UNIVERZITA PALACKÉHO V OLOMOUCI

Přírodovědecká fakulta

Katedra biochemie

Analýza struktury genomu a genetické variability triploidních banánovníků - plantainů (genom AAB)

DIPLOMOVÁ PRÁCE

Autor:	**Bc. Gabriela Majzlíková**
Studijní program:	B1406 Biochemie
Studijní obor:	Bioinformatika
Forma studia:	Prezenční
Vedoucí práce:	**Mgr. Eva Hřibová, Ph.D.**
Rok:	2020

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

V Olomouci dne 9. 5. 2020

Poděkování

Velmi ráda bych touto cestou poděkovala své vedoucí diplomové práce Mgr. Evě Hřibové, Ph.D. za odborné vedení, cenné rady, věnovaný čas, vstřícnost a trpělivost při zpracování této práce.

Bibliografická identifikace

Jméno a příjmení autora	Bc. Gabriela Majzlíková
Název práce	Analýza struktury genomu a genetické variability triploidních banánovníků - plantainů (genom AAB)
Typ práce	Diplomová
Pracoviště	Katedra biochemie
Vedoucí práce	Mgr. Eva Hřibová, Ph.D.
Rok obhajoby práce	2020

Abstrakt

Banánovník (*Musa spp.*) je jednoděložná rostlina rostoucí v tropických a subtropických oblastech Afriky, jihovýchodní Asie a jižní Ameriky. Jedlé typy se vyskytují v diploidní, triploidní a tetraploidní formě. Mezi nejvýznamnější jedlé typy banánovníku patří především škrobové banány, které vznikly hybridizací diploidních, planě rostoucích druhů *Musa acuminata* (A genom) a *Musa balbisiana* (B genom). Cílem práce byla analýza Next-Gen sekvenačních dat tří hlavních morfologických skupin škrobových banánů tzv. plantainů (genom AAB) a následná analýza jednonukleotidových polymorfismů (SNP). Plody banánovníků jsou hlavním zdrojem potravy miliónů lidí zejména v rozvojových zemích a důkladné pochopení struktury genomu a genetické variability triploidních banánovníků je účinnou strategií pro jejich pěstování a pro výběr rodičů ke křížení vylepšených odrůd.

Klíčová slova	Plantain, SNP, Next-gen sekvenační data
Počet stran	66
Počet příloh	3
Jazyk	Český

Bibliographical identification

Autor's first name and surname	Bc. Gabriela Majzlíková
Title	Analysis of genome structure and genetic variability in triploid plantains (genome AAB)
Type of thesis	Diploma
Department	Department of biochemistry
Supervisor	Mgr. Eva Hřibová, Ph.D.
The year of presentation	2020

Abstract

Banana (*Musa* spp.) is a monocotyledonous plant growing in tropical and subtropical regions of Africa, Southeast Asia and South America. Edible types occur in diploid, triploid and tetraploid forms. The most important edible types of bananas include cooking bananas, which were formed by hybridization of wild diploid species *Musa acuminata* (A genome) and *Musa balbisiana* (B genome). The diploma thesis was focused on the analysis of Next-Gen sequencing data of three main morphological groups of cooking bananas known as plantains (AAB genome) and the analysis of single nucleotide polymorphisms (SNP). Banana fruits are the main source of food for millions of people, especially in developing countries. A detailed understanding of the genome structure and genetic variability of triploid bananas is an effective strategy for their cultivation and for selecting parents to cross improved varieties.

Keywords	Plantain, SNP, Next-gen data
Number of pages	66
Number of appendices	3
Language	Czech

OBSAH

Cíle práce

Využití NGS sekvenačních dat pro analýzu struktury jaderných genomů triploidních škrobových banánovníků – afrických plantainů s genomem AAB.

Hlavní cíle zahrnovali:

- Vypracování literární rešerše na zadané téma.
- Základní zpracování 150-nt dlouhých párových illumina sekvencí vybraných zástupců triploidních plantainů (genom AAB).
- Mapování párových illumina sekvencí plantainů a jejich pravděpodobných rodičovských genomů (diploidních druhů *M. acuminata* spp. *banksii* a *M. balbisiana*) na referenční celogenomovou sekvenci banánovníku *M. Acuminata* ssp. *banksii* cv. DH Pahang.
- Identifikace jednonukleotidových polymorfismů (SNP) specifických pro A- a B-subgenomy.
- Identifikace struktury chromozomů – zastoupení A- a B-subgenomů u vybraných zástupců plantainů.

1 ÚVOD

Plantainy jsou škrobové banány patřící mezi nejvýznamnější jedlé typy triploidních banánovníků (AAB). Genom plantainů obsahuje dvě sady chromozomů pocházejících z druhu *Musa acuminata* a jednu sadu chromozomů pocházejících z *Musa balbisiana*. Kultivary této podskupiny vykazují širokou škálu morfologických znaků na jejichž základě byly popsány hlavní morfologické skupiny – typ French, Horn, French Horn a False Horn. Představují jednu z nejvýznamnějších hospodářských plodin světa a je proto důležité důkladné pochopení rozmanitosti dostupných genetických zdrojů.

Teoretická část diplomové práce obsahuje poznatky o druzích banánovníků s hlavním zaměřením na triploidní plantainy. Dále obsahuje část zabývající se problematikou technologií sekvenování nové generace (NGS) a technologiemi třetí generace, které jsou známé také pod názvem long-read sequencing. V závěrečné podkapitole teoretické části jsou popsány metody zpracování sekvenačních dat poskytnutých sekvenátory NGS.

Praktická část je zaměřena na zpracování a přípravu sekvenačních dat NGS k identifikaci jednonukleotidových polymorfismů specifických pro subgenomy A a B, s následnou analýzou struktury chromozomů a zkoumáním diverzity genomových struktur u deseti vybraných zástupců afrických plantainů.

2 SOUČASNÝ STAV ŘEŠENÉ PROBLEMATIKY

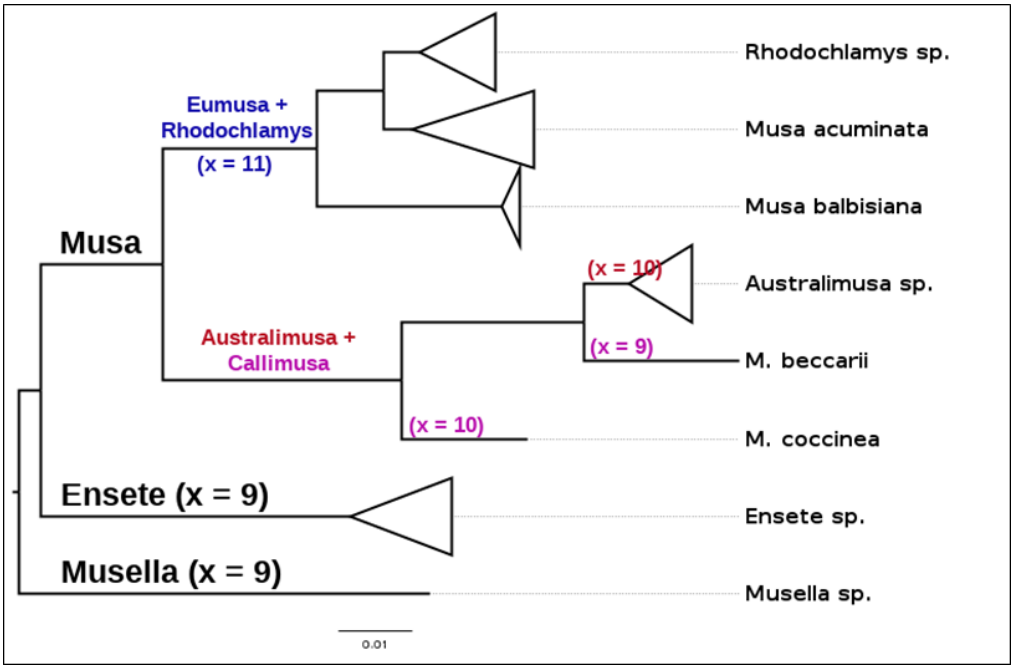
2.1 Banánovník

Banánovník (*Musa* spp.) je jednoděložná rostlina čeledi banánovníkovitých (*Musaceae*). Patří mezi rostliny vytrvalé, ale monokarpické. Banánovníky pochází z jihovýchodní Asie a západního Pacifiku a jsou jednou z prvních plodin, které byly domestikovány asi před 7 000 lety v jihovýchodní Asii (D'Hont *et al*., 2012, Davey *et al*., 2013). Dnes rostou v tropických a subtropických oblastech světa, kde jsou základním zdrojem potravy pro miliony lidí (Davey *et al*., 2013, De Langhe *et al*., 2009). Patří navíc k jedné z hlavních vývozních komodit několika rozvojových zemí a představují největší mezinárodní obchod s ovocem a mají tak důležitý socio-ekonomický význam (De Langhe *et al*., 2009, Paul *et al*., 2017).

