná ani Dupouy *et al.* (2019) s použitím sekvenovania párových knižníc pomocou technológie Illumina (prístupom mate-pairs).

Identifikovaná bola translokácia medzi chromozómami 9, 2 a 1 (kontig_1136; Obr. 19), a v kontig_1583 bola s pomocou vytvorenej pipeliny identifikovaná krátka translokácia medzi chromozómami 4/ 2 a 8/5 (Obr. 19).



Obr. 19: Vizualizácia identifikovaných translokácií pre *M. acuminata* spp. *zebrina* s D-Genies

4.2.2 Polyploidy zo sekcie *Eumusa*

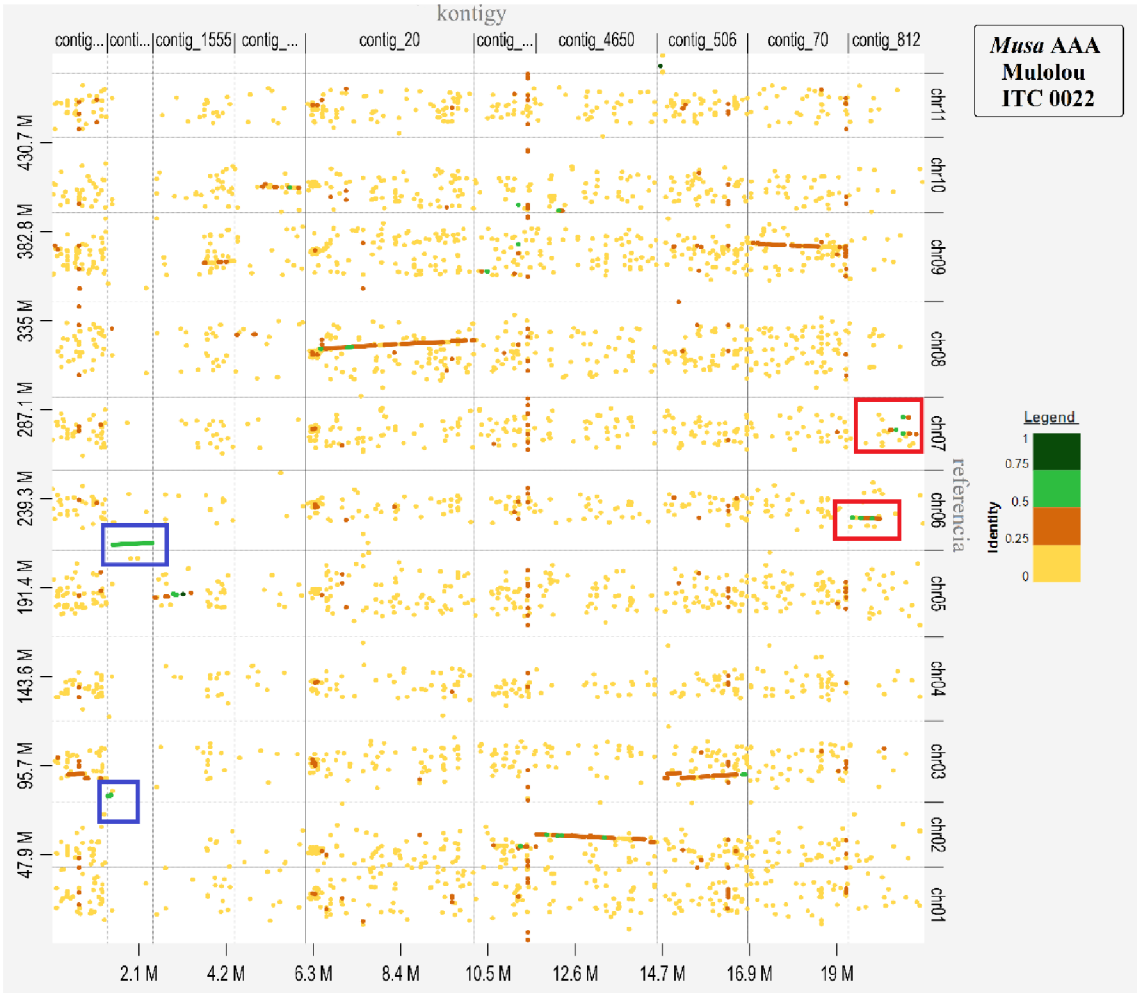
V práci sa skúmali niekoľké kultivary zo sekcie Eumusa. U ôsmych z nich je potvrdená polyploidia, počet ich chromozómov v somatických bunkách nie je teda 22 ale 33. Sú to: Mululou (ITC 0022), Nyamwihogora (ITC 0086), Nshika (ITC 0145), Intama (ITC 0153), Red Dacca (ITC 0575), Hartón Tigre (ITC 0642), 3 Hands Planty (ITC 1132) a Poyo (ITC 1482). Všetci títo zástupcovia boli zároveň analyzovaní tiež pomocou oligo paintig FISH, kolegami v Centre štruktúrnej a funkčnej genomiky rastlín, ÚEB AB ČR v Olomouci (Beránková & Hřibová, nepublikované).

U zástupcov banánovníkov africkej vysočiny (Nyamwihogora, Nshika a Intama) oligo painting FISH potvrdil prítomnosť centromerickej recipročnej translokácie medzi chromozómy 3 a 8. Túto recipročnú centromerickú translokáciu 3/8 sa nám nepodarilo presne lokalizovať, podobne ako u druhu *M. acuminata* spp. *zebrina*. Vizualizácia pre *M.* Nshika obsahuje kontig_889, ktorý primárne ukazuje slabšiu translokáciu medzi chromozómami 3 a 4 (Obr. 20). Ďalej sa ukázali možné translokácie 1/5, 2/10, 7/9 a medzi

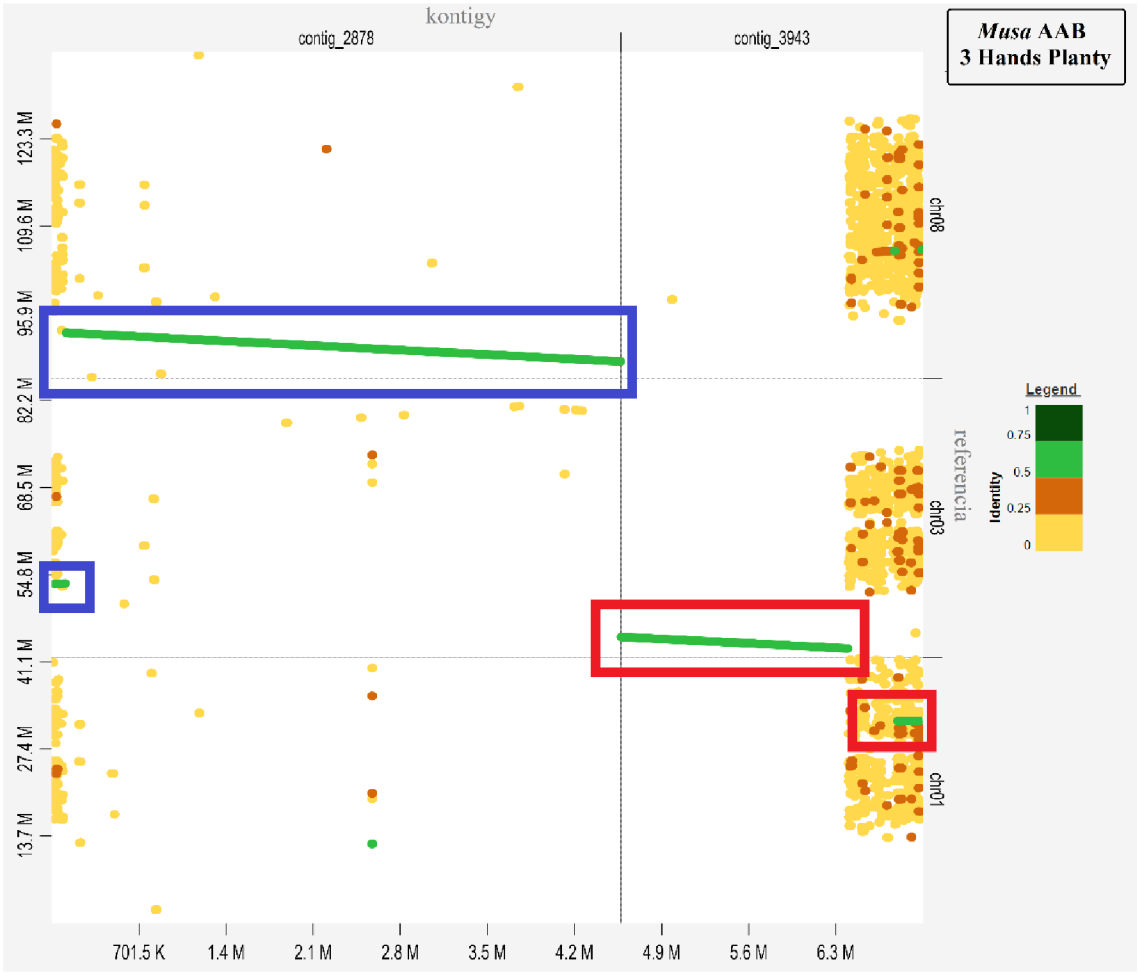
Obr. 20: Vizualizácia translokácií *M.* Nshika (ITC 0145) pomocou D-Genies: 3/4 červenou, 1/5 modrou, 2/10 zelenou, 7/9 ružovou, 7/8/9 fialovou

chromozómami 7, 8 a 9. Vykazujú však priemernú kvalitu a overenie ich pravdivosti vyžaduje hlbšie skúmanie.

U zástupcov plantajnov s genómovým zložením AAB (Mululou, Hartón Tigre a 3 Hands Planty) oligo painting FISH potvrdila prítomnosť translokácie medzi chromozómami 1 a 3, ktorá je špecifická pre druh *M. balbisiana* (B genóm). Skúmanie druhu Mulou ukázalo viacero kontigov, na ktorých sa môže nachádza translokácia (Obr. 21). Našla sa možná translokácia medzi chromozómami 6/7, 6/3. Nepodarilo sa identifikovať práve B- genómovo špecifickú translokáciu medzi chromozómami 1 a 3.



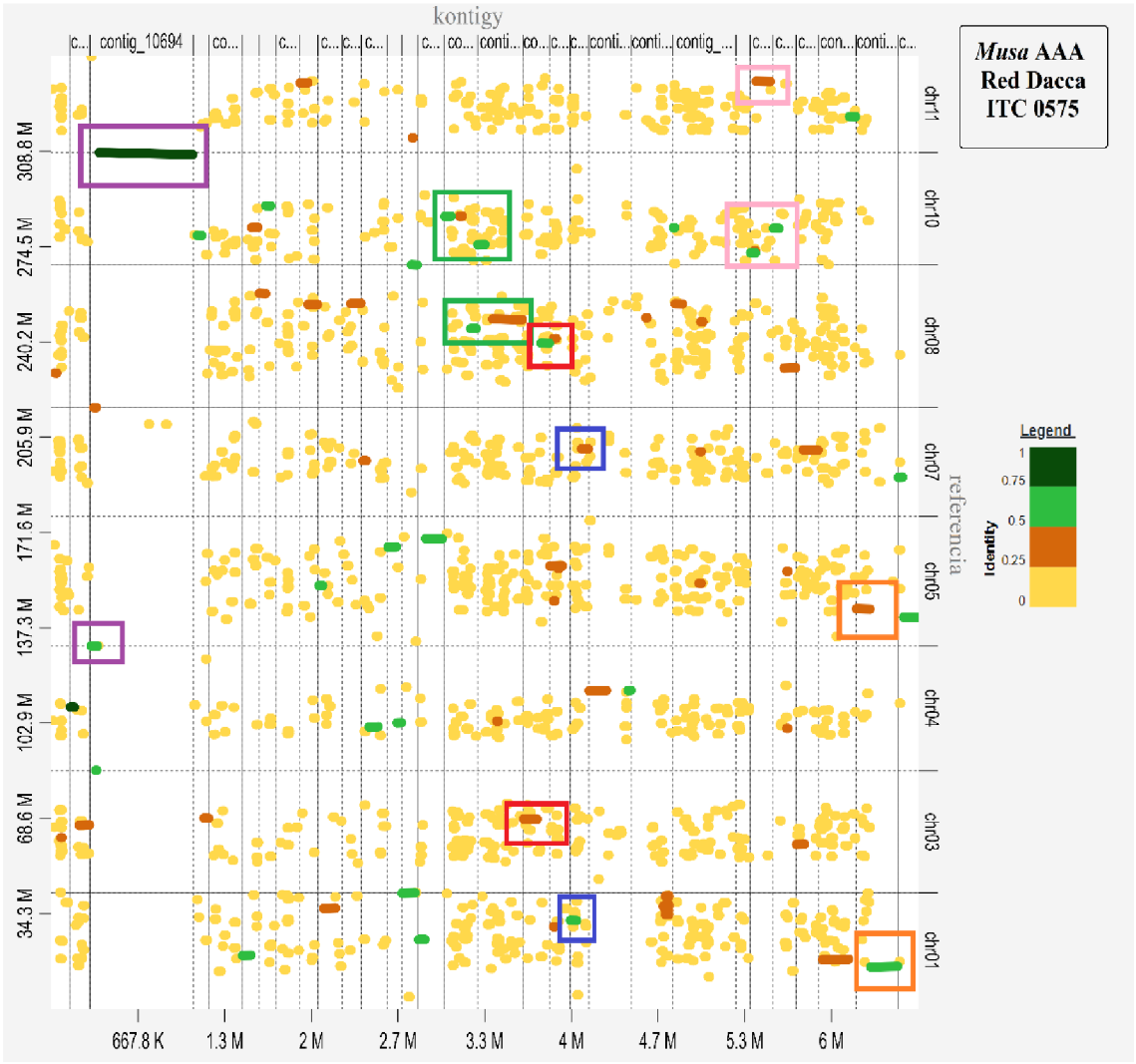
Obr. 21: Vizualizácia identifikovaných úsekoch *Musa* Mululou (ITC 0022) pomocou D-Genies: 6/7 červenou, 6/3 modrou



Obr. 22: : Vizualizácia identifikovaných úsekoch *Musa* AAB 3 Hands Planty (ITC 1132) pomocou D-Genies: translokácia 1/3 červenou a 3/8 modrou

Pre druh *M.* AAB 3 Hands Planty (ITC 1132) sa podarilo túto translokáciu vizualizovať(Obr. 22). Okrem toho sa našla možná prestavba medzi chromozómom 3 a 8.

Painting FISH pre druh Red Dacca (ITC 0575) ukázal, že obsahuje recipročnú centromerickú translokáciu 3/8, translokáciu medzi chromozómami 9/6 a štruktúru zahŕňajúcu translokáciu medzi 1/7. Nám sa podarila identifikácia prestavby medzi chromozómami 3 a 8, s identitou 25% na chromozóme 3 (Obr. 23). Ďalej sa podarila nájsť translokácia 1/7 s identitou 25% pre chromozóm 7. Možné identifikované prestavby pomocou DGenies sú 8/10; 4/5/10/11; 10/11; 5/1.

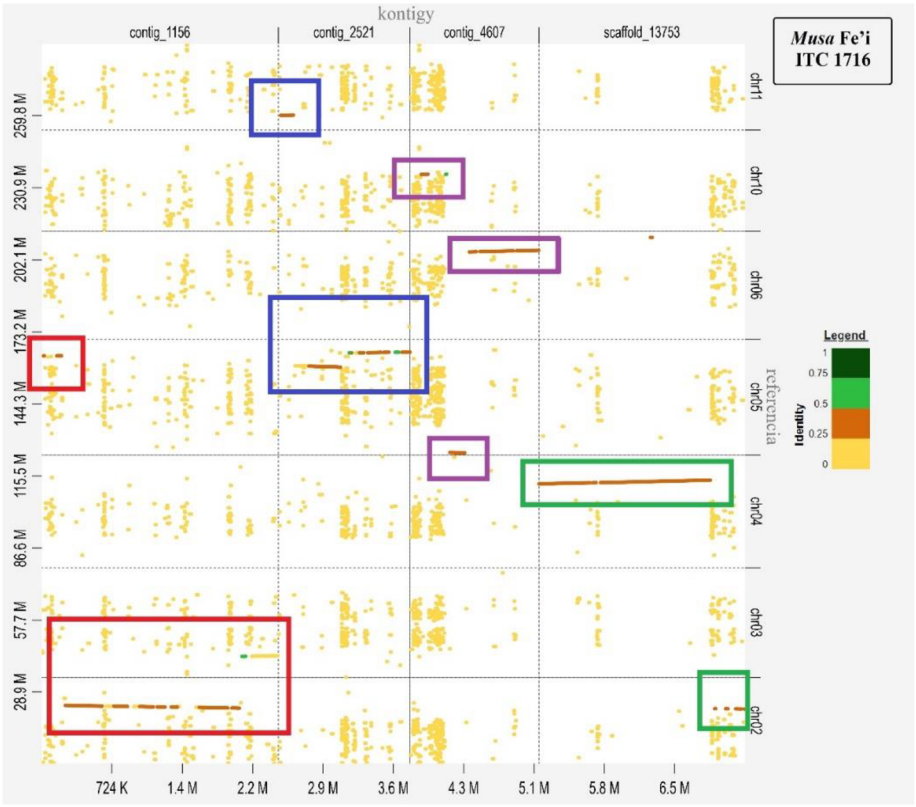


Obr. 23 Vizualizácia translokácií *M.* Red Dacca (ITC 0575) pomocou DGenies: 3/8 červenou, 1/7 modrou, 8/10 zelenou, 4/5/10/11 fialovou, 10/11 ružovou a 5/1 oranžovou

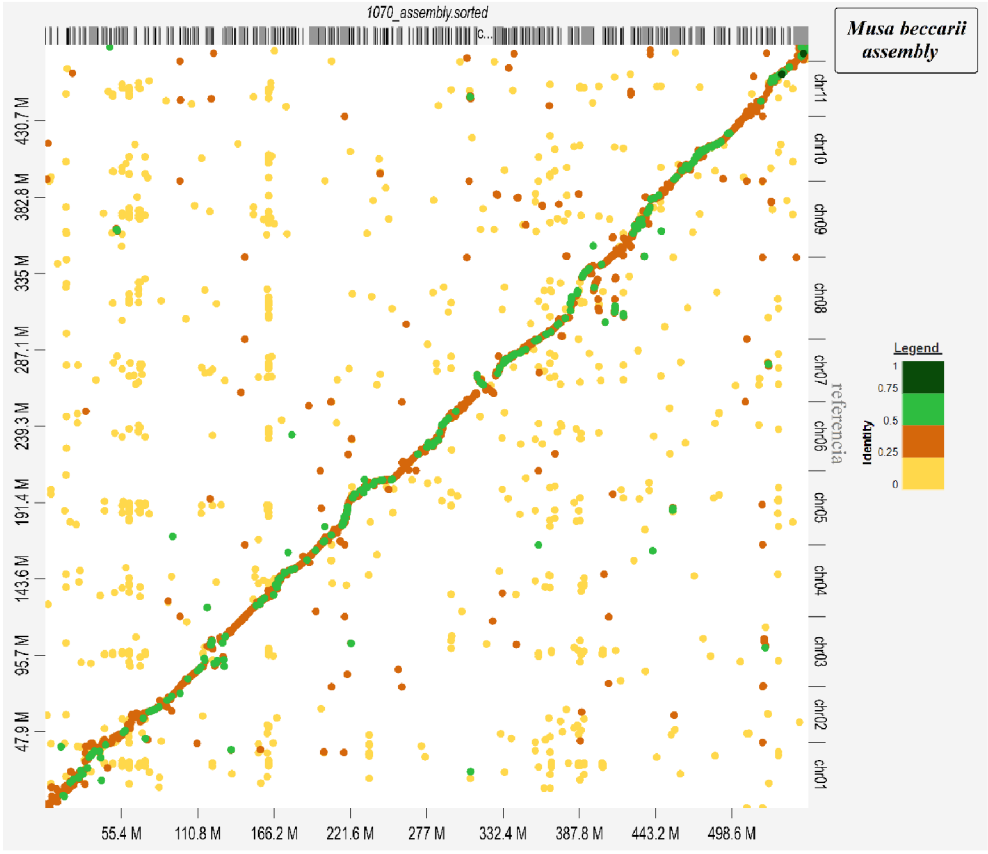
4.2.3 Plané druhy zo sekcie Australimusa

Zo sekcie Australimusa boli skúmané 3 druhy, dva diploidy *M. textilis, M. maclayi* typu Hung si a jeden jedlý triploid zástupca Fe'i banánovníkov. U triploidu Fe'i sme lokalizovali translokácie medzi chromozómami 2, 3 a 5; 11 a 5; 2 a 4; medzi chromozómami 10/6/5 (Obr. 24).

Podľa farbenia chromozómov Fe'i, banánovník obsahuje násobné translokácie medzi chromozómami: 1/11; 8/9/2/4; 8/9/7; 11/7/9 (Beránková & Hřibová, nepublikované). Nám sa nepodarilo identifikovať tieto translokácie, ale identifikovali sme iné typy, kde mnohé z nich obsahovali len krátke translokované úseky niektorých chromozómov.



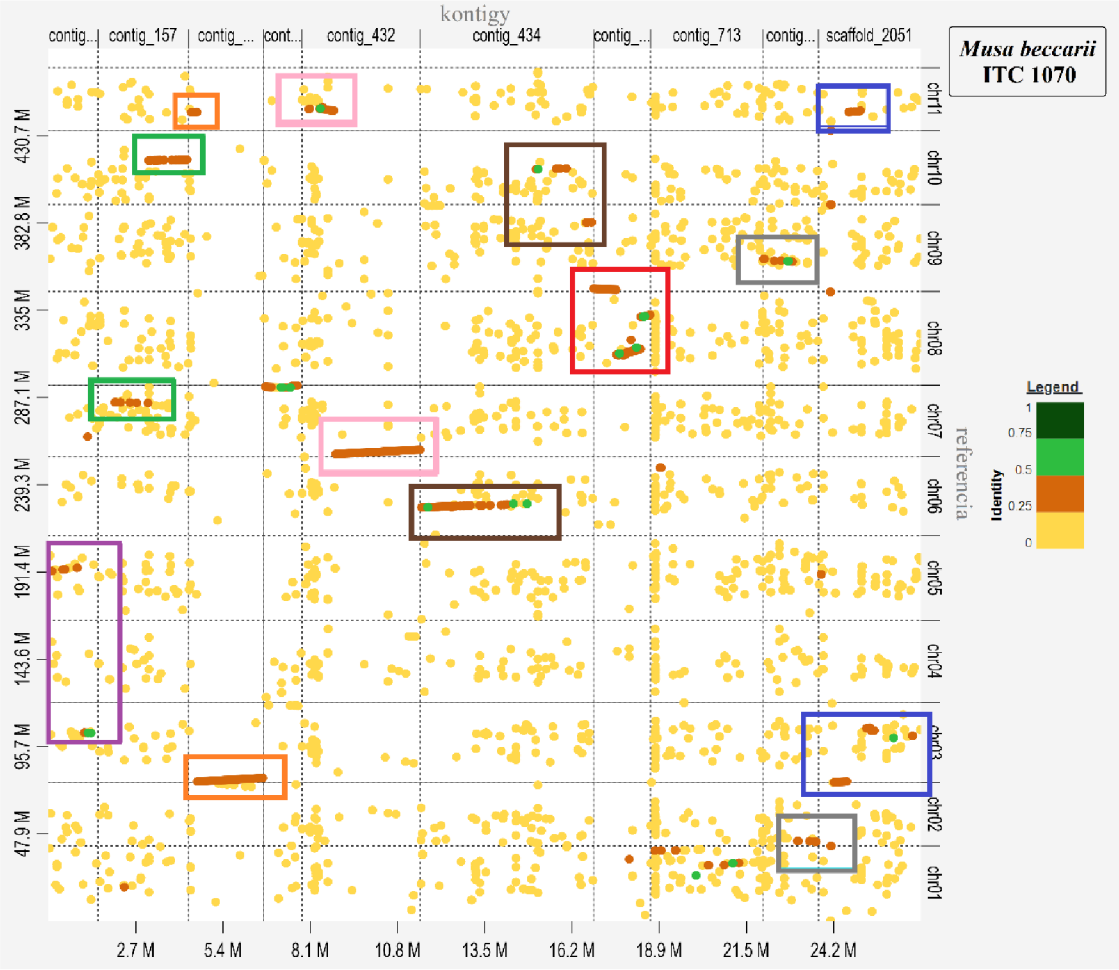
Obr. 24: Vizualizácia identifikovaných úsekoch *Musa* Fe´i (ITC 1716) pomocou D-Genies: 2/3/5 červenou, 11/5 modrou, 2/4 zelenou a 10/6/5 fialovou



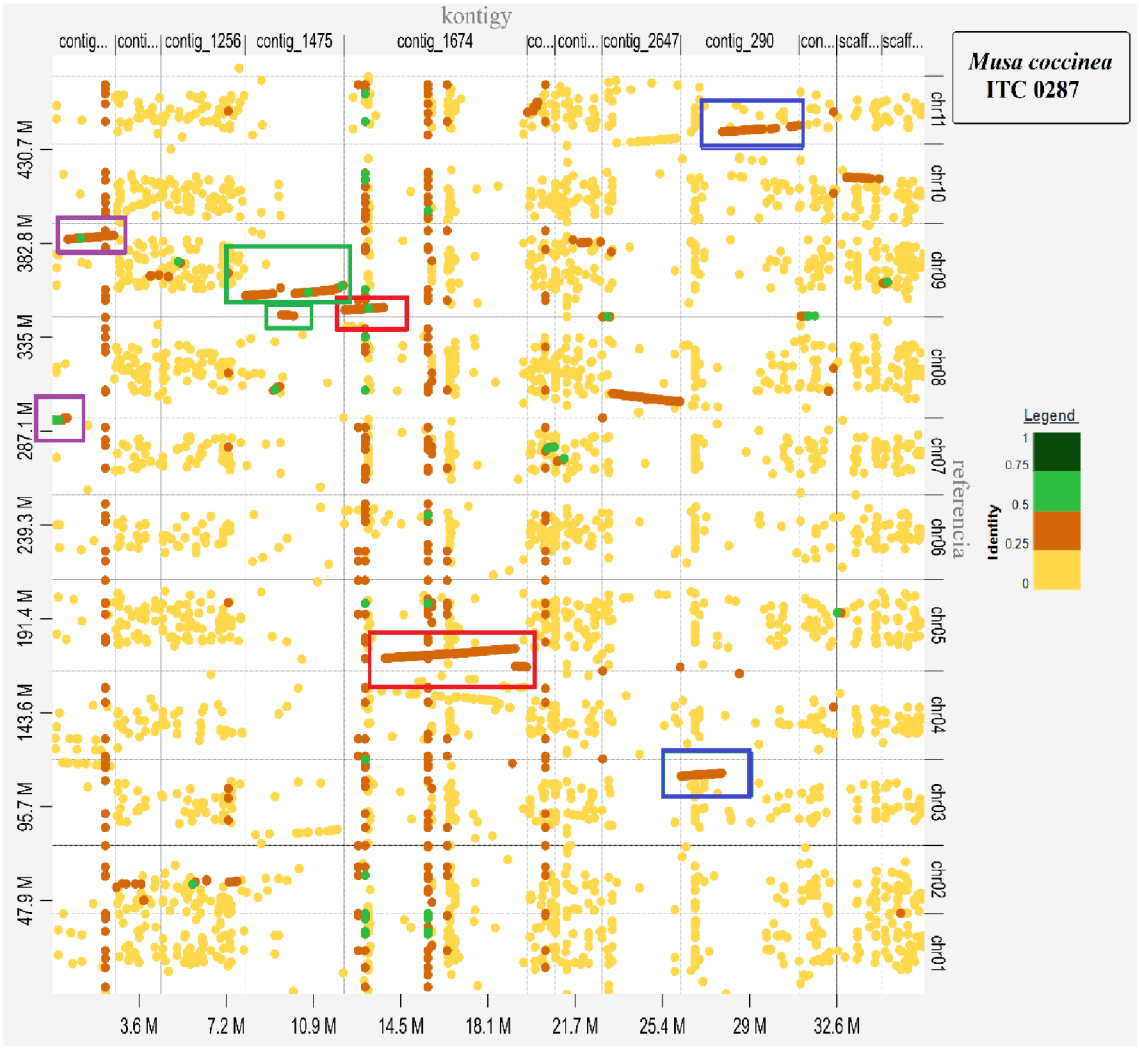
Obr. 25: Porovnanie *M. beccarii* (2n = 18) oproti *M. acuminata* DH Pahang (2n = 22) v D-Genies

4.2.4 Plané druhy zo sekcie Callimusa

Skúmané druhy zo sekcie *Callimusa* majú najmenší počet chromozómov. *Musa beccarii* má základný počet chromozómov 2n = 18. Wang *et al.* (2023) publikoval celkogenómovú sekvenciu druhu *M. beccarii,* z ktorej je zrejmé, že genóm tohto druhu je voči referenčnej sekvencii *M. acuminata* DH Pahang značne prestavený (Obr. 25). V našej štúdii, kde pre identifikáciu dlhých chromozomálnych prestavieb bola využitá čiastočná assembly zostavená z ONT sekvenačných dát, sa podarilo identifikovať tieto chromozomálne prestavby: 5/3; 10/7; 11/3; 11/7; 10/9/6; 9/8; 9/2; 11/3 (Obr. 26).



Obr. 26 Vizualizácia identifikovaných úsekoch *Musa beccarii* (ITC 1070) pomocou D-Genies: 5/3 fialovou, 10/7 zelenou, 11/3 oranžovou, 11/7 ružovou, 10/9/6 hnedou, 9/8 červenou, 9/2 šedou a 11/3 modrou



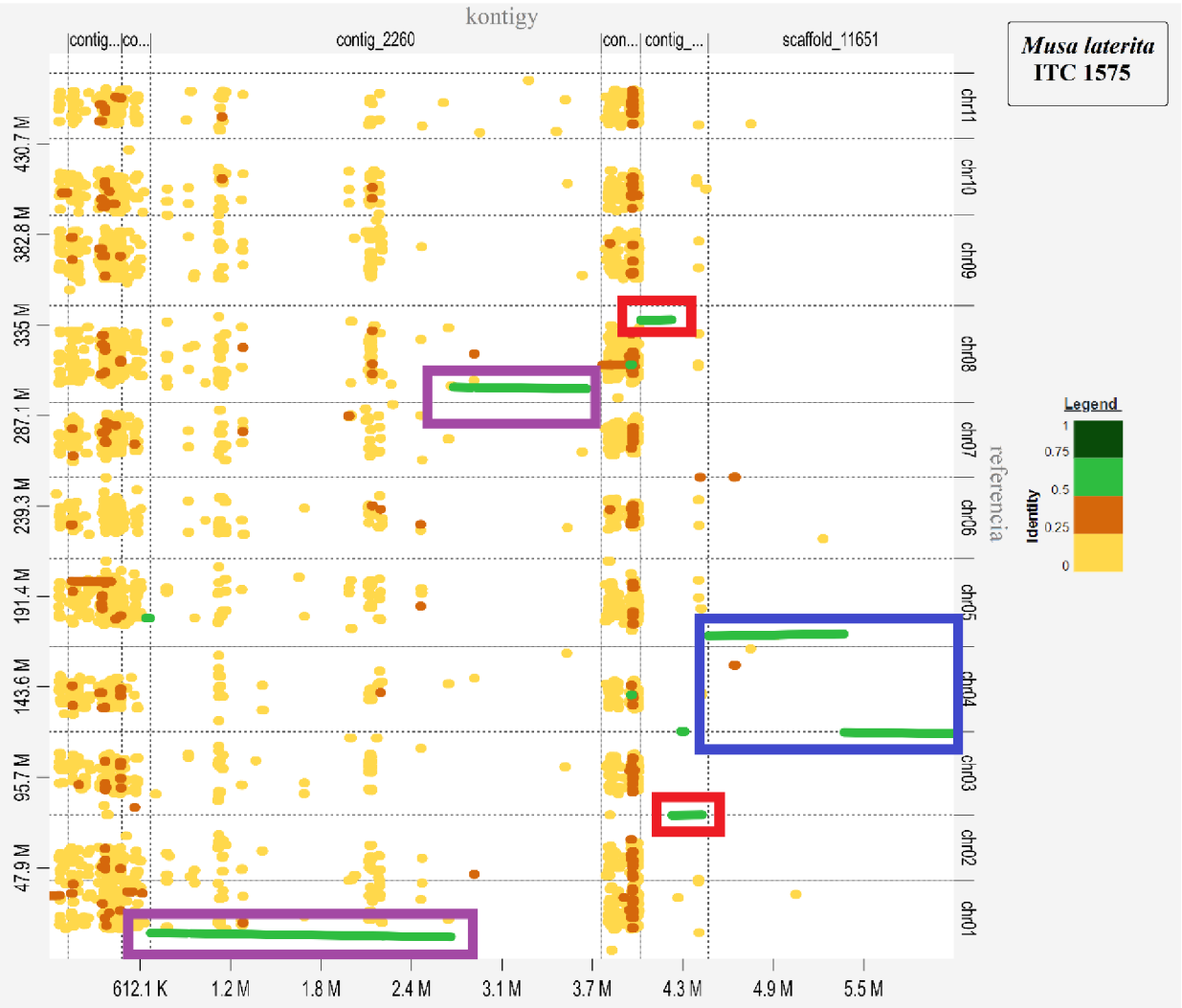
Obr. 27: Vizualizácia identifikovaných úsekoch *Musa coccinea* (ITC 0287) pomocou D-Genies: 5/9 červenou, 7/8/9 fialovou, 3/11 modrou a 9/8 zelenou

Z chromozomálnych prestavieb, ktoré boli nájdené pomocou farbenia chromozómov a ktoré odpovedajú translokáciám objaveným celogenómovej sekvencii *M. beccarii* (Wang *et al.*, 2023), sa nám podarilo nájsť tieto translokácie: 11/3; 9/8; 6/1; 11/2/3; 9/2 a 11/3.