Planě rostoucí druhy banánovníků byly dle morfologických znaků, areálu svého výskytu a základního chromozomového čísla (x) rozděleny do čtyř sekcí: *Eumusa* (x = 11; x představuje počet chromozomů), *Rhodochlamys* (x = 11), *Australimusa* (x = 10) a *Callimusa* (x = 9 nebo 10) (Cheesman *et al*., 1947, Daniells *et al*., 2001) (Obr. 1). Na základě pozdějších molekulárních analýz byly sekce *Rhodochlamys* a *Eumusa* sloučeny do sekce *Musa* a sekce *Callimusa* a *Australimusa* sloučeny do sekce *Callimusa* (Häkkinen *et al*., 2013). Převážná většina jedlých druhů banánovníku patří do sekce *Eumusa* (jinak *Musa*) a vznikly přirozeným vnitro- a mezi-druhovým křížením mezi dvěma planě rostoucími druhy *Musa acuminata* (A genom, 2n = 2x = 22) a *Musa balbisiana* (B genom, 2n = 2x = 22). Pozdější studie naznačily, že na vývoji těchto jedlých typů banánovníku se podílel i druh *Musa schizocarpa* (S genom, 2n = 2x = 22). Další skupina jedlých banánovníků vznikla nezávisle v rámci sekce *Australimusa*, jsou to tzv. Fe'i banánovníky, které jsou charakteristické vysokým obsahem β-karotenu a pěstují se jen na ostrovech Jižního Pacifiku (Davey *et al*., 2013).

Jaderný genom druhu *Musa* je relativně malý (1C ~ 500 -750 Mb, kde C je definováno jako množství DNA přítomné v jedné sadě haploidních chromosomů) (Doležel *et al*., 1994; Lysák *et al*., 1999; Asif *et al*., 2001; Kamaté *et al*., 2001; Bartoš *et al*., 2005; Čížková *et al*., 2015). Velikost haploidního genomu *Musa acuminata* se pohybuje v rozmezí 590 – 615 Mbp s predikovaným počtem genů 36 542 (Lysák *et al*., 1999, D'Hont *et al*., 2012, Bartoš *et al*., 2005). Genom A je větší než genom B

(*Musa balbisina*), jehož průměrná velikost haploidního genomu je 537 Mbp a počet genů je 35 148 (Lysák *et al.*, 1999, Wang *et al.*, 2019).



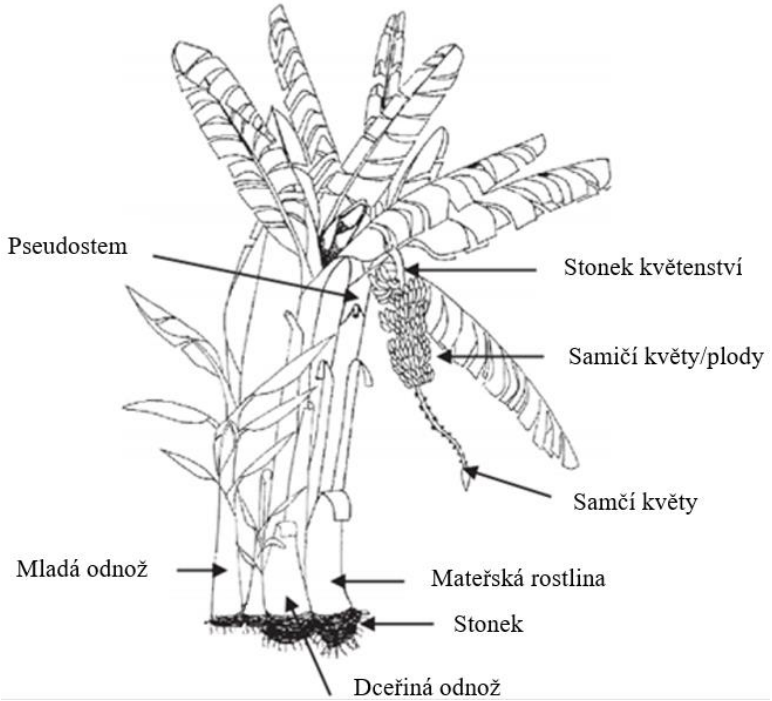
Obr. 1: Fylogenetické vztahy v rámci čeledi Musaceae. Pro konstrukci BioNJ fylogramu bylo využito ITS1-ITS2 nukleotidových sekvencí.

Jak již bylo zmíněno, většina jedlých kultivarů banánovníku vzniklo přirozenou vnitro- nebo mezi-druhovou hybridizací diploidních, planě rostoucích druhů *Musa acuminata* a *Musa balbisiana*, s možným přispěním dalších druhů, především *Musa schizocarpa*. Kombinací genomů A a B vznikly diploidní (AA, BB a AB), triploidní (AAA, AAB a ABB) a tetraploidní (AAAB, AABB, ABBB) formy jedlých banánovníků (Simmonds *et al.*, 1955). Mezi nejvýznamnější jedlé typy banánovníku patří především škrobové banány, např. plantainy (genom AAB), nebo triploidní banánovníky Africké vysočiny (tzv. East African Highland Bananas, genom AAA) a další (Kitavi *et al.*, 2016). Další významnou skupinu představují sladké typy banánů, určené pro vývoz (klon Cavendish, genom AAA) (Perrier *et al.*, 2011).

Banánovník představuje jednu z nejvýznamnějších hospodářských plodin světa. Zároveň ale čelí vážnému ohrožení četnými chorobami a škůdci, což je v případě jedlých klonů, které se množí odnožemi velký problém (Jeger *et al.*, 1995). Šlechtění brání vysoká míra sterility banánů a nedostatečná charakterizace jejich genetické diverzity. Účinnou strategií pro pěstování a výběr rodičů ke křížení vylepšených odrůd banánů je důkladné pochopení genetické rozmanitosti dostupných zdrojů.

2.2 Morfologické vlastnosti

Rostlina banánovníku je tvořena tzv. pseudostemem (nepravý stonek) (Obr. 2), který se skládá ze vzájemně překrývajících se listů, čímž vytváří pevnou válcovitou strukturu (Purseglove, 1972; Stover a Simmonds, 1987) a zajišťuje mechanickou podporu a propojení cévního systému mezi listy, kořeny a plody (Stover a Simmonds, 1987).



Obr. 2: Morfologie banánovníku (*Musa* spp.) (převzato a upraveno z Karamura *et al*., 2011)

Jednotlivé kultivary banánovníku mají odlišnou výšku a barvu pseudostemu. Pseudostemy skupiny *Musa* AB (např. „Kisubi“), AAB (plantainy, „Silk“, „Mysore“ a „Sukali Ndizi“) a ABB („Bluggoes“ a „Pisang Awak“) mají převážně žluto-zelenou barvu s mírnou růžovou pigmentací na spodní části. Barva i výška pseudostemu u banánovníků Africké vysočiny (AAA) se mění v závislosti na okolních podmínkách (Purseglove, 1975). Například klony Cavendish bývají relativně vysoké v nižších polohách, kde jsou pro ně podmínky ideální, ve vyšší nadmořské výšce jsou naopak menší (Stover a Simmonds, 1987).

Další částí rostliny banánovníku jsou postranní odnože, které vyrůstají z pupenů umístěných naproti listovému pouzdru stonku (Obr. 2). Stonek je složen z apikálního meristému, z něhož vyrůstají listy a květy. Rostliny *Musa* mají velkou listovou plochu, u plantainů se pohybuje v rozmezí od 0,68 do 0,92 m^2 (Anojulu, 1992). Jakmile

meristém přestává produkovat listy (asi 30 - 40 listů), prodlouží se internodia, postupně prorůstají středem pseudostemu, dokud se neobjeví květenství (Kamura *et al*., 2015).

Květenství (Obr. 2) je tvořeno silným stonkem, samičími květy, které vyrůstají ve dvou řadách nad sebou a jsou kryty oválnými někdy až tmavě fialovými listeny, které svým tvarem připomínají okvětní plátky. Samčí květy se nachází dále od listů na konci květenství (Purseglove, 1972).

Postupné prodlužování hlavního stonku květenství u některých druhů způsobí, že se celé květenství pod svou váhou prověsí. Listeny se otevřou a následně odpadnou, čímž se odhalí samičí květy (obr. 2). Samičí květy prochází dalším vývojem, aniž by byly opylovány nebo oplodněny. Listeny kolem samčích květů i samotné samčí květy se také otevírají a padají. Velká část stonku tak odděluje samčí pupen od plodů banánovníku. Plody se postupně zaplní, jejich hmotnost ohýbá hlavní stonek a celý trs tak u některých druhů visí svisle dolů. Ačkoli fyziologie, anatomie a vývoj klonů *Musa* jsou podobné, jejich morfologie je různá (Kamura *et al.,* 2015).