Druh *M. coccinea* zatiaľ nemá žiadne publikované translokácie. V práci sme identifikovali tieto možné translokácie: 5/9, 7/8/9, 3/11, 9/8 a inzerciu krátkeho úseku chromozómu 8 do chromozómu 9 (Obr. 27). Žiadna z týchto chromozomálnych prestavieb neodpovedá translokáciam identifikovaným pomocou farbenia chromozómov (Beránková & Hřibová, nepublikované).

4.2.4 Druhy sekcie Rhodochlamys

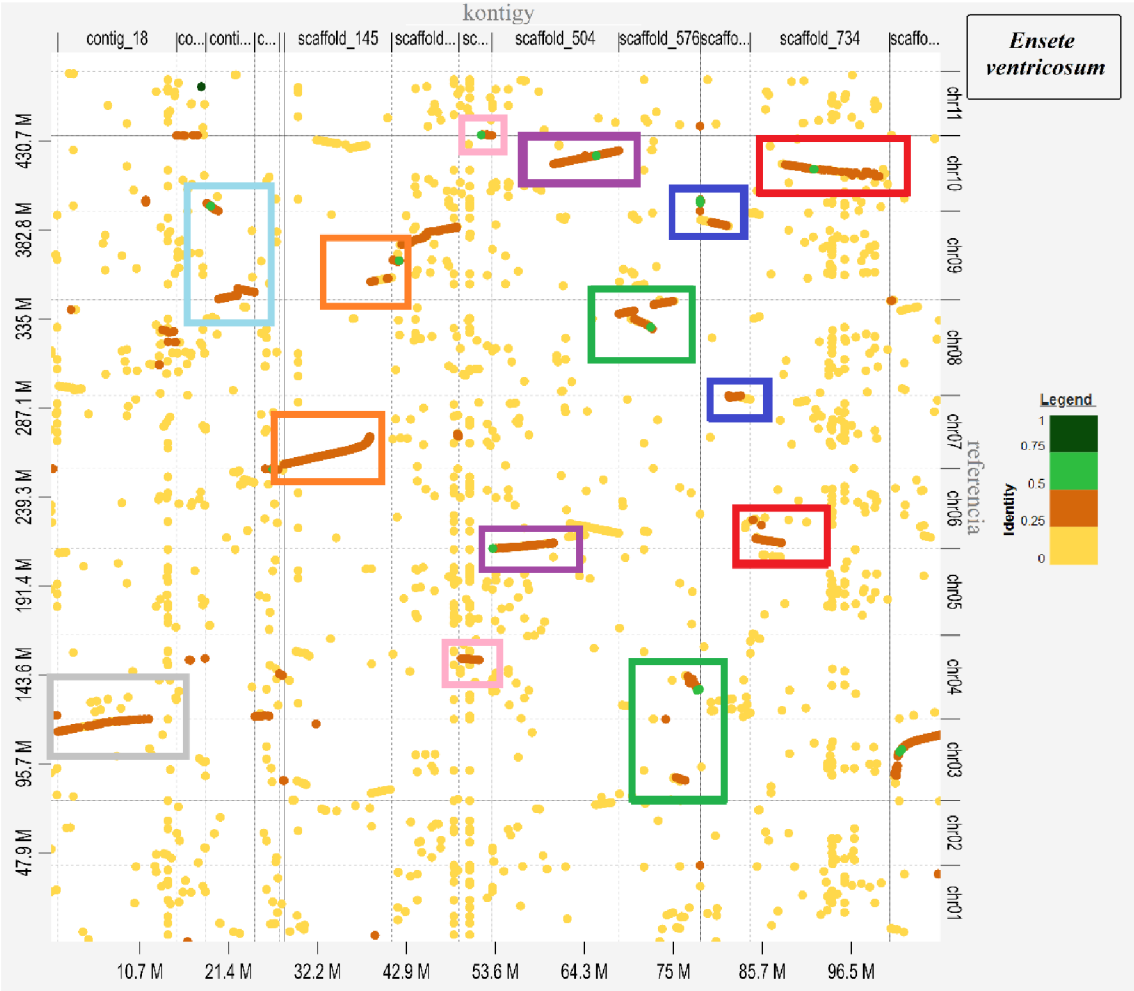
Rhodochlamys obsahuje niekoľko okrasných jedincov. V práci sa hľadali translokácie na 2 druhoch : *M. ornata* a *M. laterita*. Pre *M. laterita* (ITC 1575) boli s farbením chromozómov objavené chromozomálne prestavby 3/1; 2/8; 5/3/1 a 8/7,(Beránková & Hřibová, nepublikované). Nám sa podarilo zobraziť translokáciu medzi chromozómami 2/8/3 a pomerne veľké prestavby medzi chromozómami 5/3/4; 8 a 1 (Obr. 28), ktoré sú s najväčšou pravdepodobnosťou výsledko hybridného scaffoldingu pri zostavované čiastočnej assembly s využitím vysoko chybových ONT čítaní.



Obr. 28: Vizualizácia identifikovaných úsekoch *Musa laterita* (ITC 1575) pomocou D-Genies: 2/8/3 červenou, 5/3/4 modrou a 1/8 fialovou

4.2.5 Druh z rodu *Ensete*

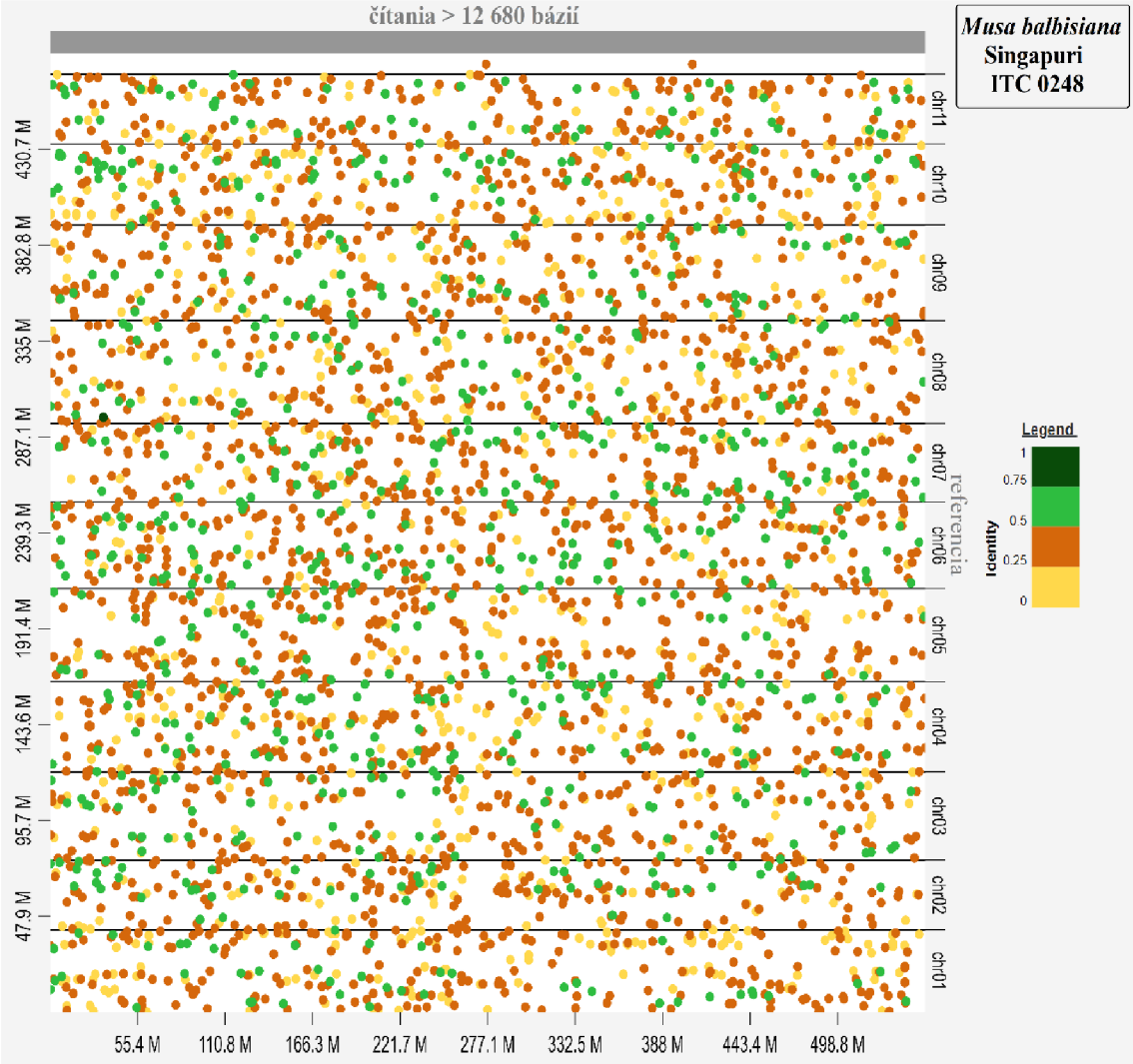
Ensete patria spolu s banánovníkmi *Musa* do čeľade *Musaceae*. Podobne ako *M. beccarii* menší počet chromozómov – 2n = 18. To znova možno vidieť vo výslednom obrázku z lokalizácie translokácií. Našli sme možné prestavby medzi chromozómami: 10/6; 10/9/8/7; 4/3/8; 10/6/5; 11/4; 7/9; 9/10 a 4/3 (Obr. 29). Translokácie medzi chromozómami 10 a 6; 4/3/8 a 11/4 sú potvrdené, ako zo zrovnania celkových referenčných genómov druhov *M. acuminata* DH Pahang a *E. glaucum* (Wang *et al.*, 2022), tak z farbenia chromozómov druhu *E. ventricosum* (Beránková & Hřibová, nepublikované).



Obr. 29: Vizualizácia identifikovaných úsekoch *Ensete ventricosum* (ITC 1387) pomocou D-Genies: 4/3 šedou, 9/10 slabo modrou, 7/9 oranžovou, 11/4 ružovou, 5/6/10 fialovou, 8/4/3 zelenou, 10/9/8/7 tmavo modrou a 6/10 červenou

4.3 Vizualizácia translokácií u jednotlivých ONT čítaní

Pre vizualizáciu translokácií z ONT čítaní bola najskôr testovaná rovnaká cesta, ako pre zostavené kontigy. Avšak výsledky neboli dostačujúce a ich vizualizácia nebola prehľadná (Obr. 30), preto sa čítania namapovali na skúmané kontigy, ktoré vznikli z cesty zostavujúcej assembly, vizualizovanej vyššie. Pre zobrazenie výsledkov bola použitá aplikácia IGV, kde sa zobrazili identifikované body zlomu podľa tabuľky zostavenej pomocou skriptu v programovacom jazyku python.

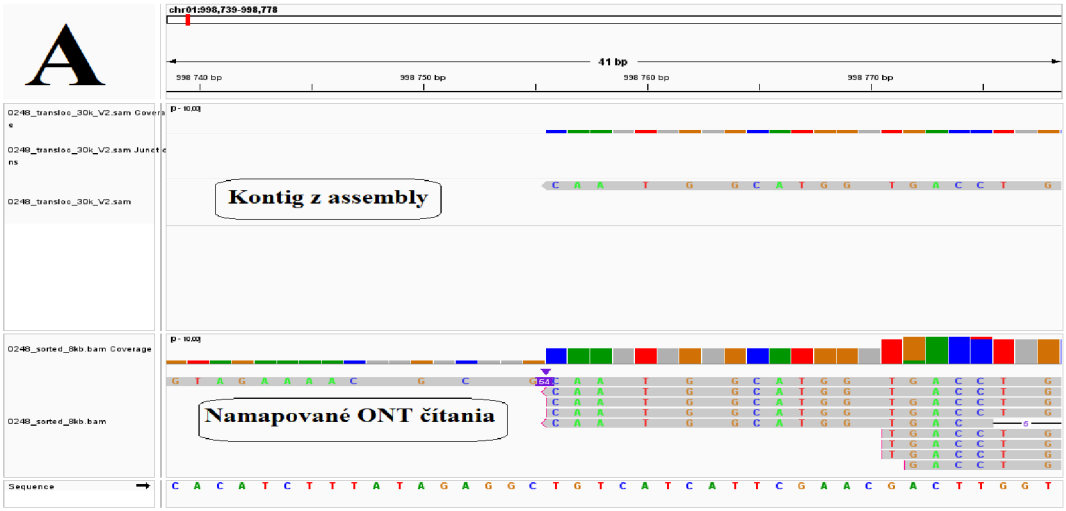


Obr. 30: Pokus o vytvorenie Dot plotu so vstupnými dátami ONT čítaní, upravených podľa vytvorenej pipeline pre assembly, pre *M. balbisiana* Singapuri (ITC 0248)

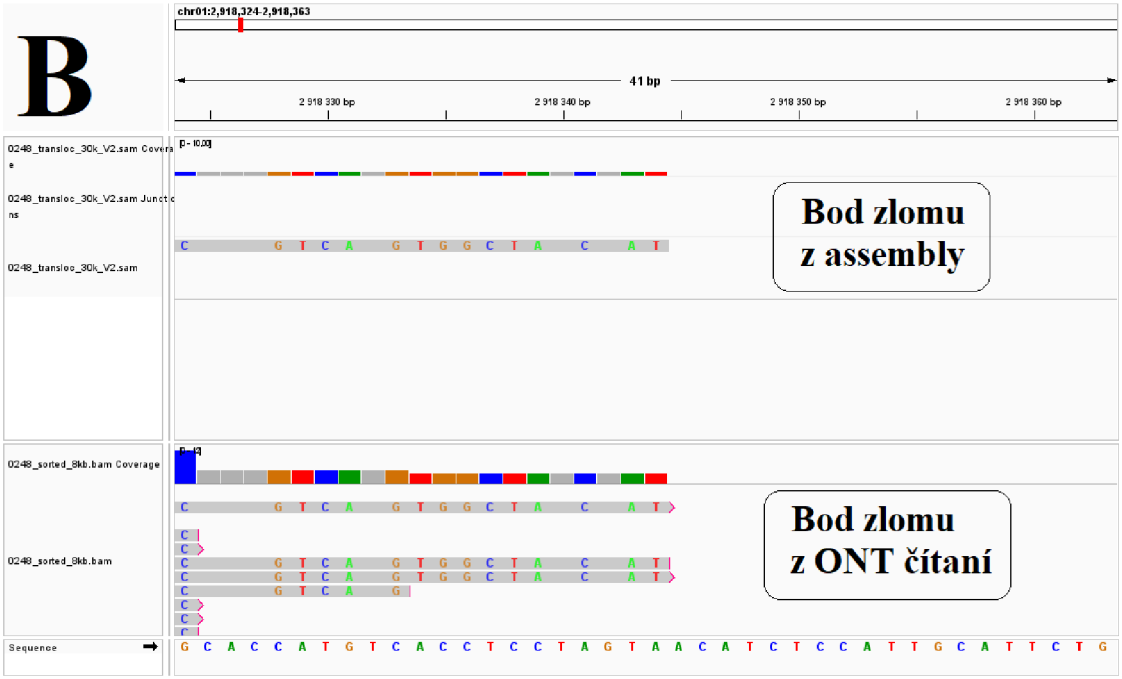
Potvrdila sa identifikácia translokácie 1/3 u *M. balbisiana* Singapuri (Tab. 3) s čítaniami ONT. Začiatočná báza na chromozóme 1 je 998 756 (Obr. 31) a bod zlomu 2 973 621 (Obr. 32).