2.3 Plantain banánovníky

Plantainy jsou škrobové typy banánovníku, které tvoří velkou skupinu banánů s více než 100 známými kultivary (Swennen, 1990). Genom plantainů (AAB) je triploidní a obsahuje dvě sady chromozomů pocházejících z druhu *M. acuminata* a jednu sadu chromozomů pocházejících z druhu *M. balbisiana*. Přestože veškeré druhy banánů, včetně všech jedlých typů vznikly v jihovýchodní Asii, rozmanitost plantainů je nejvyšší v Africe, zejména v západní a střední Africe. Tato přirozená druhotná genetická rozmanitost je pravděpodobně výsledkem selekce místními zemědělci, kteří vybrali a pěstovali přírodní mutanty odvozené z možná více než jednoho kultivaru původně přivezeného na africký kontinent. Archeologické důkazy ve formě fytolitů naznačují, že banány byly pěstovány v jižním Kamerunu již během prvního tisíciletí před naším letopočtem (Mbida *et al.,* 2000).

Kultivary této podskupiny vykazují širokou škálu morfologických znaků, od rostlin s velkým trsem a samčím pupenem, až po rostliny s pouze několika plody a bez samčích pupenů. K jejich odlišení se používá řada znaků, jako například orientace a počet trsů, velikost a tvar plodů (rovný, rovný v distální části, zakřivený, …), velikost pseudostemu, přítomnost nebo nepřítomnost samčího květenství, barva pseudostemu

(odstíny zelené, červené nebo fialové), barva slupky nezralého ovoce (odstíny zelené, červené, žluté a hnědé).

2.3.1 Morfologické skupiny plantainů

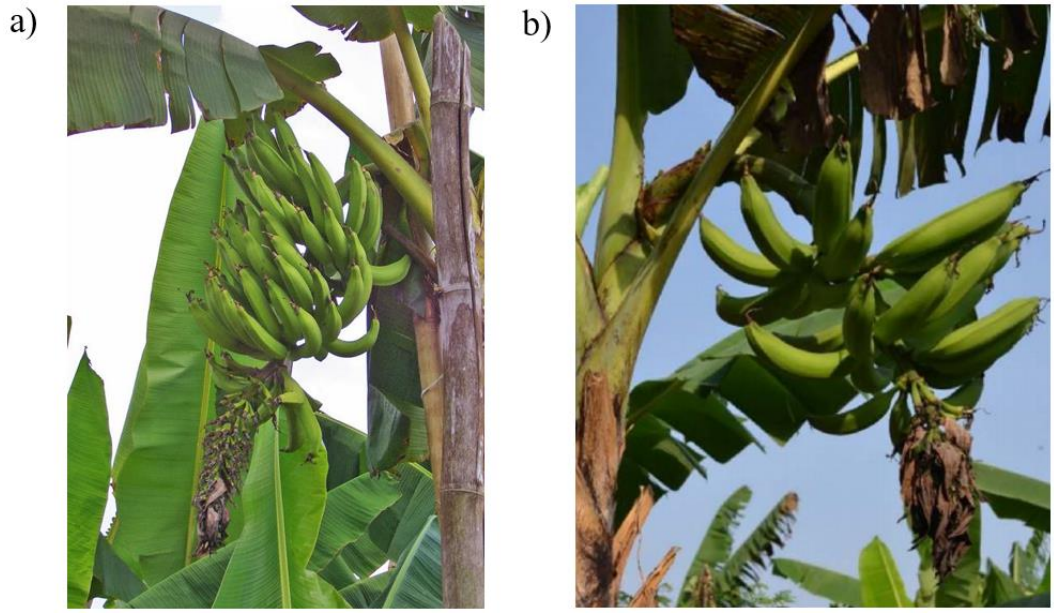
Norman Simmonds původně popsal dva typy kultivarů plantainů, které se vyznačují přítomností samčího květenství (typ French) nebo jeho nepřítomností/degradací (typ Horn) (Simmonds, 1966). Později byly rozpoznány typy False Horn (Tézenas du Montcel *et al*., 1983) a French Horn (Swennen, 1990). Studie o morfologické rozmanitosti 97 zástupců charakterizovaných v Demokratické republice Kongo popsala tři hlavní typy plantainů - French, False Horn a Horn (Adheka *et al.,* 2018), které jsou uznávány jako hlavní morfologické skupiny plantainů i v dnešní době.

Plantainy typu French (Obr. 3) mají na svém stonku stále listeny a velký samčí květenství. Běžně se podle velikosti dělí do tří kategorií - obrovské, střední a malé. Protože výška a obvod pseudostemu se liší v závislosti na okolních podmínkách, k určení velikostní třídy se používá počet listů vyprodukovaných od výsadby do doby vykvetení. Obrovské typy nají více než 40 listů, střední 32 až 38 a malé mají méně než 30 listů (Tézenas du Montcel *et al.,* 1983).

Obr. 3: Plantain typu French (z: https://www.crop-diversity.org/mgis/accession/01USA10898, 25. 3. 2020)

Většina plantainů typu French se vyznačuje velkými banánovými trsy, které obsahují mnoho plodů o relativně malé velikosti. Tyto kultivary mají dlouhý vegetativní cyklus a ve větrných podmínkách jsou náchylné k vyvrácení (Tézenas du Montcel *et al.,* 1983).

Dalšími typy plantainů jsou French Horn (Obr. 4a) a False Horn (Obr. 4b). Samčí květenství u False Horn typů v dospělosti degeneruje, samovolně odpadá a netvoří pyl (Swennen, 1990). Oba typy mají velké plody, ale trsy u French Horn jsou výrazně hustější (Adheka *et al.,* 2018).



Obr. 4: **a)** Plantain typu French Horn (z http://www.promusa.org/Plantain+subgroup#footnote7, 14. 1. 2020) **b)** Plantain typu False Horn (převzato z Adheka *et al*., 2018)

Plantainy typu Horn (Obr. 5) produkují velmi málo plodů rozdělených na jeden až pět trsů (výjimečně osm až deset). Květenství obsahuje pouze samičí květy. Samčí květenství zcela chybí, stonek je ukončen za posledním trsem (Adheka *et al.,* 2018).

Obr. 5: Plantain typu Horn (převzato z Adheka *et al.*, 2018)

2.3.2 Vznik a domestikace plantainů

Vznik a proces domestikace mezidruhových kultivarů banánů je stále předmětem zkoumání. Obecně přijímaná teorie předpokládá vznik všech vnitro i mezidruhových jedlých banánovníků v oblasti JV Asie (Thajsko, Malajsie, Indonésie) v průběhu doby ledové (Holocénu), kdy hladina moří byla podstatně nižší a tyto dnešní poloostrovy a ostrovy byly vzájemně propojeny pevninou (Sand, 1989, Denham, 2004, Denham, 2010, Kagy *et al*., 2016). Takto vzniklé a později selektované jedlé typy byly lidskou migrací v následujícím období přivezeny do Afriky, odkud se díky obchodu s otroky dostali až do střední a jižní Ameriky (D'Hont *et al*., 2012, Häkkinen, 2013, Janssens *et al*., 2016). Jiná teorie vzniku jedlých typů banánovníku dnes pěstovaných na Africkém kontinentu, je založena na transportu rostlin *Musa balbisiana* do Tichomoří, kde došlo k hybridizaci s *Musa acuminata* a vzniklé hybridní typy banánovníků se lidskou migrací rozšířili do západní Afriky (Perrier *et al.*, 2011).

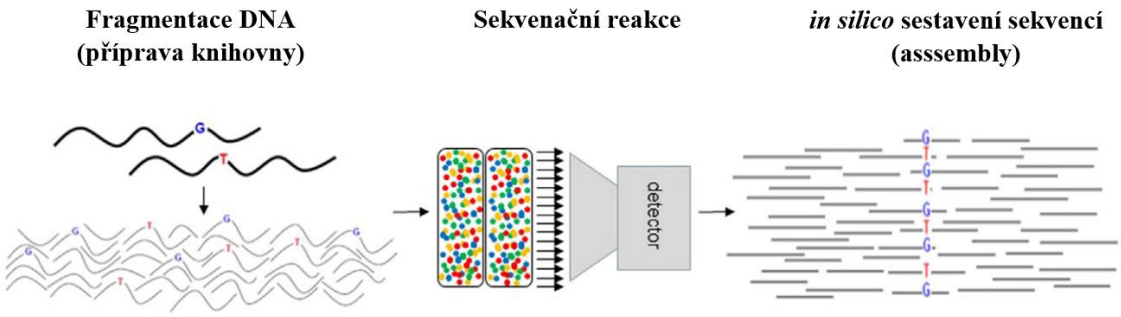
Na základě přenosu mimo-jaderných organel u planých i jedlých zástupců rodu *Musa* byl navržen postup vzniku triploidních plantainů s genomem AAB. V prvním kroku došlo k mezidruhové hybridizaci dvou diploidních planě rostoucích druhů *Musa acuminata* ssp. *banksii* a *Musa balbisiana*, který dal vzniknout mezidruhovému diploidnímu hybridu s genomovým složením AB. Díky neúplné kompatibilitě dvou různých genomů, tento mezidruhový hybrid AB produkoval v meióze neredukované 2n gamety, které byly oplodněny haploidní gametou druhu *Musa acuminatou* ssp. *banksii* (AA), což vedlo k výslednému triploidnímu potomstvu AAB. Tento proces

mohl být o to komplikovanější, že mohl obsahovat několik kol zpětných křížení, které byly již dříve popsány u dalších jedlých kultivarů banánů (De Langhe *et al.,* 2010). Analýzy chromozomového párování v meióze u planých diploidních druhů naznačují, že různé druhy *Musa* a stejně tak i různé poddruhy *Musa acuminata* se liší svou chromozomální strukturou (přítomností rozličných přestaveb), což má za následek narušení meiózy u hybridů a přispívá k jejich ztrátě plodnosti (Dodds a Simmonds 1948; Shepherd, 1999; Jeridi *et al.,* 2012, Martin *et al.,* 2017, Dupouy *et al.,* 2019, Šimoníková *et al.,* 2019, Šimoníková *et al.,* 2020 in press).

Dosud nebyly přesně identifikovány rozdíly na genetické úrovni, ale publikace z roku 2019 (Baurens *et al.,* 2019) naznačuje, že jednotliví mezidruhoví hybridi, včetně plantainů mohou obsahovat rozdílný podíl subgenomů A a B. Studie byla zaměřena na charakterizaci genomové struktury diploidních, triploidních a tetraploidních A/B mezidruhových kultivarů a na studium dopadu těchto genomových struktur vzhledem k segregaci a rekombinaci chromozomů. S využitím sekvenační technologie Illumina a produkci tzv. mate-paired sekvencí (viz níže) byl vyvinut postup pro porovnávání A a B specifických jednonukleotidových polymorfismů (SNP) a vyjádření jejich vzájemného zastoupení v mezidruhových genomech banánovníků (VcfHunter program, Garsmeur *et al.* 2018, Baurens *et al.* 2019).

2.4 Sekvenování nové generace, NGS technologie

Sekvenování nové generace (NGS) je definováno jako technologie umožňující určit v jediném experimentu sekvenci molekuly DNA s celkovou délkou větší než 1 milión párů bází. Důležitou vlastností NGS je možnost sekvenovat stovky/tisíce genů nebo dokonce celý genom v jednom experimentu, čehož je dosaženo masivně paralelním přístupem, kdy centrální část procesu sestává z velkého počtu sekvenačních reakcí prováděných paralelně na fragmentované DNA ve velmi malých objemech, poté je výsledek jednotlivých reakcí čten optickým nebo elektronickým detektorem a posledním krokem je sestavení výstupních produktů (čtení - readů) pro stanovení sekvence molekuly (molekul) DNA před fragmentací (Obr. 6).



Obr. 6: Obecné schéma NGS (převzato a upraveno z Płoski, 2016)

Termín NGS zdůrazňuje nárůst čtení v porovnání s tradičním sekvenováním DNA vyvinutým Sangerem v roce 1975 (Sanger a Coulson, 1975), který má i přes zlepšení, která byla od té doby zavedená, výkon omezený stále na ~ 75 kb (Płoski, 2016).

Důležitým rysem NGS je mnohonásobné sekvenování každé báze cílové sekvence. Kolikrát byla daná pozice v NGS experimentu sekvenována (tj. počet čtení obsahujících tuto pozici), se nazývá pokrytí (coverage). Získání více čtení pokrývajících stejný cíl je nezbytné pro odstranění náhodných chyb během sekvenování (Ebbert *et al.*, 2016, Gargis *et al.*, 2012, Huang *et al.*, 2015).

2.4.1 Příprava NGS knihovny

Kroky potřebné k přípravě DNA pro NGS analýzu se souhrnně nazývají příprava knihovny. Knihovny NGS jsou specifické pro jednotlivé platformy, takže knihovnu připravenou pro jednu platformu nelze použít na jiné platformě, pokud není kompatibilní (obvykle pochází od stejného výrobce). Knihovny NGS mohou být připraveny přímo z cílové DNA (obvykle celkové genomové DNA) nebo z produktů polymerázové řetězové reakce (PCR) (Chen *et al.*, 2018, Head *et al.*, 2014, Seitz *et al.*, 2015). Pro sekvenování je důležité, aby byly molekuly DNA v knihovně opatřeny adaptérovými sekvencemi. Pokud je knihovna připravena pomocí PCR, nejjednodušším způsobem je začlenění adaptérů do PCR primerů tak, aby se staly součástí produktů PCR, které jsou pak připraveny pro sekvenování. PCR se pro přípravu knihovny používá například v případě, kdy je třeba analyzovat pouze málo genů či exonů (Tewhey *et al.*, 2009), respektive při velmi omezeném množství počáteční DNA (od 2 ng). PCR přístup snižuje rozmanitost knihovny a vytváří tzv. duplikáty, tedy více fragmentů, které jsou kopiemi jedné molekuly. Duplikáty snižují kvalitu

sekvenování (Bansal, 2017), mohou chybně naznačovat homozygotnost anebo amplifikovat náhodnou chybu do té míry, že ji lze ve výsledku přijmout jako skutečnou variantu. Řešení těchto problémů poskytují protokoly a soupravy, které umožňují vytvářet knihovny bez PCR (http://www.illumina.com, http://www.biospace.com) (Płoski, 2016).

Pokud příprava knihovny není založena na PCR, ale i v případě využití PCR přístupu, je první fází fragmentace DNA. Dalším krokem je ligace adaptérů (adaptérových sekvencí). Adaptéry jsou sekvence o délce 4 až 10 bp, které slouží jako čárové kódy či indexy. Díky adaptérovým sekvencím je tak možné rozlišit různé vzorky sekvenované dohromady (Li *et al.,* 2019). Obzvláště efektivní je použití dvojitého indexování, obvykle ve strategii používající samostatný adaptér pro každý ze dvou párovaných koncových čtení. Vzorky jsou poté identifikovány kombinací dvou adaptérových sekvencí. Tradiční metody fragmentace jsou založeny na sonikaci, jedná se o technologie Adaptive Focused Acoustics™, která je patentovaná společností Covaris (http://covarisinc.com) nebo Adaptive Cavitation Technology of Diagenode (http://www.diagenode.com/en/index.php). Sonikace ale způsobuje poškození konců molekul DNA, což vyžaduje opravu pomocí enzymů. Pokrokem v přípravě NGS knihoven NGS je enzymová reakce s transpozázou, která současně katalyzuje fragmentaci DNA i začlenění adaptéru/značky/tagu (Adey *et al.,* 2010). Tento proces značení/tagování redukuje množství materiálu potřebného pro konstrukci knihovny, urychluje samotnou přípravu knihovny a umožňuje snadnou automatizaci.

2.4.2 Sestavování sekvence: de novo sekvenování a resekvenování

Důležitým krokem analýzy je sestavení sekvence z krátkých čtení generovaných platformami NGS. Existují dva v zásadě odlišné přístupy: *de novo* sekvenování a resekvenování. *De novo* sekvenování vyžaduje vysoké pokrytí (alespoň 30 x), aby bylo zajištěno dostatečné překrývání jednotlivých čtení pro sestavení cílové sekvence. Tento způsob sestavování sekvence je výpočetně náročný, protože všechna čtení musí být porovnána vůči sobě, aby byla nalezena homologní místa v jednotlivých sekvenačních čteních, ze kterých je následně sestavena dlouhá DNA sekvence, tzv. kontig. Další problém v *de novo* sekvenování způsobují repetitivní oblasti v genomu. Sekvence těchto oblastí je obtížné z krátkých čtení odvodit, a především dobře sestavit do dlouhých kontigů (Lian *et al,* 2014).

Při resekvenování se pro správné skládání sekvenačních čtení do výsledné sekvence-kontigu používá tzv. referenční sekvence. Referenční sekvence je konsenzuální sekvence sloužící jako obecný rámec cílové sekvence s jejími nejrozšířenějšími variantami (Schneeberger *et al.*, 2011, Tsai *et al.*, 2016, Zarrella *et al.*, 2019). Resekvenování se provádí obvykle na nižší pokrytí (~ 10 x). Dnes se tato strategie používá často v kombinaci s tzv. redukcí komplexity genomu, kdy jsou přednostně sekvenovány nemethylované, potenciálně kódující oblasti genomů (např. metody GBS a RADseq nebo další) a používá se k identifikaci specifických jednonukleotidových polymorfismů, vhodných např. pro analýzu genetické variability či identifikaci specifických molekulárních markerů (Huang *et al.*, 2009, Uitdewilligen *et al.*, 2013, Trebbi *et al.*, 2019, Jeffries *et al.*, 2016, Aggeli *et al.*, 2018, Malinsky *et al.*, 2018, Nyine *et al.*, 2018). Jednotlivá sekvenční čtení jsou mapována na referenční sekvenci obvykle s velkou jistotou, to znamená, že pochází z dané části genomu a oblasti, které se po zarovnáni liší (obvykle jednonukleotidové, či velmi krátké polymorfismy), jsou označovány jako varianty.