Tab. 3: Tabuľka možných bodov zlomu, začiatočných pozícií, dĺžky úsekov a identity pre identifikované translokácie pre *M. balbisiana* Singapuri (ITC 0248), zvýraznená je začiatočná pozícia na chromozóme 1 a bod zlomu

Názov	Chro-mozóm	Začiatočná pozícia	Breakpoint	Dĺžka	Identita
contig_2274	chr01	998 756	1 364 195	365 440	0,586
contig_2274	chr01	2 480 884	2 718 084	237 201	0,809
contig_2274	chr01	2 718 718	**2 918 344**	199 627	0,749
contig_2274	chr01	**1 959 796**	2 136 486	176 691	0,825
contig_2274	chr01	1 561 659	1 697 489	135 831	0,759
contig_2274	chr01	1 368 818	1 481 586	112 769	0,775
contig_2274	chr01	2 138 442	2 235 992	97 551	0,794
contig_2274	chr01	902 506	998 643	96 138	0,716
contig_2274	chr01	2 923 576	2 973 621	50 046	0,864
contig_2274	chr03	35 752 612	36 097 409	344 798	0,727
contig_2274	chr03	36 138 014	36 393 486	255 473	0,716
contig_2274	chr03	36 138 014	36 393 486	255 473	0,716
contig_2274	chr03	37 140 888	37 310 759	169 872	0,805
contig_2274	chr03	35 507 824	35 642 559	134 736	0,722
contig_2274	chr03	36 405 157	36 502 427	97 271	0,864
contig_2274	chr03	36 504 616	36 679 668	175 053	0,434
contig_2274	chr03	35 299 508	35 390 449	90 942	0,780
contig_2274	chr03	35 684 372	35 749 528	65 157	0,799



Obr. 31: Identifikácia začiatočnej pozície translokácie 1/3 *M. balbisiana* Singapuri (ITC 0248) na chromozóme 1



Obr. 32: Identifikácia bodu zlomu translokácie 1/3 *M. balbisiana* Singapuri (ITC 0248) na chromozóme 1 s ONT čítaniami, zobrazená v IGV

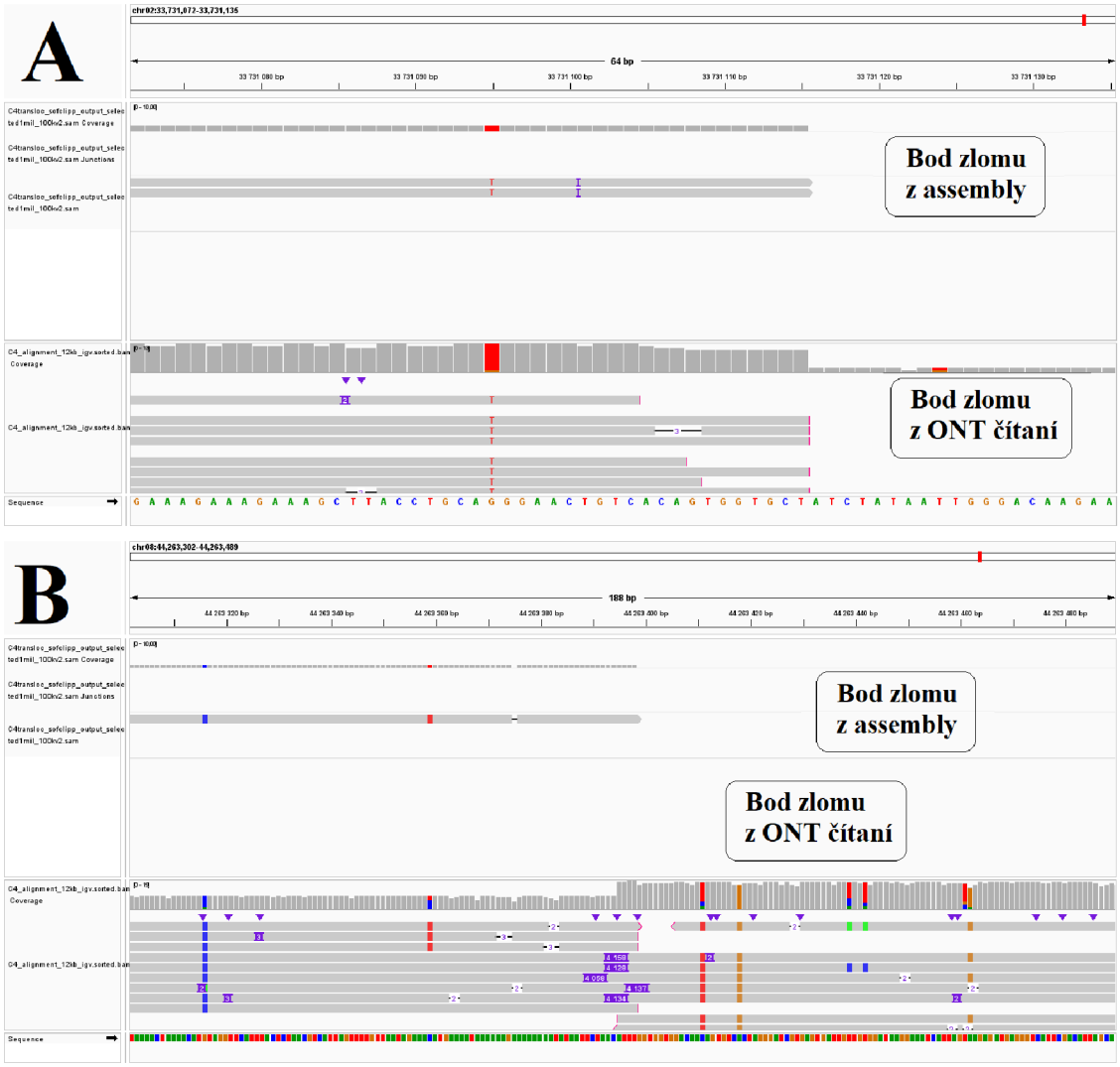
Pre druh *M. acuminata burmanniccoides* Calcutta 4 bola úspešne potvrdená translokácia 2/8 pomocou IGV vizualizácie namapovaných čítaní na kontigy (Obr. 33). Pre translokáciu 1/11 sa potvrdila možná translokácia 1/11 nájdeným začiatkom na chromozóme 1 a bodom zlomu na chromozóme 11. Bod zlomu na chromozóme 1 sa nenašiel na predpokladanom mieste (Tab. 4).

Tab. 4: Tabuľka možných bodov zlomu, začiatočných pozícií, dĺžky úsekov a identity pre identifikované translokácie pre *M. acuminata burmanniccoide* Calcutta 4 (ITC 0249)

Kontig	Chro-mozóm	Začiatočná pozícia	Bod zlomu	Dĺžka	Identita
contig_169	chr02	1 946 453	2 066 857	120 405	0,576
contig_169	chr06	18 324 112	18 419 413	95 302	0,360
contig_4212	chr01	21 386 719	21 514 172	127 454	0,474
contig_4212	chr01	21 190 929	21 326 289	135 361	0,387
contig_4212	chr03	28 369 634	28 518 700	149 067	0,481
contig_4212	chr05	6 297 094	6 472 507	175 414	0,891
scaffold_9311	chr01	7 682 374	7 901 302	218 929	0,862
scaffold_9311	chr01	7 491 627	7 677 084	185 458	0,898
scaffold_9311	chr11	7 377 117	7 581 506	204 390	0,807
scaffold_9311	chr11	7 377 117	7 581 506	204 390	0,807
scaffold_9703	chr02	32 133 044	32 869 108	736 065	0,876

Tab. 4: Pokračovanie tabuľky možných bodov zlomu, začiatočných pozícií, dĺžky úsekov a identity pre identifikované translokácie pre *M. acuminata burmanniccoide* Calcutta 4 (ITC 0249)

Kontig	Chro-mozóm	Začiatočná pozícia	Bod zlomu	Dĺžka	Identita
scaffold_9703	chr02	33 362 626	33 573 610	210 985	0,761
scaffold_9703	chr02	33 573 604	33 731 115	157 512	0,822
scaffold_9703	chr02	33 573 604	33 731 115	157 512	0,822
scaffold_9703	chr08	44 263 395	45 045 148	781 754	0,908
scaffold_9703	chr08	43 861 173	44 051 073	189 901	0,842
scaffold_9703	chr08	43 861 173	44 051 073	189 901	0,842
scaffold_9703	chr08	44 149 979	44 263 398	113 420	0,505



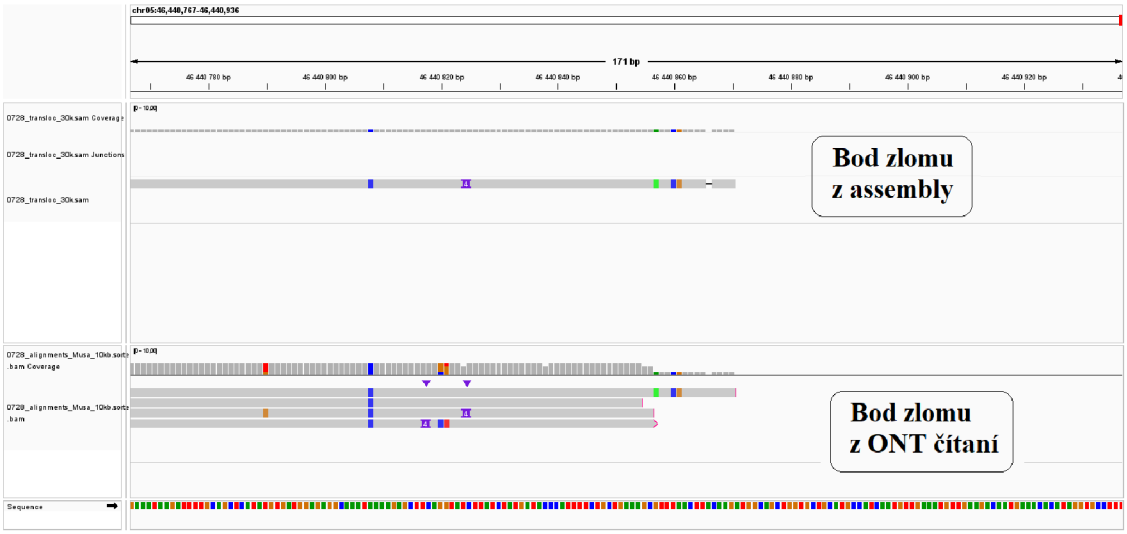
Obr. 33: Vizualizácia bodu zlomu translokácie 2/8 na chromozóme 2 (A) a 8 (translokácia 2/8) pre *M. acuminata burmanniccoides* Calcutta 4 s ONT čítaniami v IGV

Tab. 5: Tabuľka možných bodov zlomu, začiatočných pozícií, dĺžky úsekov a identity pre identifikované translokácie pre *M. acuminata zebrina* Maia Oa (ITC 0728)

Názov	Chro-mozóm	Začiatočná pozícia	Breakpoint	Dĺžka	Identita
contig_1136	chr01	41 635 119	41 740 330	105 212	0,167
contig_1136	chr01	41 635 119	41 740 330	105 212	0,167
contig_1136	chr01	41 603 796	41 633 889	30 094	0,464
contig_1136	chr09	47 481 316	47 580 682	99 367	0,869
contig_1136	chr09	46 923 041	46 987 862	64 822	0,862
contig_1136	chr09	47 445 511	47 480 973	35 463	0,225
contig_1583	chr02	34 619 128	34 791 027	171 900	0,706
contig_1583	chr02	34 619 128	34 791 027	171 900	0,706
contig_1583	chr04	652 864	821 330	168 467	0,910
contig_1583	chr04	518 269	**650 259**	131 991	0,890
contig_8151	chr05	45 030 707	45 057 154	26 448	0,728
contig_8151	chr05	46 431 806	**46 440 870**	9 065	-1,719
contig_8151	chr08	51 191 506	**51 209 159**	17 654	-0,890

Pre druh *M. acuminata zebrina* Maia Oa (ITC 0728) sme lokalizovali body zlomu (Tab. 5) pre možnú translokáciu 5/8. Podarilo sa nájsť aj zodpovedajúce čítania, ktoré následne boli vizualizované úspešne v IGV (Obr. 34).

Výsledky pre ostatné skúmané kultivary sa nachádzajú v zložke IGV v prílohe 6. Zložky obsahujú obrázky stiahnuté z IGV, nie celé súbory BAM, kvôli ich veľkosti (niekoľko GB). Ďalej obsahujú tabuľky s možnými bodmi zlomu. Pre ilustráciu bola vybraná len 1 ilustrácia pre kultivar.



Obr. 34: Bod zlomu translokácie 5/8 na chromozóme 8 u druhu *M. acuminata zebrina* Maia Oa

5 DISKUSIA

Sekvenovanie metódou Oxford Nanopore môže poskytovať veľmi dlhé čítania prekonávajúce problémy s použitím krátkych čítaní. Dokáže charakterizovať aj dlhé štruktúrne varianty a vysoko repetetívne oblasti. ONT čítania sa dajú použiť efektívne na zostavenie *de novo* genómov, kde ich dĺžka pomáha generovaniu dlhých a súvislých kontigov. Ďalšiu metódu poskytujúce dlhé čítania je PacBio sekvenovanie (Eid *et al.*, 2009). Oproti ONT sekvenovaniu, ktoré generuje sekvenicie s relatívne vysokým počtom chýb (u rastlín okolo 14% chybných bázií v jednotlivých čítaniach), najnovší prístup PacBio HiFi generuje sekvenačné čítania s vysokou kvalitou – kvalitou Illumina sekvenovania (99,9% presnosť). Nedávno prišla firma Oxford Nanopore Technologies s novou verziou sekvenačnej flow-celly, ktorá je v kombinácií s novým kitom prípravy ONT sekvenačných knižníc schopná (v prípade ľudskej DNA) generovať sekvencie s vysokou kvalitou (>94%; https://nanoporetech.com). Vzhľadom k tomu, že v našej práci sa dáta využité pre identifikáciu translokácií a ich breakpointov začali generovať pred viac než 1,5 rokom, boli využité vtedy dostupné kity pre ONT sekvenovanie. Generované ONT čítania mali tak relatívne vysokú mieru chybovosti. ONT bola zvolená pre získanie sekvencií vďaka svojej rýchlej dostupnosti, finančnej aj časovej.

V rámci diplomovej práce boli porovnané dva metodické prístupy s cieľom identifikácie translokačných miest - bodov zlomov, sprevádzajúcich evolúciu karytotypov pri čeľadi *Musaceae*. Na tento účel bola vytvorená bioinformatická pipeline, ktorá umožňuje identifikovať dlhé chromozómové prestavby (dlhšie ako 30 kb) v dostupných ONT dátach. Vytvorená pipleine bola aplikovaná ako na identifikáciu dlhých štruktúrnych prestavieb v jednotlivých ONT početných, tak v na čiastočne zostavených genómových sekvenciách z vygenerovaných ONT dát – v dlhých kontigoch a scaffoldoch. Všetky skúmané druhy boli zároveň analyzované metódou oligo painting FISH, ktorá umožňuje identifikáciu veľkých chromozómových translokácií – voči referenčnému genómu *M. acuminata* DH Pahang.

Bolo zistené, že priame mapovanie surových ONT sekvenačných čítaní na referenčný genóm *M. acuminata* DH Pahang neumožnilo jednoznačnú identifikáciu ONT čítania nesúcich bod zlomu a nie je tak možné tento priamy prístup využiť na identifikáciu dlhých chromozomálnych prestavieb, a ani pre charakterizáciu bodu zlomu u známych

štrukturálnych prestavieb (zistených napr. maľovaním chromozómov). Príčinou je s najväčšou pravdepodobnosťou vysoká chybovosť analyzovaných ONT čítania. Ako bolo spomenuté vyššie, od apríla toho roku je dostupná nová sekvenačná flow cell, ktorá v kombinácii s novou generáciou kitov pre prípravu ONT knižníc môže poskytovať ONT čítanie blížiace sa kvalite illumina sekvenovania. Vzhľadom na to, že rastlinné druhy obsahujú vysoké percento metylovaných oblastí, a navyše obsahujú veľké množstvo rôznych typov metylácií, s ktorými si ONT technológie nevedia do veľkej miery poradiť, nedá sa predpokladať, že by nová generácia ONT sekvenovania u rastlín poviedla k skokovému zlepšeniu kvality ONT čítanií. Alternatívu by tak mohla predstavovať aplikácia PacBio HiFi sekvenovania, ktoré by malo byť od apríla-mája toho roku cenovo dostupnejšia, a to vďaka vývoju nového stroja (PacBio Revio), ktorý umožňuje generovať 15 x väčší objem dát za kratšiu dobu (360 Gb HiFi sekvenačných čítanie za deň) (https://www.pacb.com/revio/).

Druhý metodický prístup bol preto založený na identifikácii dlhých štruktúrnych prestavieb v čiastočných assembly vytvorených z ONT dát. Predpokladali sme, že aj množstvo ONT dát, zodpovedajúce asi 15 x pokrytiu genómu, povedie k zostaveniu dlhých kontigov či scaffoldov, v ktorých by sa potenciálne dlhé genómové prestavby dali lepšie identifikovať. Navyše, v rámci procesu assembly z ONT dát dochádza tiež k vylepšeniu kvality sekvencií u výsledných kontigov, čo by viedlo k lepšiemu percentu namapovaných sekvencií (kontigov a scaffoldov) na referenčnej sekvencii *M. acuminata* DH Pahang.