Ve srovnání s *de novo* sekvenováním vyžaduje resekvenování nižší pokrytí a je výpočetně jednodušší. Resekvenování je efektivní pro detekci variant kratších než délka čtení, jako jsou jednonukleotidové polymorfismy (SNP) nebo krátké inzerce/delece. Naopak detekce větších variant, jako jsou varianty počtu kopií (CNV, délka > 1000 bp) nebo dokonce větších strukturních chromozomálních variant, je náročnější nebo dokonce nemožná, stejně jako detekování variant v repetitivních oblastech (Płoski, 2016).

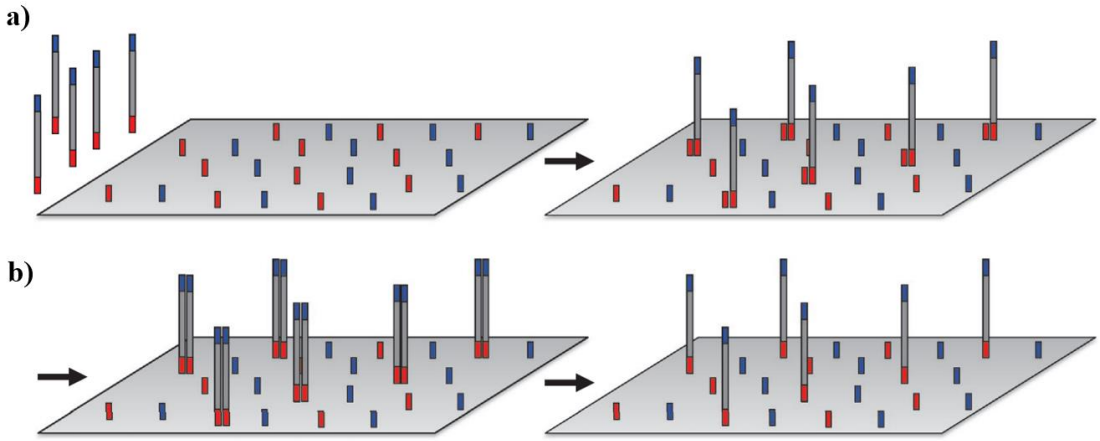
2.4.3 Platformy Illumina

Platformy Illumina jsou založeny na fluorescenčním sekvenování jednotlivých molekul DNA po amplifikaci na pevném nosiči. Tento přístup vyvinula v roce 2006 společnost Solexa, kterou následně získala Illumina.

Příprava knihovny pro platformy Illumina původně zahrnovala mechanickou fragmentaci DNA, enzymatickou opravu konců, přidání jediné adeninové báze na 3´ konec fragmentů DNA a ligaci adaptérů. Jiný postup přípravy sekvenačních knihoven je založen na značení katalyzovaném transpozázou (Adey *et al.*, 2010). Adaptéry Illumina obsahují oblasti pro navázaní primerů a takzvané P5 a P7 vazebné oblasti, které jsou komplementární k oligonukleotidům navázaným na povrch

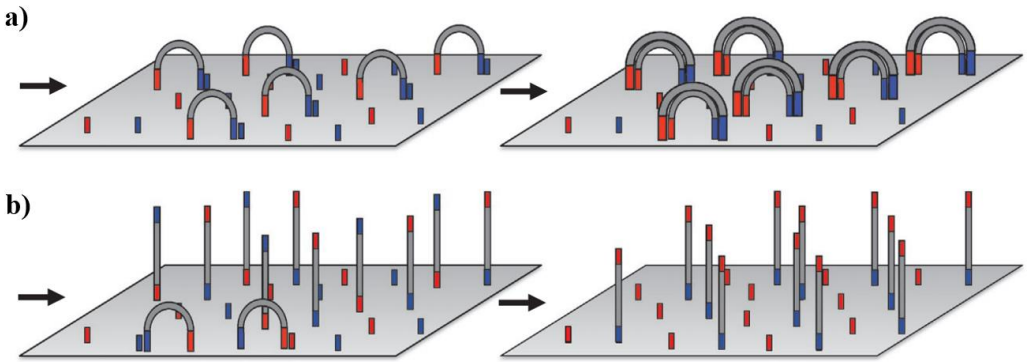
průtokové komůrky (flow cell), která umožňuje přístup k činidlům a optickému zobrazení (Lebl *et al.*, 2012).

Předpokladem pro sekvenování je hybridizace knihovny jednořetězcové DNA na oligonukleotidy navázané na průtokovou komůrku (Obr. 7a). Tato relativně slabá nekovalentní vazba je následně převedena na silné kovalentní vazby syntézou komplementárního (reverzního) řetězce. Následuje promývání a původně navázaný řetězec DNA (templát) je odstraněn (Obr. 7b) (Płoski, 2016).



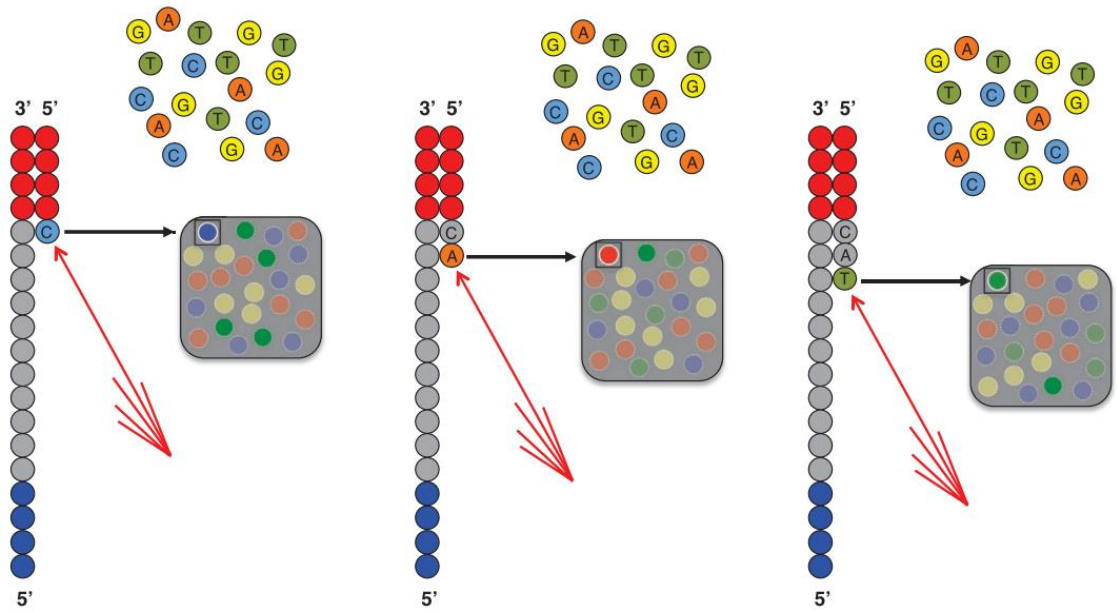
Obr. 7: a) navázání DNA knihovny, b) syntéza reverzního řetězce (převzato z Vilgis a Deigner, 2018)

Dalším krokem je můstková amplifikace, což je cyklický proces, při kterém se replikují molekuly DNA navázané na průtokovou komůrku a vytváří se takzvané klastry. Během můstkové amplifikace se jednořetězcová molekula převrátí a vytvoří můstek navázáním na sousední komplementární primer. Po syntéze polymerázou se vytvoří dvouvláknový můstek, který po denaturaci poskytne reverzní kopii původního fragmentu DNA (forward) kovalentně navázaného na povrch průtokové komůrky (Obr. 8a). Proces se cyklicky opakuje a v posledním kroku se reverzní řetězce odštěpí a zůstanou homogenní klastry přibližně s 1000 forward vlákny (Obr. 8b) (Illuminna, 2010, Płoski, 2016, Voelkerding *et al.*, 2009, Buermans a den Dunnen, 2014).



Obr. 8: a) můstková amplifikace b) odstranění reverzních řetězců (převzato z Vilgis a Deigner, 2018)

Vlastní sekvenování DNA pomocí platforem Illumina se provádí metodou sekvenování pomocí syntézy (sequencing-by-synthesis, SBS) (Obr. 9). Reakce začíná hybridizací primeru na k němu komplementární sekvenci adaptéru. Následně se opakují kroky: (1) přidání DNA polymerázy se čtyřmi fluorescenčně značenými nukleotidy, (2) zobrazování, zaznamenání fluorescence a (3) odštěpení fluorescenčních barviček a terminátoru. Nukleotidy jsou reverzibilně blokovány (zakončeny) a jednotlivě značeny fluorescenčně, což zajišťuje, že během každého cyklu je primer rozšířen pouze o jednu bázi a že tato báze může být identifikována a skenována pomocí fluorescence (Illumina, 2010, Buermans a den Dunnen, 2014).



Obr. 9: Sekvenování pomocí syntézy (SBS) (převzato z Vilgis a Deigner, 2018)

V závislosti na aplikaci a konkrétní platformě lze provádět 36 až 301 cyklů, což umožňuje sekvence dlouhé 35 až 300 bp.

Všechny platformy Illumina podporují párové-koncové sekvenování (pair-end sequencing). Sekvenování druhého konce molekuly DNA následuje poté, co je osekvenován první (forward) řetězec. Čtení (read 1) z forward řetězce je odděleno a je provedena můstková amplifikace. Následně je forward řetězec z průtokové komůrky odštěpen a je osekvenován reverzní řetězec, opět metodou SBS (Płoski, 2016, Vilgis a Deigner, 2018).

2.4.3.1 Illumina zařízení

První platformou byl Genome Analyzer (GA), původně nabízený společností Solexa, kterou získala společnost Illumina v roce 2007. Přestože je GA stále používána, je do značné míry nahrazena nástroji HiSeq (HiSeq 1000 a 2000), anebo jejich hardwarově upgradovanými verzemi (HiSeq 1500, HiSeq 2500 a HiSeq 3000, HiSeq 4000), stejně jako MiSeq a MiSeqDx. Důležitým vylepšením HiSeq 3000/4000 je vzorovaná průtoková komůrka s nanočásticemi, která řídí tvorbu klastrů, což zajišťuje jejich optimální hustotu. Stroje MiSeq patří do kategorie stolních sekvenátorů a MiSeqDx je zaměřen na klinické aplikace. Porovnání platforem NGS Illumina je uvedeno v tabulce 1.

Tab. 1: Porovnání Illumina platforem (převzato a upraveno z Płoski, 2016)

	Max. velikost výstupu (Gb)	**Max. počet čtení (M, párová čtení)**	**Max. délka čtení (bp)**
Genome Analyzer IIx	95	300	2 x 150
HiSeq 2500	1000	4000	2 x 125
HiSeq 2500 rapid mode/single flow cell	90	300	2 x 150
HiSeq 3000/4000	750/1500	2500/5000	2 x 150
NextSeq 500	120	400	2 x 150
MiSeq	15	25	2 x 300

Další platformou firmy Illumina je NovaSeq 6000 System, který využívá osvědčenou technologii NGS, poskytuje více typů průtokových komůrek (SP, S1, S2, S4) (Tab. 2) a umožňuje kombinovat různé délky čtení, díky čemuž představuje jednoduchou, škálovatelnou a spolehlivou vysoce výkonnou sekvenační platformu

Illumina, která poskytuje vynikající kvalitu dat (výstup o velikosti až 6 Tb) za méně než 2 dny (Illumina, 2019).

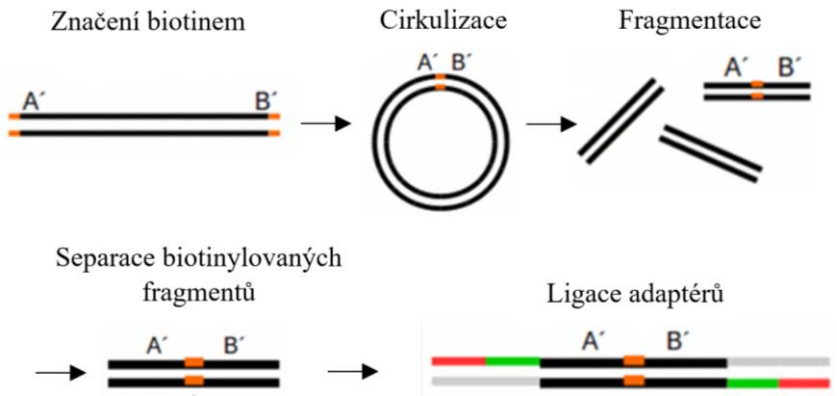
Tab. 2: Specifikace průtokových komůrek NovaSeq 600 (převzato a upraveno z Illumina, 2019)

Velikost výstupu průtokových komůrek	**SP**	**S1**	**S2**	**S4**
2 x 50 bp	65-80 Gb	134-167 Gb	333-417 Gb	/
2 x 100 bp	/	266-333 Gb	667-833 Gb	1600-2000 Gb
2 x 150 bp	200-250 Gb	400-500 Gb	1000-1250 Gb	2400-3000 Gb
2 x 250 bp	325-400 Gb	/	/	/

2.4.4 Nevýhody metod NGS

Technologie NGS mají určité nevýhody. Jedním z hlavních omezení jsou jejich relativně krátká čtení. Genomy často obsahují četné repetitivní sekvence, které jsou delší než čtení NGS, a to může vést k nesprávně sestaveným sekvencím a mezerám ve výsledných kontizích (Goodwin *et al.,* 2016, Salzberg a Yorke, 2005). Mimo jiné se pomocí krátkých čtení hůře detekují delší strukturní varianty (Weischenfeldt *et al.,* 2013). Technologie NGS mají také omezenou schopnost charakterizovat transkripční izoformy generované alternativním sestřihem.

Illumina zmírnila problémy spojené s krátkými délkami čtení vytvářením tzv. mate-pair knihoven. Jedná se o přístup, kdy jsou sekvenovány fragmenty nacházející se daleko od sebe (až do 25 kb) (Korbel *et al.,* 2007). Nejprve je provedena fragmentace, jejíž výsledkem jsou přiměřeně dlouhé fragmenty DNA. Konce DNA fragmentů jsou značeny biotynem a molekuly jsou cirkularizovány (obr. 8). Cirkularizovaná DNA je poté fragmentována na malé kousky (300–500 bp) a biotinylované fragmenty jsou shromážděny a purifikovány afinitním zachycením pomocí magnetických kuliček potažených streptavidinem. V dalším kroku jsou na konce zachycených fragmentů naligovány sekvenční adaptéry (Obr. 10) a je provedeno párové-koncové sekvenování (Gao a Smith, 2015).



Obr. 10: Příprava mate-pair knihovny (převzato a upraveno z: https://www.ecseq.com/support/ngs/what-is-mate-pair-sequencing-useful-for)

Mate-pair sekvenování tak do jisté míry umožňuje překonat omezení NGS spojená s dlouhými repetitivními úseky.

Jako alternativní přístup představila společnost Illumina v roce 2014 sadu pro přípravu knihoven pro syntetické dlouhé čtení (synthetic long reads, SLR) (Voskoboynik *et al.,* 2013). Přístup je však pracný, nákladný a poskytuje stále relativně malé délky čtení. Bylo proto potřeba vyvinout metody, které by lépe zvládly výše uvedené problémy.

2.5 Sekvenování třetí generace (Long-Read Sequencing)

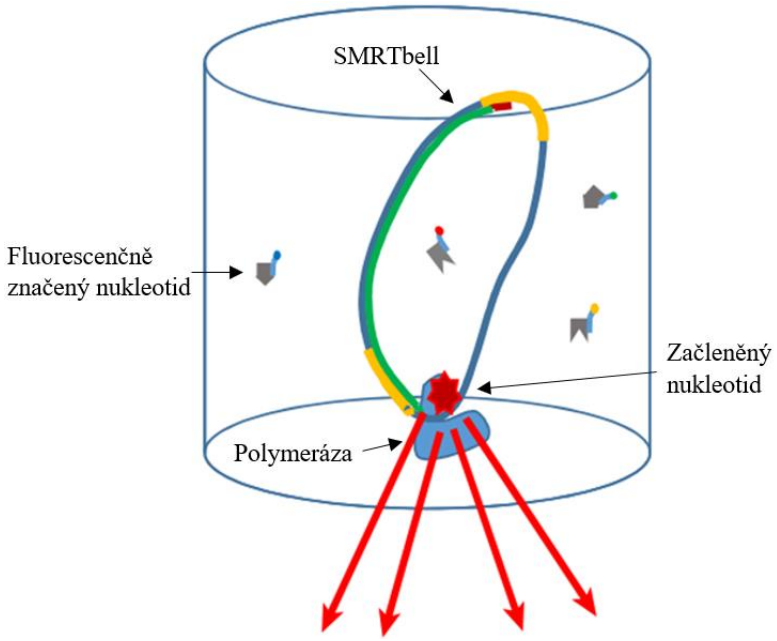
Krátce po vzniku NGS se objevily technologie sekvenování třetí generace (TGS). Charakteristickým rysem TGS je sekvenování jedné molekulou (single-molecule sequencing, SMS) a sekvenování v reálném čase (na rozdíl od NGS, kde je sekvenování pozastaveno po každém začlenění báze) (Schadt *et al.,* 2010). První technologie SMS, komercializovaná společností Helicos Biosciences, připomínala sekvenování Illumina, ale bez můstkové amplifikace (Puskarev *et al.,* 2009). Metoda byla ale relativně pomalá, nákladná a produkovala krátká čtení (32 bp). První skutečná technologie TGS byla uvedena na trh v roce 2011 společností Pacific Biosciences (PacBio) a je označována jako „sekvenování jedné molekuly v reálném čase“ (single-molecule real-time, SMRT) (Eid *et al.,* 2009). V roce 2014 zavedla společnost Oxford Nanopore Technologies (ONT) sekvenování pomocí nanoporů (Jain *et al.,* 2015). Nejen, že SMRT a nanopore nevyužívají proces amplifikace, ale produkují dlouhá čtení.