Oba dva predpoklady sa potvrdili - výsledná kvalita sekvencií v assembly dosahovala 96% a získané kontigy boli mnohonásobne presahujúce N50 ONT čítania. Najlepšie výsledky boli dosiahnuté pri analýze diploidných druhov, ktoré boli blízko príbuzné (pochádzali z rovnakej sekcie) referenečnému genómu *M. acuminata* DH Pahang. Avšak aj tu nastal problém s identifikáciou translokácie pochádzajúcej z centromérnych oblastí (translokácia 3/8 u druhu *M. acuminata zebrina*). Táto reciproká centromerická translokácia nebola jednoznačne identifikovaná. Najpravdepodobnejšou príčinou je jednak relatívne nízke pokrytie ONT čítaní (cca 15 x) a predovšetkým lokalizácia traslokačného bodu zlomu v centromerickej oblasti chromozómov, ktoré sú najvariabilnejšími oblasťami genómov. V prípade centromérov, sú to obvykle zložité oblasti, ktoré obsahujú veľké množstvo rôznych typov repetitívnych DNA sekvencií, niekedy tandemovo repetetívneho charakteru, a je tak veľmi ťažké presne identifikovať

kolineárne sekvencie z iných druhov či poddruhov a identifikovať štruktúrne zmeny v týchto genomových oblastiach. Prítomnosť bodu zlomu v centromerickej oblasti bola tiež príčinou neúspechu identifikácie reciprokej cetromerickej translokácie medzi chromozómom 3 a 8 aj u triploidných jedlých banánovníkov Východoafrickej vysočiny (Nyamwihogora, Nshika a Intama). Táto centromerická translokácia nebola identifikovaná ani v práci Dupouy *et al*. (2019) s použitím sekvenovania párových knižníc pomocou technológie Illumina (prístupom mate-pairs).

Medzi ďalšie analyzované triploidné klony patrili zástupcovia tzv. plantain banánovníkov, čo sú alotriploidné klony s genómom AAB. U všetkých troch analyzovaných druhov bola pomocou metódy maľovania chromozómov objavená len jedna dlhá translokácia medzi chromozómami 1 a 3, špecifická pre B-subgenome (Šimoníková *et al*. 2019 & 2020). Vzhľadom k tomu, 6e tieto genómy obsahujú tri vzájomne divergované subgenómy, a zostavenie (assembly) čiastočných ONT dát u klonov Mululou a Hartón Tigre viedlo k vytvoreniu veľmi fragmentovanej assembly, ktoré znemožnila zostavenie dlhých kontigov obsahujúcich daný bod zlomu. Oproti tomu, pre klon 3 Hand Planty bolo získané mnohonásobne väčšie množstvo ONT čítaní (13 709 551 čítaní), ktoré umožnilo zostaenie kvalitnejšej a viac kompletnej assembly. Mapovanie kontigov a scaffoldov získaných zostavením ONT dát klonu 3 Hand Planty umožnilo identifikovať kontigy nesúce translokáciu 1/3 a 3/1. Tento príklad ukazuje, že pri vysoko hetozygotných rastlinných druhoch, ako je prevažná časť nami študovaných zástupcov banánovníka, je relatívne nízke (10 – 15 x coverage) sekvenačné prekrytie ONT dátami nepoužiteľné na jednoznačnú identifikáciu dlhých štruktúrnych zmien v genóme. Ďalšou možnou bariérou brániacou identifikácii, je rozdielna veľkosť genómu.

Pre prekonanie týchto problémov pre dobrú identifikáciu prestavieb u vysoko heterozygotných alebo polyploidných druhov rastlín by bolo potreba viacerých krokov. Vytvoriť lepšiu a hlavne kvalitnejšiu assembly, čiže pracovať s vyšším pokrytím sekvenačných dát, alebo využiť nižšie pokrytie (10 – 15 x) sekvenačnými dátami dlhých čítaní s vysokou kvalitou. To by sa mohlo dosiahnuť použitím PacBio Hifi čítania, prípadne začlenením Hi-C dať pre zostavenie dlhých scaffoldov, a overenie správneho zostavenia chromozomálnych sekvencií by sa dalo dosiahnuť využitím Bionano optického mapovania (Belser *et al*. 2021).

Zaujímavým prípadom, dokladajúcim už vyššie spomínané, je M. beccarii (sekcia Callimus, ITC 1070), druh s najväčším genómom. V našich dátach mal najväčšiu priemernú dĺžku čítania a jeho assembly malo tak lepšiu dĺžku kontigov. Aj keď ONT dáta mali menšie pokrytie (10x menej čítaní), vysoká miera homozygotnosti viedla k zostaveniu kvalitnejších a dlhších kontigov/scaffoldov a tým pádom aj k zvýšeniu pravdepodobnosti identifikácie chromozomálnych prestavieb.

Pomocou nášho *in silico* prístupu pre identifikáciu genómových prestavieb v assembly študovaných zástupcov banánovníka sa nám tiež podarilo identifikovať možné doposiaľ neodhalené translokácie. Tieto prestavby nemuseli byť identifikované s farbením chromozómov, kvôli nízkemu rozlíšeniu fluorescenčnej in situ hybridizácie. Ich správnosť by bolo vhodné potvrdiť alebo vyvrátiť s využitím ďalších komplexných metodických prístupov, vyššie spomenutých – najlepšie vytvorením tzv. chromosome-scale assembly.

Všeobecne sa dá povedať, že výsledky našej práce ukazujú, že aj s relatívne nízkym pokrytím vysoko chybových ONT čítaní (10 – 15 x) možno u vysoko homozygotných druhov využiť pre relatívne kvalitné assembly, v ktorej sa dajú identifikovať dlhé genómové prestavby, a vieme určiť približne aj bod zlomu – presnosť identifikácie bodu zlomu závisí od homológie s referenčnou genómovou sekvenciou. Čím fylogeneticky bližší je genóm referenčnej sekvencie študovaných druhom, tým presnejšia bude identifikácia budov zlomu. Oproti tomu, pri vysoko heterozygotných alebo polyploidných rastlinných druhoch, využitie rovnakého pokrytia ONT dátami, na 1C alebo 1Cx genóm, viedlo k zosadeniu fragmentovanej assembly a nemožnosti identifikácie dlhých genómových prestavieb, či identifikácii bodu zlomu. Ďalším, všeobecný problém predstavujú genómové prestavby, ktoré nastali vo vysoko variabilných centromerických oblastiach chromozómov. Opäť, nezávislé vytvorenie kvalitných chromosome-scale assembly umožní jednoznačnú identifikáciu všetkých dlhých chromozomálnych prestavieb v rámci širokého spektra príbuzných rastlinných druhov.

6 ZÁVER

Diplomová práca je zameraná na identifikáciu chromozomálnych translokácií v rode *Musa* spp. pomocou sekvencií získaných sekvenačnou metódou Oxford Nanopore. Literárna rešerš je zameraná na oblasť banánovníkov, ich genóm, sekvenačnú metódu ONT a bioinformatickú analýzu vzniknutých dát. Praktická časť pojednáva o ceste (pipeline), ktorá umožňuje spracovanie ONT dát a ich použitiu k identifikácií chromozomálnych translokácií vyskytujúcich sa v *Musa* spp.. Získané dáta umožňujú vizualizáciu v DGenies.

Cieľom experimentálne časti bolo vytvorenie cesty umožňujúcej identifikáciu chromozomálnych prestavieb, presnejšie bodov zlomu, pomocou ONT čítaní a ich assembly.

U sekvenačných dát najskôr prebehlo volanie báz (basecalling) na serveroch Centra štruktúrnej a funkčnej genomiky rastlín, ktoré sú súčasťou MetaCentra, kde prebiehali aj následne použité príkazy. Vytvorili sme assembly pre vybrané kultivary a následne ich použili pre mapovanie na referenčnú genómovú sekvenciu *M. acuminata* spp. *malaccensis* (klon DH Pahang, verzia 4). Namapované assembly kontigy boli spracované pomocou série príkazov používajúcich nástroj SAMtools. Výsledné vybrané kontigy boli znova mapované na referenciu a upravené, aby mali aspoň minimálnu dĺžku 30 kb. Pre vizualizáciu získaných kontigov bol vytvorený skript pre extrakciu sekvencií a tie boli následne vložené spolu s referenčnou sekvenciou do webového nástroja DGenies. Ten vytvoril Dot plot, znázorňujúci identifikované translokácie.

Táto cesta bola vyskúšaná aj na ONT čítaniach, u ktorých prebehlo len volanie báz, a nedosiahla požadovaných výsledkov. Ukázalo sa, že samostatné čítania na túto prácu nie sú vhodné, a že vytvorenie assembly je dôležitým krokom v identifikácií translokácií pomocou ONT sekvenovania. Čítania sme napriek tomu využili pre vizualizáciu bodov zlomu v IGV. Najskôr sme čítania namapovali na podozrivé kontigy získané z assembly, upravili ich so SAMtools a namapovali na referenciu. Potom sa vybrali len tie čítania, ktoré presahovali priemernú dĺžku čítania stanovenú s NanoPlot. Pre orientáciu v IGV sa použili tabuľky vytvorené so skriptom, ktorý extrahoval dôležité informácie zo SAM súboru obsahujúceho kontigy vykazujúce translokáciu.

Vytvorená cesta sa ukázala ako pomerne účinná, a pri kvalitnej assembly dokáže identifikovať translokácie v genóme. ONT však stále robia problémy vysoko variabilné oblasti s prítomnosťou tandemovo repetetívnej DNA – napr. centromerické oblasti, ktorým nedokážeme identifikovať chromozomálne prestavby. Nové, doposiaľ neidentifikované chromozomálne prestavby, ktoré sa použitím našej novo vytvorenej programovovej pipeline podarilo identifikovať, bude potreba ďalej potvrdiť pomocou ďalších metód alebo vylepšenia vstupných dát.

7 LITERATÚRA

Acevedo, S. A., Carrillo, Á. J. D., Flórez-López, E., & Grande-Tovar, C. D. (2021). Recovery of Banana Waste-Loss from Production and Processing: A Contribution to a Circular Economy. *Molecules*, *26*(17). https://doi.org/https://doi.org/10.3390/molecules26175282

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol.*, *215*(3), 403-413. https://doi.org/10.1016/S0022-2836(05)80360-2

Ashton, P. M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., Wain, J., & O'Grady, J. (2015). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology*, *33*(3), 296-300. https://doi.org/10.1038/nbt.3103

Baurens FC, Martin G, Hervouet C, Salmon F, Yohomé D, Ricci S, Rouard M, Habas R, Lemainque A, Yahiaoui N, D'Hont A. (2019). Recombination and Large Structural Variations Shape Interspecific Edible Bananas Genomes. *Molecular Biology Evolution*. 1;36(1):97-111. doi: 10.1093/molbev/msy199

Bayley, H. (2015). Nanopore sequencing: from imagination to reality. *Clinical Chemistry*, *61*(1), 25–31. https://doi.org/10.1373/clinchem.2014.223016

Belser, C., Baurens, F. -C., Noel, B., Martin, G., Cruaud, C., Istace, B., Yahiaoui, N., Labadie, K., Hřibová, E., Doležel, J., Lemainque, A., Wincker, P., D'Hont, A., & Aury, J. -M. (2021). Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. *Communications Biology*, *4*(1): 1047. https://doi.org/https://doi.org/10.1038/s42003-021-02559-3

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Bleidorn, C. (2016). Third generation sequencing: Technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*, *14*(1), 1-8. https://doi.org/10.1080/14772000.2015.1099575

Borrell, J. S., Biswas, M. K., Goodwin, M., Blomme, G., Schwarzacher, T., Heslop-Harrison, J. S. (P.), Wendawek, A. M., Berhanu, A., Kallow, S., Janssens, S., Molla, E. L., Davis, A. P., Woldeyes, F., Willis, K., Demissew, S., & Wilkin, P. (2019). Enset in Ethiopia: a poorly characterized but resilient starch staple. *Annals of Botany*, *123*(5), 747–766. https://doi.org/https://doi.org/10.1093/aob/mcy214

Cifuentes, M., Grandont, L., Moore, G., Chèvre, A. M., & Jenczewski, E. (2010). Genetic regulation of meiosis in polyploid species: new insights into an old question. *New Phytologist*, *186*(1), 29-36. https://doi.org/10.1111/j.1469-8137.2009.03084.x

Čížková J, Hřibová E, Christelová P, Van den Houwe I, Häkkinen M, Roux N, Swennen R, Doležel J.(2015). Molecular and Cytogenetic Characterization of Wild Musa Species. *PLoS One*, 10(8):e0134096. doi: 10.1371/journal.pone.0134096

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience, 10(2). https://doi.org/https://doi.org/10.1093/gigascience/giab008

Dash, P. K., & Rai, R. (2016). Translating the "Banana Genome" to Delineate Stress Resistance, Dwarfing, Parthenocarpy and Mechanisms of Fruit Ripening. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.01543

Davey, M. W., Gudimella, R., Harikrishna, J. A., Wan Sin, L., Khalid, N., & Keulemans, J. (2013). "A draft Musa balbisiana genome sequence for molecular genetics in polyploid, inter- and intra-specific Musa hybrids." *BMC Genomics*, *14*:683. https://doi.org/10.1186/1471-2164-14-683

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. Bioinformatics, 34(15), 2666–2669. https://doi.org/https://doi.org/10.1093/bioinformatics/bty149

D-Genies. https://dgenies.toulouse.inra.fr/run (01.05.2023).

de Koning, W., Miladi, M., Hiltemann, S., Heikema, A., Hays, J. P., Flemming, S., van den Beek, M., Mustafa, D. A., Backofen, R., Gruning, B., & Stubbs, A. P. (2020). NanoGalaxy: Nanopore long-read sequencing data analysis in Galaxy. *GigaScience*, *9*(10), 1-7. https://doi.org/10.1093/gigascience/giaa105

D'Hont A, Denoeud F, Aury JM., Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, et al. (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. (2012). *Nature*, vol.488, 213–217. https://doi.org/10.1038/nature11241

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., & Bettman, B. *et al.* (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, Vol 323(5910), 133-138. https://doi.org/10.1126/science.1162986

Feng, Y., Zhang, Y., Ying, C., Wang, D., & Du, C. (2015). Nanopore-based fourth-generation DNA sequencing technology. *Genomics Proteomics Bioinformatics*, *13*(1), 4-16. https://doi.org/10.1016/j.gpb.2015.01.009

Fu, S., Wang, A., & Au, K. F. (2019). A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biology*, *20*(1):26. https://doi.org/https://doi.org/10.1186/s13059-018-1605-z

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular ecology resources*, *11*(5), 759-769. https://doi.org/10.1111/j.1755-0998.2011.03024.x

Guo, J., Fu, H., Yang, Z., Li, J., Jiang, Y., Jiang, T., Liu, E., & Duan, J. (2021). Research on the Physical Characteristic Parameters of Banana Bunches for the Design and Development of Postharvesting Machinery and Equipment. *Agriculture*, *11*(4) :362. https://doi.org/https://doi.org/10.3390/agriculture11040362

Gong, L., Wong, C. -H., Idol, J., Ngan, C. Y., & Wei, C. -L. (2019). Ultra-long Read Sequencing for Whole Genomic DNA Analysis. *Journal of Visualized experiments*, (145). https://doi.org/10.3791/58954

Gong, L., Wong, C. -H., Cheng, W. -C., Tjong, H., Menghi, F., Ngan, C. Y., Liu, E. T., & Wei, C. -L. (2018). Picky Comprehensively Detects High Resolution Structural Variants in Nanopore Long Reads. *Nat Methods*, *15*(6), 455-460. https://doi.org/10.1038/s41592-018-0002-6

Hornblower, B., Coombs, A., Whitaker, R. D., Kolomeisky, A., Picone, S. J., Meller, A., & Akeson, M. (2007). Single-molecule analysis of DNA-protein complexes using nanopores. *Nature Methods*, *4*(4), 315-7. https://doi.org/10.1038/nmeth1021

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, *82*(11), 801-811. https://doi.org/https://doi.org/10.1016/j.humimm.2021.02.012

Chen, P., Gu, J., Brandin, E., Kim, Y. -R., Wang, Q., & Branton, D. (2004). Probing Single DNA Molecule Transport Using Fabricated Nanopores. *Nano Letters*, *4*(11), 2293–2298. https://doi.org/doi: 10.1021/nl048654j

Cherf, G. M., Lieberman, K. R., Rashid, H., Lam, C. E., Karplus, K., & Akeson, M. (2012). Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nature Biotechnology*, *30*(4), 344-8. https://doi.org/10.1038/nbt.2147

Ip, C. L. C., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., Leggett, R. M., Eccles, D. A., Zalunin, V., Urban, J. M., Piazza, P., Bowden, R. J., Paten, B., Mwaigwisya, S., Batty, E. M., Simpson, J. T., Snutch, T. P., Birney, E., Buck, D., et al. (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis [version 1; peer review: 2 approved]. *F1000Research*, *4:1075*. https://doi.org/10.12688/f1000research.7201.1

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community, *17*(1), 239. https://doi.org/10.1186/s13059-016-1103-0

Kasianowicz, J. J., Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, *93*(24), 13770-3. https://doi.org/10.1073/pnas.93.24.13770

Lieberman, K. R., Cherf, G. M., Doody, M. J., Olasagasti, F., Kolodji, Y., & Akeson, M. (2010). Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *Journal of the American Chemical Society*, *132*(50), 17961-72. https://doi.org/10.1021/ja1087612

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, *32*(14), 2103-10. https://doi.org/doi: 10.1093/bioinformatics/btw152

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 25(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Lima, L., Marchet, C., Caboche, S., Da Silva, C., Istace, B., Aury, J. -M., Touzet, H., & Chikhi, R. (2020). Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data. *Brief Bioinform.*, *21*(4), 1164-1181. https://doi.org/10.1093/bib/bbz058

Karamura, D., Karamura, E., & Blomme, G. (2011). General Plant Morphology of Musa. In M. Pillay & A. Tenkouano, *Banana Breeding: Progress and Challenges* (pp. 1-20). CRC Press.