2.5.1 Pacific Biosciences, SMRT sekvenování

Technologie, kterou využívá PacBio se nazývá SMRT sekvenování (Puskarev *et al.,* 2009). Tato technologie používá uzavřený kruhový templát dsDNA zvaný SMRTbell (Obr. 11), který je vytvořen ligací vlásenkových adaptérů na konce cílové molekuly dsDNA (Voskoboynik *et al.,* 2013).

Obr. 11: Kruhová templátová DNA – SMRTbell (převzato z Dijk *et al.,* 2018)

Platforma používá průtokovou komůrku se 150 000 pikolitrovými jamkami. V každé jamce je jediná polymeráza imobilizována na dně, kde replikuje cílovou molekulu DNA. Během replikačního procesu dochází při začlenění fluorescenčně značených nukleotidů k uvolnění fluorescenčního signálu (Obr. 12) a tyto záblesky jsou zaznamenány kamerovým systémem v reálném čase.



Obr. 12: replikační proces sekvenování SMRT (převzato a upraveno z Dijk *et al.,* 2018)

Protože SMRTbell tvoří uzavřený kruh, poté, co polymeráza replikuje jedno vlákno cílové dsDNA, může pokračovat užitím adaptéru a poté druhého vlákna jako templátu. Pokud je životnost polymerázy dostatečně dlouhá, mohou být oba řetězce sekvenovány vícekrát za vzniku jediného nepřetržitého dlouhého čtení,

které lze poté rozdělit na více částí rozpoznáním a vyříznutím adaptérových sekvencí (Dijk *et al.,* 2018).

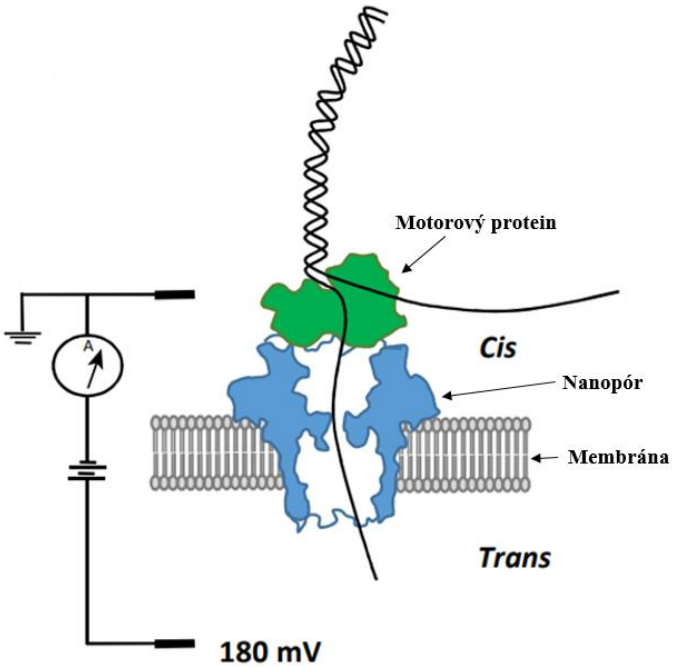
SMRT sekvenování má i nevýhody. Jako například velké množství výchozího materiálu potřebného pro přípravu knihovny a vysoká míra chyb při jednom průchodu (15 %) způsobená rychlostí záblesků a kamera proto není schopna záblesky fluorescenčního signálu zaznamenat (Dijk *et al.,* 2018).

2.5.2 Oxford Nanopore Technology

Myšlenka použití nanopórů k sekvenování jednořetězcových molekul DNA nebo RNA vznikla na konci 80. let (Deamer *et al.,* 2016). Avšak z důvodu technických překážek byly první úspěšné výsledky sekvenování hlášeny až v roce 2012 (Manrao *et al.,* 2012). O dva roky později společnost Oxford Nanopore vydala svůj první nanopórový sekvenátor MinION.

Průtokové komůrky v kombinaci s nejnovějšími verzemi kitů pro přípravu knihovny mohou produkovat až 20 Gb sekvenčních dat. Také se zvýšila rychlost translokace, od 30 do 450 bp za sekundu (Brown a Clarke, 2016).

K sekvenování dochází v průtokové komůrce, která umožňuje masivní paralelní sekvenování. Průtoková komůrka má dvě komory (*cis* a *trans*), naplněné iontovými roztoky. Komory jsou vzájemně odděleny elektricky odolnou membránou, která obsahuje jednotlivé nanopóry (MinION obsahuje až 2 000 pórů, PromethION 12 000) (Dijk *et al.,* 2018). Na nanopóry jsou navázány helikázy, sloužící jako motorové proteiny, které vážou jednořetězcovou DNA (ssDNA) a řídí translokaci DNA přes nanopór (Wu a Spies, 2013) (Obr. 13). Jakmile DNA projde nanopórem, motorový protein se oddělí a nanopór je připraven přijmout další fragment (Dijk *et al.,* 2018).



Obr. 13: Průtoková komůrka Oxford Nanopore (převzato a upraveno z Dijk *et al.*, 2018)

Během sekvenování jsou napóry pod stálým elektrickým proudem. Proud je závislý na velikosti vstupu do póru (pokud se v póru něco nachází, aktuální proud se změní). Sledují se změny elektrického proudu během postupného procházení DNA nanopórem. Pro každý nukleotid je pór otevřen jinak a pro každý nukleotid se tak mění i elektrický proud. Výsledný signál je poté dekódován a je poskytnuta výsledná sekvence DNA.

Sekvenování Oxford Nanopore vyžaduje přípravu knihovny, ve které jsou konce fragmentů DNA opraveny a následně jsou k nim naligovány adaptéry. Adaptéry jsou komplexy DNA-protein s pevně vázanou polymerázou a enzymem helikázou, který zajišťuje postupný pohyb DNA skrz pór pomocí západkového mechanismu. Pro zvýšenou přesnost čtení byly vyvinuty mechanismy pro sekvenování druhého vlákna poté, co první vlákno prošlo pórem. Za tímto účelem byl vyvinut systém $1D^2$ (Dijk *et al.*, 2018).

Na rozdíl od SMRT sekvenování není délka čtení Oxford Nanopore omezena samotnou technologií, ale spíše délkou molekul DNA, které mají být sekvenovány. Proto za předpokladu, že DNA je dostatečně kvalitní, lze získat extrémně dlouhé produkty (až 1Mb) (Jain *et al.*, 2018).

Nevýhodou Oxford Nanopore je vysoká míra chyb (15 %) (Jain *et al.*, 2017), pro kvalitnější sestavu sekvence se tak využívá kombinace s Illuminou. Oproti SMRT

sekvenování nemá Oxford Nanopore možnost sekvenovat stejný řetězec vícekrát a pro vyšší přesnost tak vyvinuli metodu, která podporuje průchod druhého vlákna do pórů poté, co prošlo první vlákno. Zdvojnásobuje se tak ale doba průchodu molekuly pórem. Další nevýhodou jsou časté změny softwarových verzí, průtokových komůrek a kitů (Dijk *et al.,* 2018).

2.6 In silico analýza NGS dat

Moderní technologie posunuly studium biologie rostlin na vyšší úroveň, kvůli rychlému nárůstu sekvenovaných genomů mnoha druhů rostlin (Schuster, 2008, Govindaraj *et al.,* 2015). Je tak vidět obrovský dopad výzkumu rostlinných genomů na zlepšení ekonomicky významných rostlin a znalost biologie rostlin (Feuillet *et al.,* 2011).

Sekvenování genomu bylo revolucionizováno obzvláště technologiemi Sekvenování nové generace, které rychle produkují obrovské množství dat při relativně nízkých nákladech a je obrovskou výzvou analyzovat toto velké množství dat a interpretovat biologický význam (Lee *et al.,* 2011).

Vývoj NGS sekvenování DNA tak dramaticky změnil proces objevování jednoduchých nukleotidových polymorfismů a malých inzercí a delecí z nukleotidových sekvenačních dat a nahradilo tak starší a nákladnější metody zahrnující subklonování nebo amplifikaci polymerázovou řetězovou reakcí (PCR) s následným Sanger sekvenováním. Výsledkem je bezprecedentní charakterizace genetických variant v nemocných i zdravých tkáních všech organismů (The 1000 Genomes Project Consortium, 2012, Hoadley *et al.,* 2014).

Jednonukleotidové polymorfismy (SNP) jsou definovány jako změny jedné báze ve specifické poloze nukleotidů a jsou široce distribuovány v genomech, jak v kódujících, tak nekódujících oblastech všech organismů. Jedná se o nejrozšířenější typ variant genů a mohou být odpovědné za specifické rysy nebo fenotypy, a mohou také poskytnout informace o evoluční historie druhů. Jsou vhodnými molekulárními markery (McCouch *et al.,* 2010, Guajardo *et al.,* 2020, Agarwal *et al.,* 2008, Trebbi *et al.,* 2019, Wondji *et al.,* 2007, Leache a Oaks, 2017).