Karamura, E.B. and D.A. Karamura. 1995. Banana morphology. Part 2: The aerial shoot. In: Bananas and plantains, S.R. Gowen, ed., 190–205. London: Chapman and Hall.

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

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, *vol.3*, 1-8. https://doi.org/10.1016/j.bdq.2015.02.001

MacBryde, B. (2009). Consensus document on the biology of bananas and plantains (Musa sp.) Vol. 43. Paris. *OECD Environment Health Safety Publications*, 43:1-87.

Manrao, E. A., Derrington, I. M., Laszlo, A. H., Langford, K. W., Hopper, M. K., Gillgren, N., Pavlenok, M., Niederweis, M., & Gundlach, J. H. (2012). Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature Biotechnology*, *30*(4), 349-53. https://doi.org/10.1038/nbt.2171

Martin, G., Baurens, F. -C., Hervouet, C., Salmon, F., Delos, J. -M., Labadie, K., Perdereau, A., Mournet, P., Blois, L., & Dupouy, M. (2020). Chromosome reciprocal translocations have accompanied subspecies evolution in bananas. *Plant Journal*, *104*(6), 1698–1711. https://doi.org/10.1111/tpj.15031

Martin G, Baurens FC, Droc G, Rouard M, Cenci A, Kilian A, Hastie A, Doležel J, Aury JM, Alberti A, Carreel F, D'Hont A. (2016). Improvement of the banana "Musa acuminata" reference sequence

using NGS data and semi-automated bioinformatics methods. *BMC Genomics*, 17:243. doi: 10.1186/s12864-016-2579-4.

Mason, C. E., & Elemento, O. (2012). Faster sequencers, larger datasets, new challenges. *Genome Biology*, *13*(3). https://doi.org/10.1186/gb-2012-13-3-314

Meller, A., Nivon, L., Brandin, E., Golovchenko, J., & Branton, D. (2000). Rapid nanopore discrimination between single polynucleotide molecules. *Comparative Study*, *97*(3), 1079-1084. https://doi.org/10.1073/pnas.97.3.1079

Nussbaum, R.L., Mcinnes, R.R., Willard, H.W. (2004). *Klinická genetika*. 6.vydanie. Triton. Praha.

Oxford Nanopore Technologies. https://nanoporetech.com (01.05.2023).

PacBio Revio. https://www.pacb.com/revio/ (15.05.2023).

Paggi, M., & Spreen, T. (Eds.). (2003). Overview of the World Banana Market. In T. E. Josling & T. G. Taylor, *BANANA WARS The Anatomy of a Trade Dispute* (pp. 7-17). CABI Publishing.

Pevsner, J. (2015). *Bioinformatics and Functional Genomics* (3rd edition). John Wiley & Sons.

Ploetz, R. C. (2015). Fusarium Wilt of Banana. *Phytopathology*, *vol.105*(Number 12), 1512–1521. https://doi.org/https://doi.org/10.1094/PHYTO-04-15-0101-RVW

Ploetz, R. C., Kepler, A. K., Daniells, J., & Nelson, S. C. (2007). Banana and plantain—an overview with emphasis on Pacific island cultivars. *Species profiles for Pacific Island agroforestry*, *1*, 21-32, [Internet]. Avaiable: http://www.bananenzeug.ch/wp-content/uploads/2018/06/banana-plantain-overview.pdf .

Pushkarev, D., Neff, N. F., & Quake, S. R. (2009). Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, *27*(9), 847–850. https://doi.org/10.1038/nbt.1561

Quick, J., Quinlan, A. R., & Loman, N. J. (2015). A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *GigaScience*, *4*(1), s13742–015–0043–z. https://doi.org/https://doi.org/10.1186/s13742-015-0043-z

Rekha, A., & (Eds.). (2016). History, Origin, Domestication, and Evolution. In S. Mohandas & K. V. Ravishankar, *Banana: Genomics and Transgenic Approaches for Genetic Improvement* (pp. 3-11). Springer Singapore.

Robinson, J. C., & Saúco, V. G. (2010). *Bananas and Plantains, 2nd Edition* (2nd Edition). CAB International.

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

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*, 461–468. https://doi.org/https://doi.org/10.1038/s41592-018-0001-7

Sheperd, K. (1999). Cytogenetics of the genus Musa. Montpellier, France

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, *19*(R2), 227-240. https://doi.org/10.1093/hmg/ddq416.

Schiessl, S. -V., Katche, E., Ihien, E., Chawla, H. S., & Mason, A. S. (2019). The role of genomic structural variation in the genetic improvement of polyploid crops. *The Crop Journal*, *vol.7*(2), 127–140. https://doi.org/https://doi.org/10.1016/j.cj.2018.07.006

Schneider, G. F., & Dekker, C. (2012). DNA sequencing with nanopores. *Nature Biotechnology*, *30*(4), 326–328. https://doi.org/10.1038/nbt.2181

Si, W., & Aksimentiev, A. (2017). Nanopore Sensing of Protein Folding. *ASC Nano*, *11*(7), 7091–7100. https://doi.org/doi: 10.1021/acsnano.7b02718

Simmonds, N. W. (1962). *The Evolution of the Bananas* (1st ed.). Longmans.

Sović, I., Šikić, M., Wilm, A., Fenlon, S. N., Chen, S., & Nagarajan, N. (2016). Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun*, *7:* 11307. https://doi.org/10.1038/ncomms11307

Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G., & Bayley, H. (2009). Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences*, *106*(19), 7702-7707. https://doi.org/10.1073/pnas.0901054106

Stover, R. H., & Simmonds, N. W. (1987). *Bananas*. Wiley.

Šimoníková, D., Němečková, A., Čížková, J., Brown, A., Swennen, R., Doležel, J., & Hřibová, E. (2020). Chromosome Painting in Cultivated Bananas and Their Wild Relatives (Musa spp.) Reveals Differences in Chromosome Structure. *International Journal of Molecular Sciences*, *21*(21) :7915. https://doi.org/10.3390/ijms21217915

Šimoníková, D., Němečková, A., Karafiátová, M., Uwimana, B., Swennen, R., Doležel, J., & Hřibová, E. (2019). Chromosome Painting Facilitates Anchoring Reference Genome Sequence to Chromosomes In Situ and Integrated Karyotyping in Banana (Musa Spp.). *Frontiers in Plant Science*, *10*, :1503. https://doi.org/10.3389/fpls.2019.01503

Turner, D. W., Fortescue, J. A., & Thomas, D. S. (2007). Environmental physiology of the bananas (Musa spp.). *Brazilian Journal of Plant Physiology*, *19*(4), 463-484. https://doi.org/10.1590/S1677-04202007000400013

Varongchayakul, N., Song, J., Meller, A., & Grinstaff, M. W. (2018). Single-molecule protein sensing in a nanopore: a tutorial. *Chemical Society Reviews*, *47*(23), 8512–8524. https://doi.org/doi: 10.1039/c8cs00106e

Wang, Z., Rouard, M., Droc, G., Heslop-Harrison, P.J.S., Ge X.-J.. (2023). Genome assembly of Musa beccarii shows extensive chromosomal rearrangements and genome expansion during evolution of Musaceae genomes. *Gigascience*. Volume 12,giad005. doi: 10.1093/gigascience/giad005

Wang, Z., Rouard,R., Biswas, M.K., Droc, G., Cui, D., Roux, N., Baurens, F.-Ch., Ge, X.-J., Schwarzacher, T., *at al.* (2022). A chromosome-level reference genome of Ensete glaucum gives insight into diversity and chromosomal and repetitive sequence evolution in the Musaceae, *GigaScience,* Volume 11, giac027, https://doi.org/10.1093/gigascience/giac027

Wang, Y., Zhao, Y., Bollas, , Bollas, A., Wang, Y., & Au, K. F. (2021) Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**, 1348-1365,https://doi.org/10.1038/s41587-021-01108-x

Wang, Z., Zhang, J., Jia, C. H., Liu, J. H., Li, Y. Q., Yin, X. M., Xu, B. Y., & Jin, Z. Q. (2012). De novo characterization of the banana root transcriptome and analysis of gene expression under Fusarium oxysporum f. sp. Cubense tropical race 4 infection. *BMC Genomics*, *13* :650. https://doi.org/10.1186/1471-2164-13-650

Wang, H., Dunning, J. E., Huang, A. P. -H., Nyamwanda, J. A., & Branton, D. (2004). DNA heterogeneity and phosphorylation unveiled by single-molecule electrophoresis. *Proceedings of the National Academy of Sciences*, *101*(37), 13472-7. https://doi.org/10.1073/pnas.0405568101

Wu, R. 1970. Nucleotide sequence analysis of DNA. I. Partial sequence of the cohesive ends of Bacteriophage lambda and 186 DNA. *Journal of Molecular Biology* **51**, 501–521. 10.1016/0022-2836(70)90004-5

Xiong, Z., Gaeta, R. T., & Pires, J. C. (2011). Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica

napus. *Proceedings of the National Academy of Sciences (PNAS)*, *108*(19), 7908-7913. https://doi.org/https://doi.org/10.1073/pnas.1014138108

8 POUŽITÉ SKRATKY

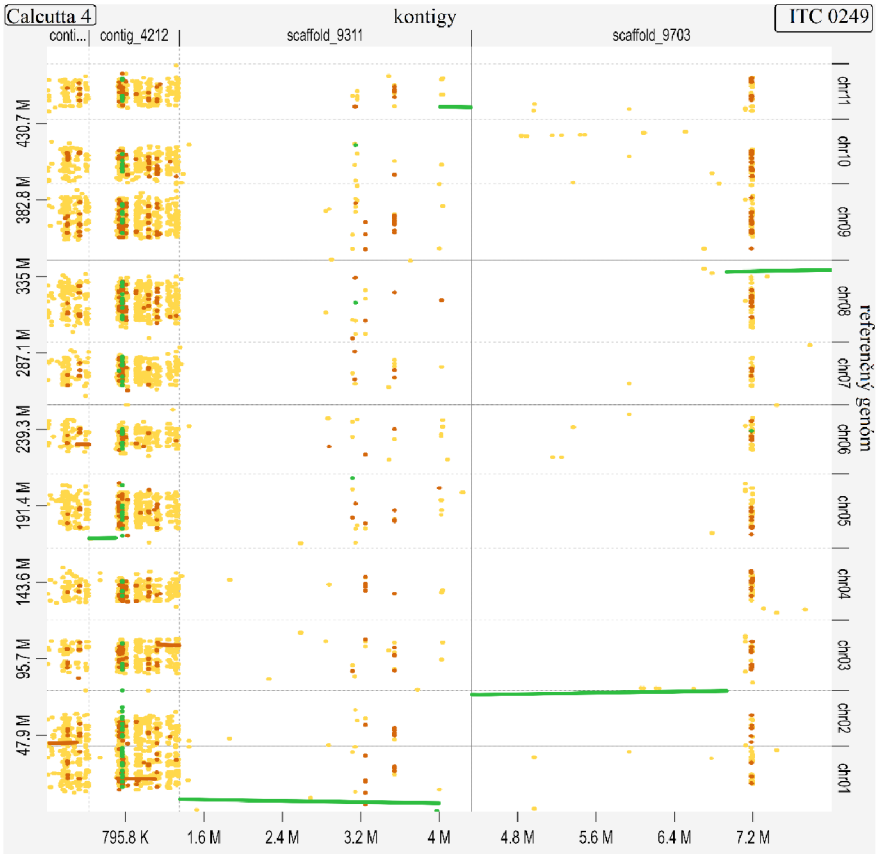
BAM	Binary File Format (binárna verzia SAM súboru)
bp	pár bázií
GB	Giga Byte
FISH	Fluorescence *in situ* hybridization (Fluorescenčná *in situ* hybridizácia)
ONT	Oxford Nanopore Technologies
SAM	Sequence Alignment Map (mapa sekvenčného zarovnania)
SNP	single nucleotide polymorfism

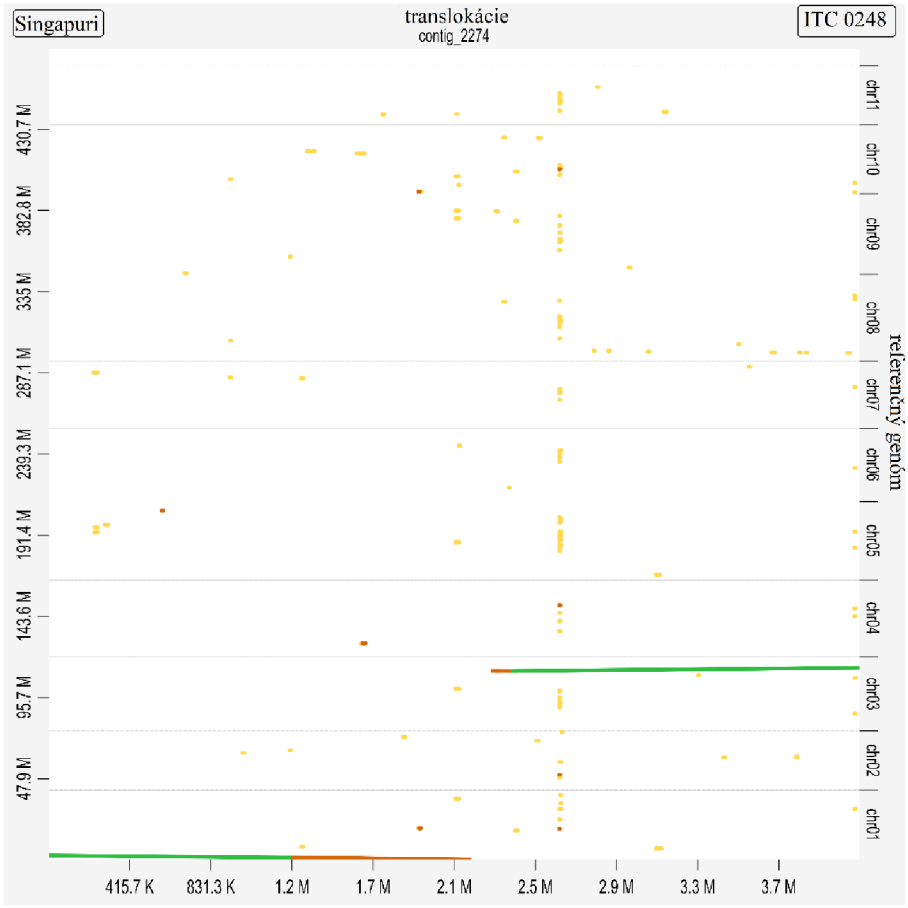
9 PRÍLOHY

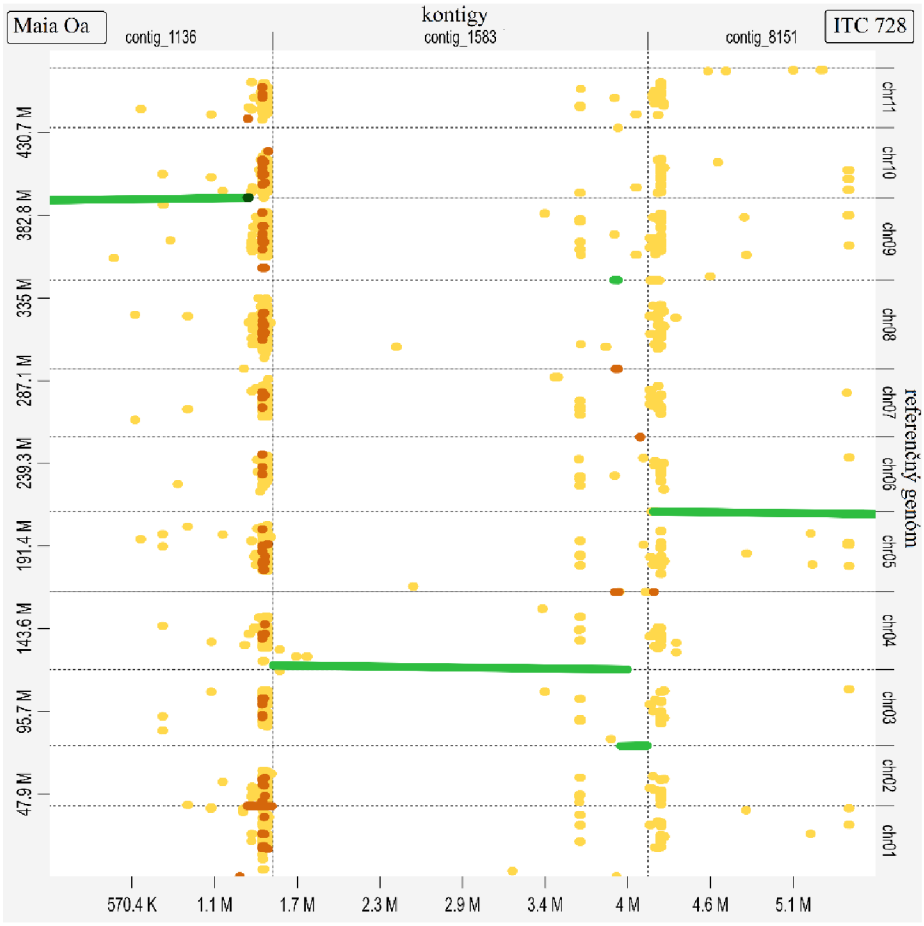
e-Príloha: Simona_Martikánová_diplomová_práca.zip:

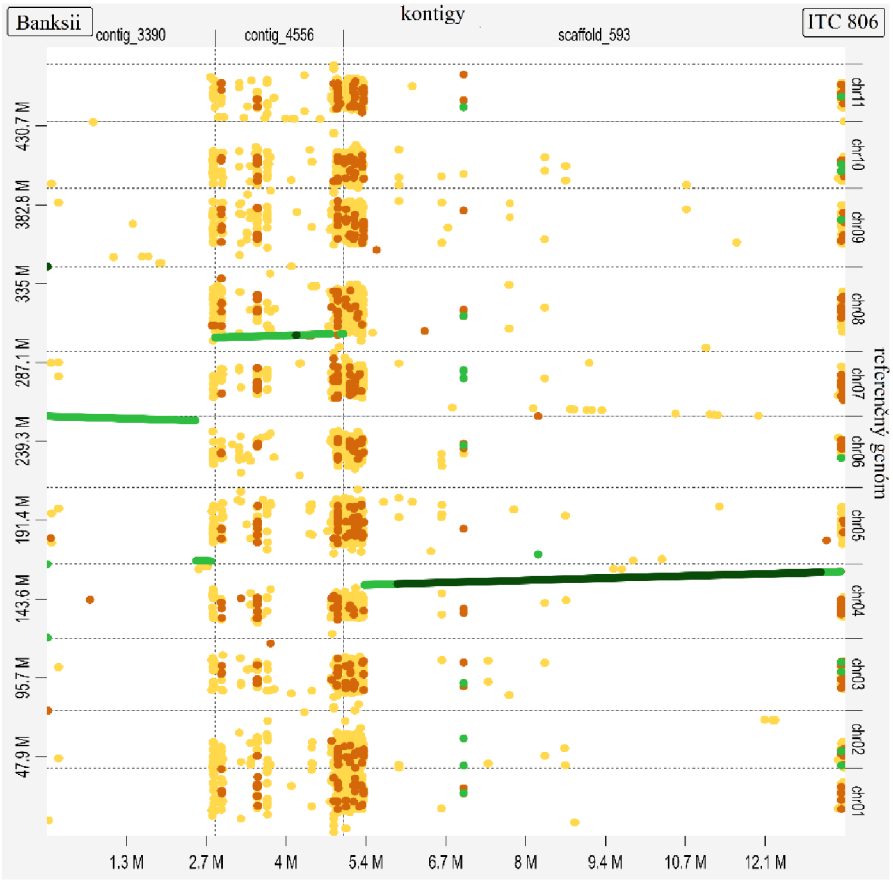
- **Príloha 1:** ukážka skriptu pre basecalling
- **Príloha 2:** ukážka skriptu pre Flye
- **Príloha 3:** skript pre získanie DNA sekvencií vo FASTA formáte zo SAM
- **Príloha 4:** skript pre získanie informácií (body zlou, dĺžka, ...) o súbore s možnými translokáciami v tabuľkovej forme
- **Príloha 5:** Nanoplot.xlxs – tabuľka s výsledkami pre NanoPlot skúmaných druhov
- **Príloha 6:** zložky 0022, 0145, 0248, 0249, 0287, 0370, 0539, 0575, 0614, 0642, 0728, 0806, 1070, 1132, 1387, 1482, 1575, 1716 obsahujúce: **zložka IGV** s obrázkami lokalizovaných bodov zlomu pre vybranú translokáciu, tabuľku výsledkov získanú s **Prílohou 4**

Príloha 7 – Vizualizácia translokácií s assembly pre sekciu Eumusa – Dgenies Dot plot

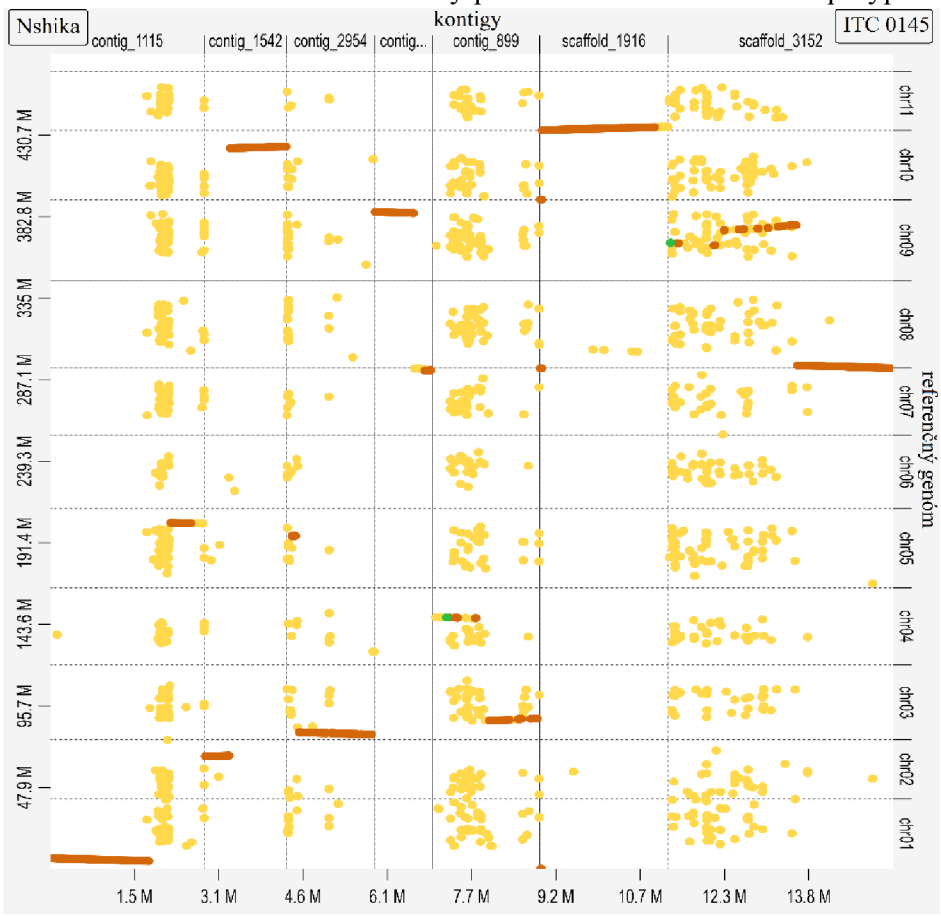


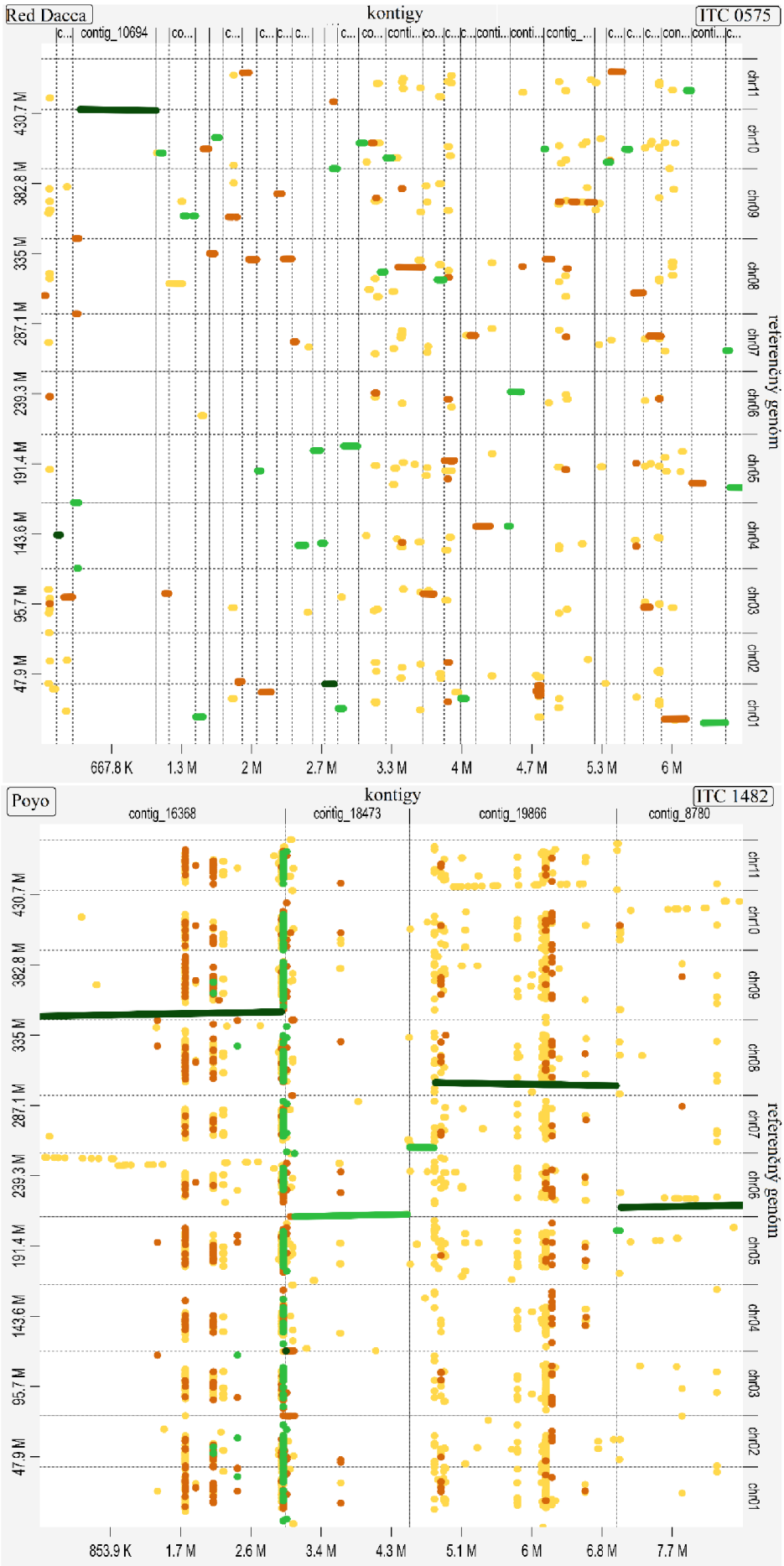

Singapuri
translokácie
contig_2274
ITC 0248
referenčný genóm
chr11
chr10
chr09
chr08
chr07
chr06
chr05
chr04
chr03
chr02
chr01
430.7 M
382.8 M
335 M
287.1 M
239.3 M
191.4 M
143.6 M
95.7 M
47.9 M
415.7 K
831.3 K
1.2 M
1.7 M
2.1 M
2.5 M
2.9 M
3.3 M
3.7 M

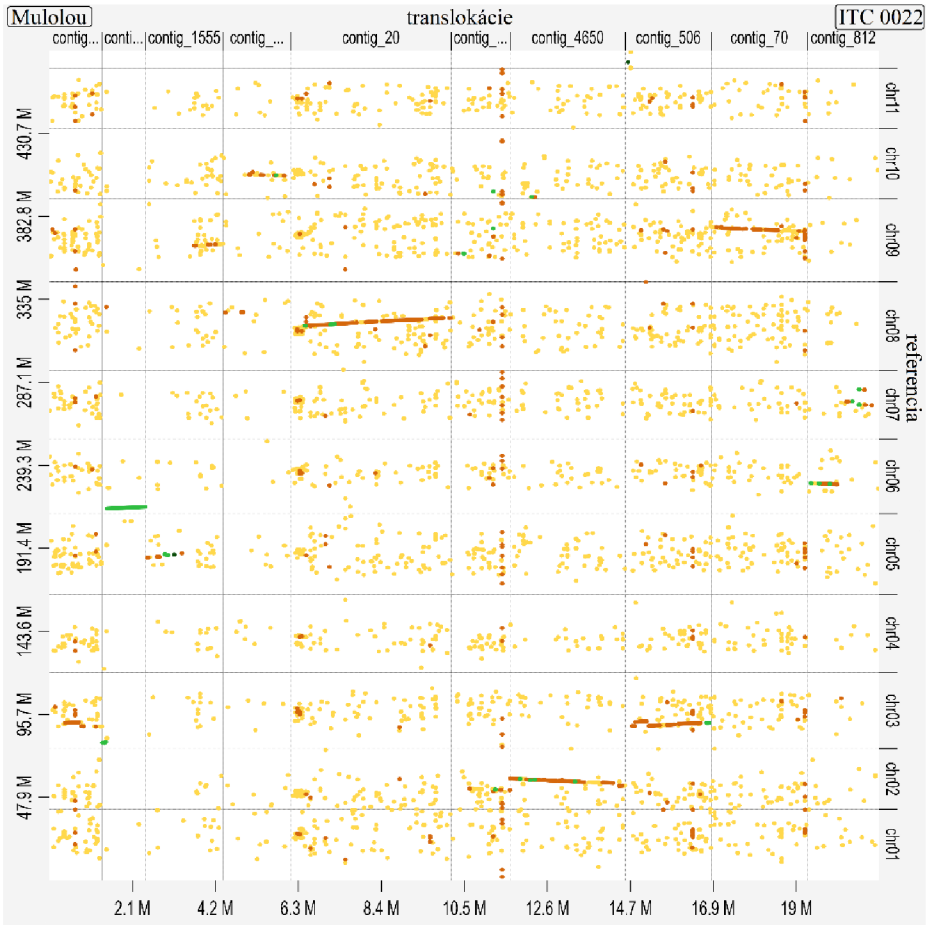

Maia Oa
kontigy
contig_1136
contig_1583
contig_8151
ITC 728
referenčný genóm
chr11
chr10
chr09
chr08
chr07
chr06
chr05
chr04
chr03
chr02
chr01
430.7 M
382.8 M
335 M
287.1 M
239.3 M
191.4 M
143.6 M
95.7 M
47.9 M
570.4 K
1.1 M
1.7 M
2.3 M
2.9 M
3.4 M
4 M
4.6 M
5.1 M

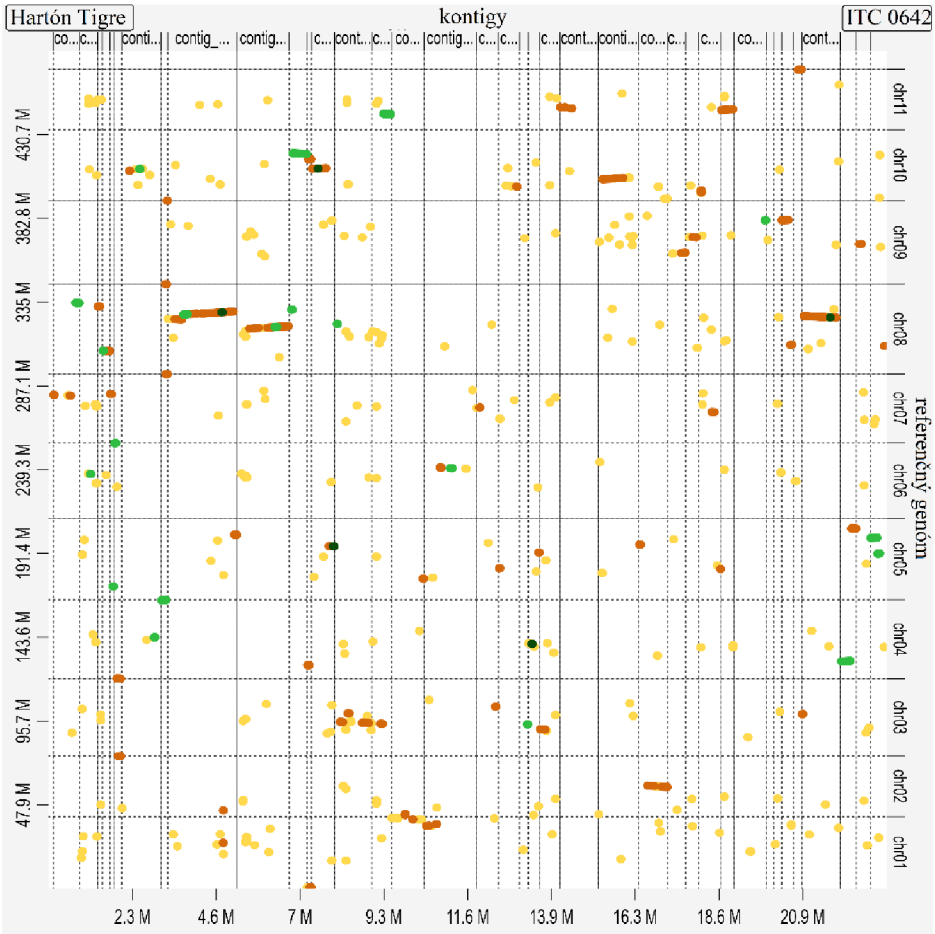


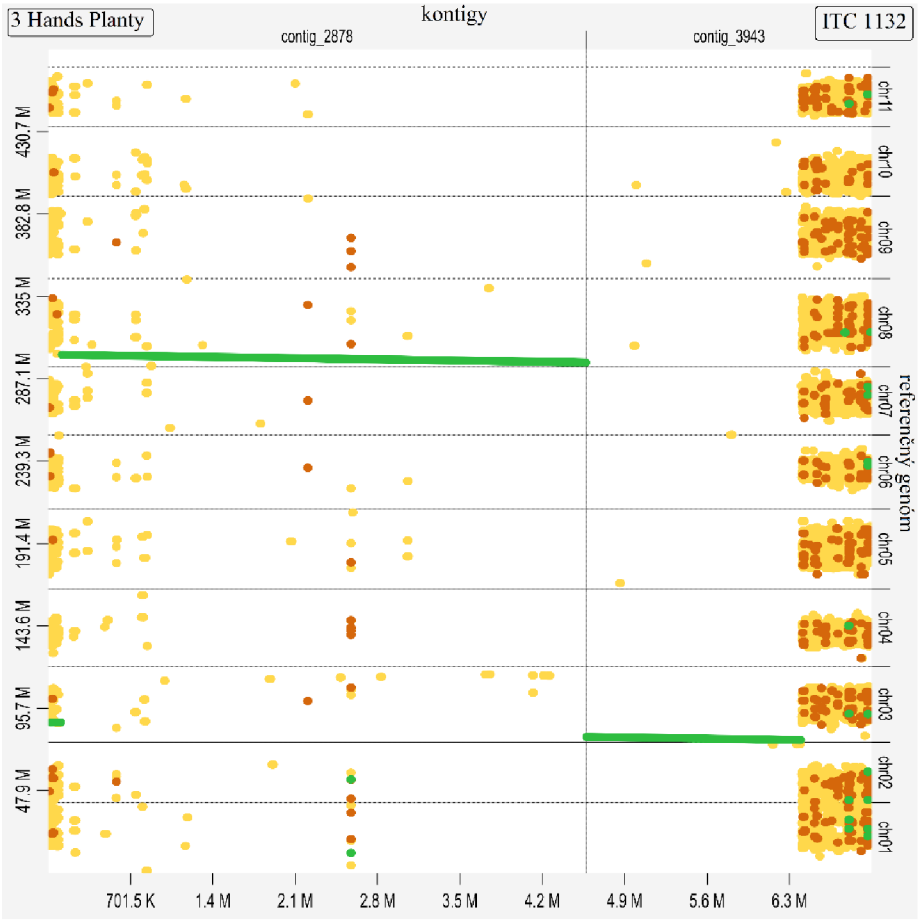
- Vizualizácia translokácií s assembly pre sekciu *Eumusa* – Plané polyploidy




Red Dacca
kontigy
ITC 0575
contig_10694
referenčný genóm
chr11
chr10
chr09
chr08
chr07
chr06
chr05
chr04
chr03
chr02
chr01
430.7 M
382.8 M
335 M
287.1 M
239.3 M
191.4 M
143.6 M
95.7 M
47.9 M
667.8 K
1.3 M
2 M
2.7 M
3.3 M
4 M
4.7 M
5.3 M
6 M
Poyo
kontigy
ITC 1482
contig_16368
contig_18473
contig_19866
contig_8780
referenčný genóm
853.9 K
1.7 M
2.6 M
3.4 M
4.3 M
5.1 M
6 M
6.8 M
7.7 M

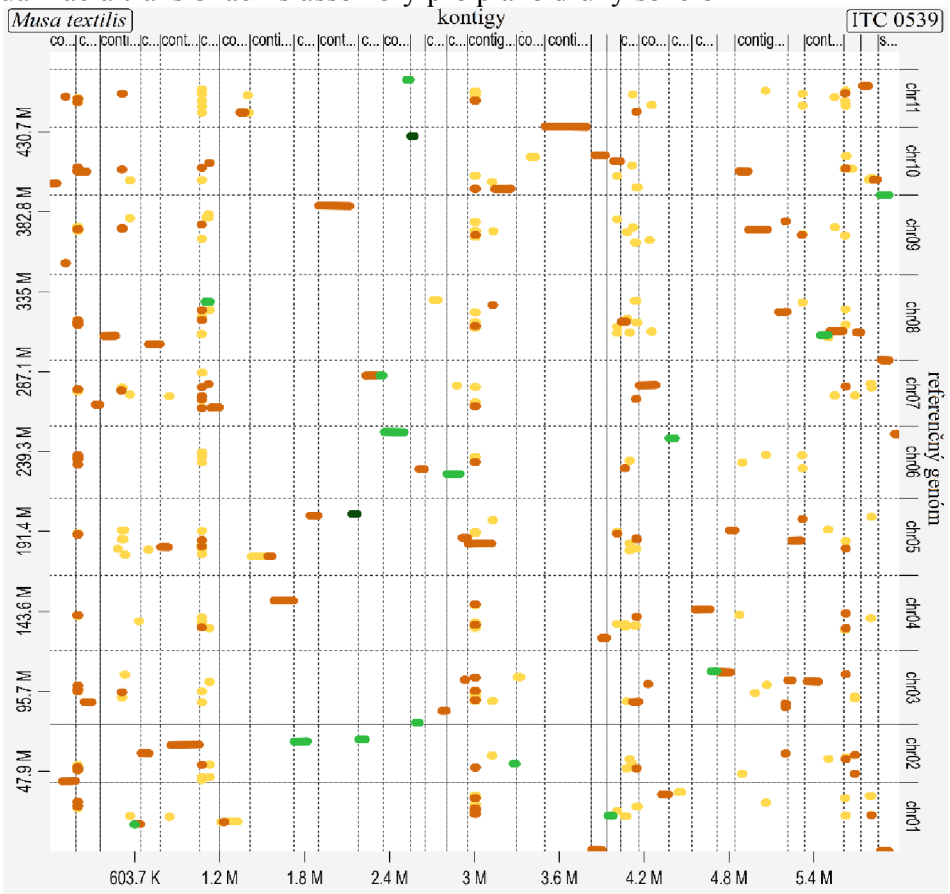

Mulolou
translokácie
ITC 0022
contig...
conti...
contig_1555
contig_...
contig_20
contig_...
contig_4650
contig_506
contig_70
contig_812
chr11
chr10
chr09
chr08
chr07
chr06
chr05
chr04
chr03
chr02
chr01
referencia
430.7 M
382.8 M
335 M
287.1 M
239.3 M
191.4 M
143.6 M
95.7 M
47.9 M
2.1 M
4.2 M
6.3 M
8.4 M
10.5 M
12.6 M
14.7 M
16.9 M
19 M

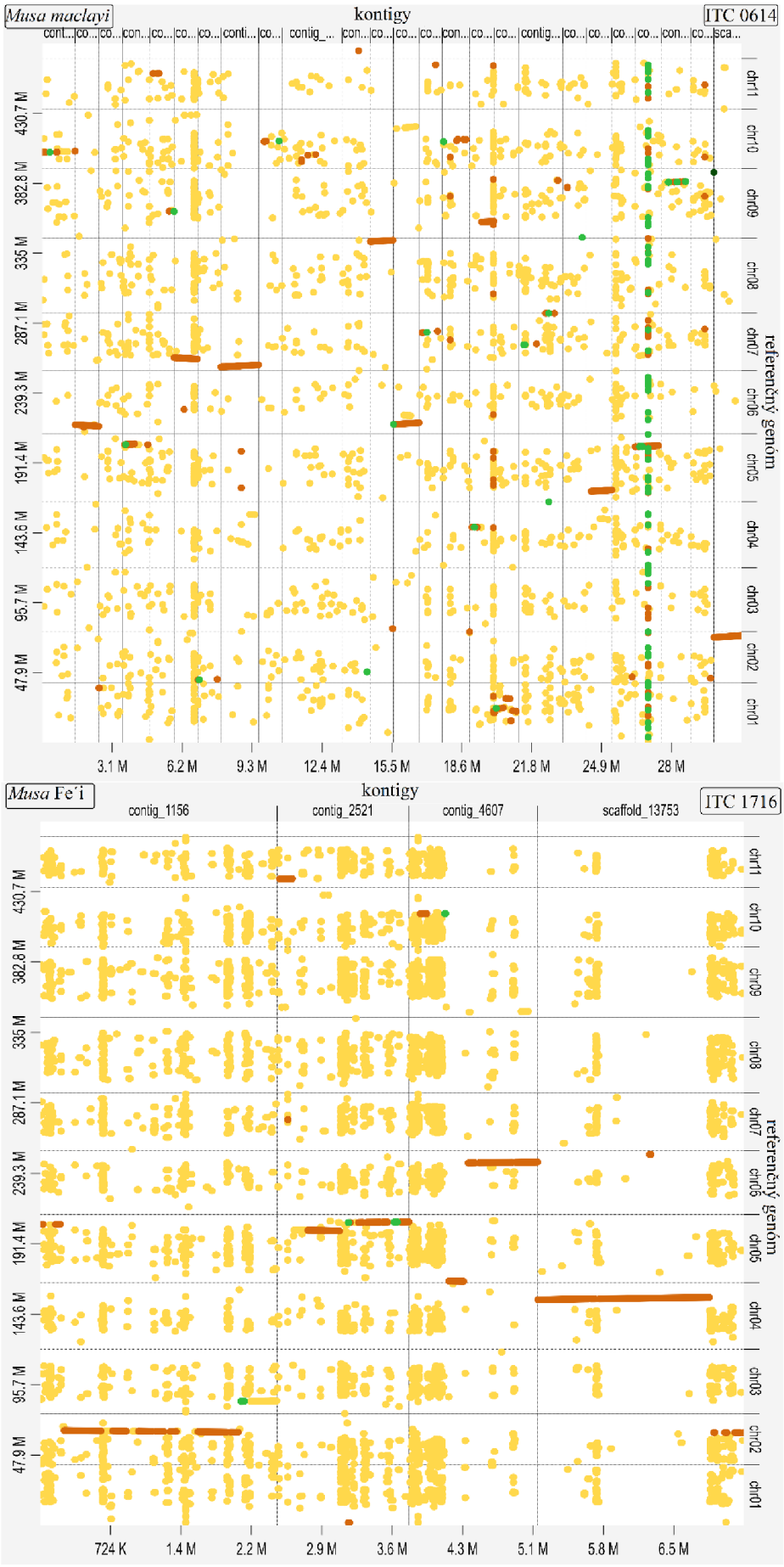

Hartón Tigre
kontigy
ITC 0642
referenčný genóm
2.3 M
4.6 M
7 M
9.3 M
11.6 M
13.9 M
16.3 M
18.6 M
20.9 M



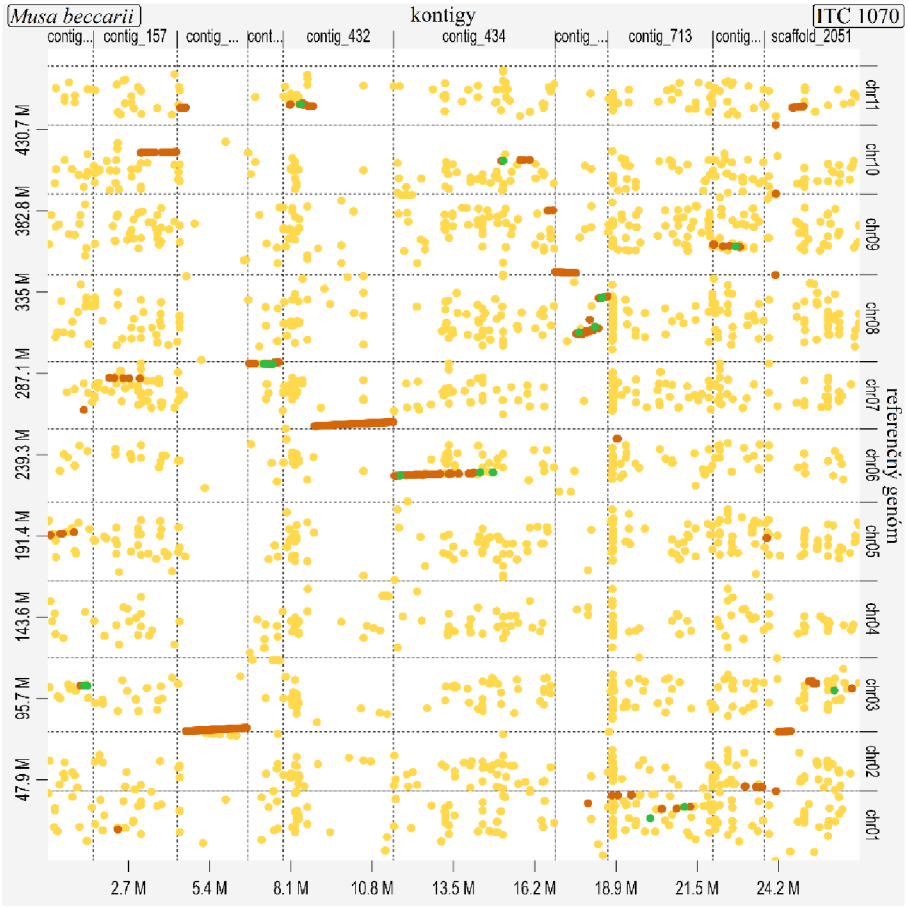
- Vizualizácia translokácií s assembly pre plané druhy sekcie *Rhodochlamys*

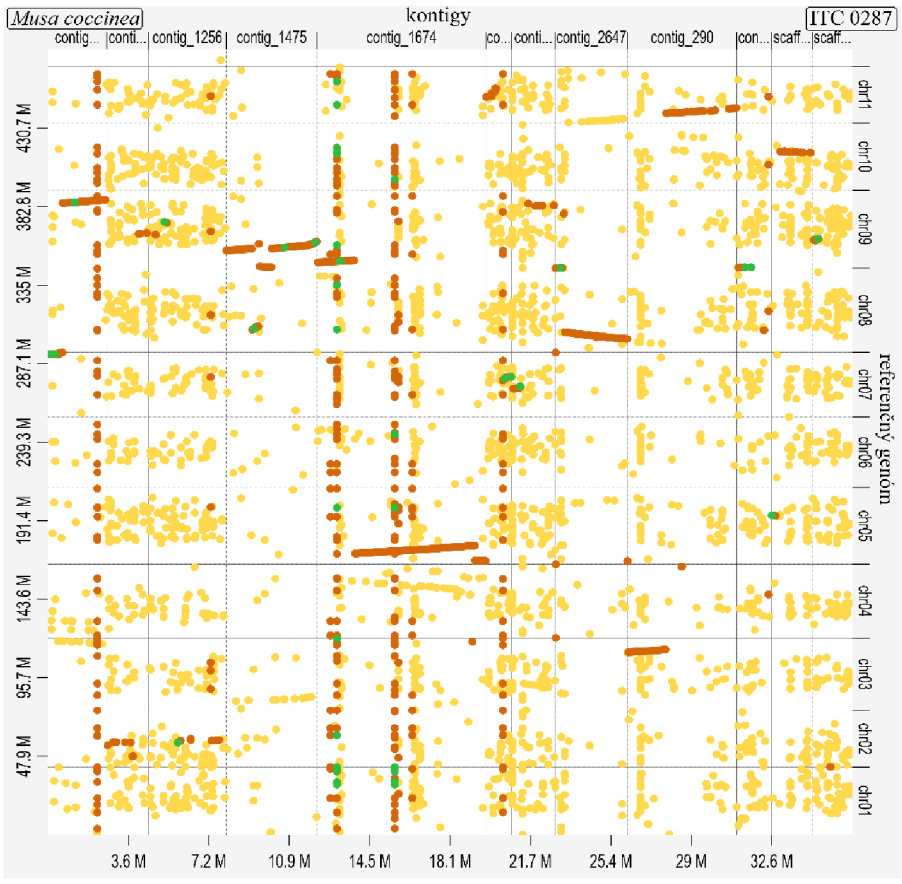
- Vizualizácia translokácií s assembly pre plané druhy sekcie *Australimusa*




Musa maclayi
kontigy
ITC 0614
referenčný genóm
Musa Fe'i
kontigy
contig_1156
contig_2521
contig_4607
scaffold_13753
ITC 1716
referenčný genóm

- Vizualizácia translokácií s assembly pre plané druhy sekcie *Callimusa*





- Vizualizácia translokácií s assembly pre *Ensete*