Velký počet SNP lze u druhů identifikovat pomocí vysoce výkonných technologií NGS, jako je sekvenování DNA asociované s restrikčním místem (RAD-seq) (Baird *et al.,* 2008), metodou genotypování pomocí sekvenování (GBS) (Elshire *et al.,* 2011) a sekvenováním amplifikovaných fragmentů specifických lokusů (SLAF-

seq) (Sun *et al.,* 2013). SNP markery se tedy používají pro hodnocení genetické rozmanitosti, studii molekulární evoluce a genetického mapování u plodin (Agarwal *et al.,* 2008).

Pro *de novo* identifikaci variant ze sekvenačních dat byly vyvinuty bioinformatické nástroje a postupy obvykle sestávají ze tří částí: úprava čtení (read processing) odstraněním bází s nízkou kvalitou, dále mapování a zarovnání (mapping and alignment) čtení na referenční sekvenci pomocí softwarů jako jsou například Bowtie (Langmead *et al.,* 2009), BWA (Li *et al.,* 2009) a MAPQ (Li *et al.,* 2008) a posledním krokem je určování variant (variant calling) (Xu, 2018).

Určování variant je klíčovým krokem analýzy NGS. Kvalita vstupních dat přímo ovlivňuje následnou analýzu, jako je detekce specifických genů. Aby bylo dosaženo vysoce kvalitního určení variant a genotypizace, jsou sekvenační čtení nejprve mapována na referenční genom/sekvenci a tyto výsledné soubory se poté aplikují programy sloužící k určení variant (Li *et al.,* 2018). Dobré výsledky poskytují například SAMtools (Li, 2009), Genome Analysis Toolkit (GATK) (McKenna *et al.,* 2010) a Platypus (Rimmer *et al.,* 2014).

Algoritmy sloužící k určování variant mají za cíl řešit technické potíže, jako jsou homopolymerní chyby, náhodné mutace, inzerce a delece (indely), chybné namapování a zkreslené PCR výsledky. Obecně existují dva přístupy (Albers *et al.,* 2011). První přístup využívá bayesovské metody pro modelování sekvenačních chyb a identifikaci variant. Kandidátní varianty jsou tak úspěšně generovány pomocí bayesovských metod přímo z výsledků nezávislého mapování každého čtení na referenční sekvenci. Tento přístup je velmi účinný pro detekci SNP, ale může vykazovat chyby, pokud jsou čtení namapována do oblastí vedle kandidátních indelů (Li *et al,* 2018). Na tomto principu pracují například softwary SAMtools a GATK-UnifiedGenotyper (DePristo *et al.,* 2011).

Druhý přístup je založený na sestavování sekvencí. Ve výpočtu nejprve provádí *de novo* sestavení krátkých čtení v rámci oken dané fixní délky pro vytvoření kandidátních haplotypů. Následně je provedeno pravděpodobnostní srovnání výsledných haplotypů s referenční sekvencí. Kandidátní haplotyp, který má nejvyšší pravděpodobnostní hodnoty, je považován za skutečnou sekvenci daného okna a je v něm provedeno určení/volání variant. Tento přístup může řešit chybné namapování okolních indelů a také identifikovat velké indely. Přístup je přesnější, ale vzhledem k extrémně vysoké výpočetní složitosti a velkému počtu kandidátních

haplotypů, vyžaduje v porovnání s prvním přístupem mnohem delší dobu pro výpočet (Li *et al,* 2018). Na druhém přístupu je založen GATK-HaplotypeCaller (Poplin *et al.,* 2017).

Existuje další metoda, která kombinuje oba přístupy a je považována za bayesovskou haplotypovou metodu (software Platypus).

Analýza NGS dat zahrnuje také detekci strukturních variant (SV), do kterých se řadí dlouhé inzerce, duplikace, delece, inverze a translokace. S podstatným pokrokem v technologiích sekvenování a analytických strategiích byla definice SV rozšířena tak, aby zahrnovala varianty dlouhé až 50 bp (Mills *et al.,* 2011, Sudmant *et al.,* 2015, Zarrei *et al.,* 2015).

Obecně lze způsob detekce SV klasifikovat do čtyř různých algoritmů: Read-Pair (RP), Split-Read (SR), Read-Depth (RD) a Assembly (AS). Některé nástroje kombinují více než jeden algoritmus, aby se zvýšila specifičnost a citlivost (Ye *et al.,* 2016).

Při párovém-koncovém sekvenování se očekává, že DNA fragmenty budou mít specifickou velikost inzertu (Korbel *et al.,* 2007). Metody Read-Pair jsou založeny na identifikaci namapovaných párových čtení na referenční genom, jejichž vzdálenosti jsou výrazně odlišné od předem stanovené průměrné velikosti inzertu (Pirooznia *et al.,* 2015). Tato čtení mohou být namapována dále nebo blíže od sebe, mohou být v obrácené orientaci nebo v nesprávném pořadí anebo jsou mapovány na různé chromozomy. Každá z těchto možností je indikátorem různých typů strukturních variant (Escaramís *et al.,* 2015). Mezi nástroje používající metodu RP, patří například PEMer (Korbel *et al.,* 2009) a BreakDancer (Chen *et al.,* 2009).

Metoda Split-Read využívá čtení generovaná párovým-koncovým sekvenováním, kde pouze jedno z dvojice párových čtení je spolehlivě namapováno na referenční genom a druhé se buď zcela nebo částečně nedokáže na referenci namapovat (Zhang *et al.,* 2011). Nemapovaná čtení jsou potenciálním zdrojem zlomových bodů na úrovni páru bází. Mapovaná čtení, která se rozprostírají přes zlomový bod strukturních variací, poskytují přesné počáteční a koncové polohy delecí či inzertů. Nástroje používající Split-Read metodu, jako například Pindel (Ye *et al.,* 2009) a Prism (Jiang *et al.,* 2012), jsou schopny tyto zlomové body identifikovat, ale mají omezenou schopnost identifikovat velké strukturní varianty. Software Prism toto omezení podstatně překonává, použitím modifikovaného Needleman-Wunsch algoritmu (Jiang *et al.,* 2012).

Read-Depth metody jsou založeny na hypotéze, že existuje korelace mezi hloubkou pokrytí a počtem kopií v dané oblasti genomu (Teo *et al.,* 2012). RD metody detekují přesný počet variant počtu kopií (CNV), na rozdíl od RP a SR metod, které detekují pouze jejich polohu. Prvním krokem detekce pomocí metody RD je mapování čtení na referenční genom a následný výpočet hloubky pokrytí využitím předdefinového okna. Hloubka čtení v každém okně je poté normalizována, aby se odstranil potenciální šum kvůli GC a repetitivním oblastem (Boeva *et al.,* 2011; Janevski *et al.,* 2012). Následně je provedena predikce statistické významnosti určení CNV variant a v posledním kroku je použito filtrování (Janevski *et al.,* 2012; Zhao *et al.,* 2013). Nástroje, které tuto metodu využívají, patří například CNV-seq (Xie a Tammi, 2009) a BIC-seq (Xi *et al.,* 2011).

Teoreticky mohou být všechny formy genetických variant včetně CNV detekovány pomocí sestavení celogenomové sekvence (sequence assembly) z krátkých čtení (pokud jsou čtení dostatečně dlouhá a přesná). Metody assembly nejprve vygenerují kontigy a scaffoldy, které se poté porovnávají s referenčním genomem, aby se objevily strukturní varianty (Nijkamp *et al.,* 2012; Teo *et al.,* 2012). Eukaryotní genomy ale obsahují významný podíl repetic a duplikací, což způsobuje, že jsou metody AS méně přesné a složitější, neboť v těchto složitých oblastech vykazují špatnou výkonnost (Xi *et al.,* 2012).

3 EXPERIMENTÁLNÍ ČÁST

3.1 Sekvenační data

Izolace genomové DNA, příprava sekvenačních knihoven, a samotné sekvenování platformou Illumina (NovaSeq 6000 System), které vedlo k získání párových 150-nt dlouhých sekvenačních čtení bylo provedeno kolegy z Centra strukturní a funkční genomiky rostlin, Ústavu experimentální botaniky AV ČR, v. v. i. a CR Haná v Olomouci.

Jako referenční genomová sekvence banánovníku byla použita celogenomová sekvence dihaploidního druhu *M. acuminata* ssp. *malaccensis,* klon DH Pahang (verze 2; Martin *et al.*, 2016).

Všechny bioinformatické programy využité v diplomové práci byly spuštěny na serverech Centra strukturní a funkční genomiky rostlin, které jsou zároveň součástí MetaCentra (Virtuální organizace MetaCentrum VO).

V rámci diplomové práce bylo analyzováno deset zástupců afrických plantainů (4 zástupci typu French, 4 zástupci typu False horn, a 2 zástupci typu Horn) (Tab. 3) a dva diploidní plané druhy – pravděpodobné rodičovské genomy *M. acuminata* ssp. *banksii* ITC0904 (donor subgenomu A) a *M. balbisiana* ITC0248 (donor subgenomu B).

Tab. 3: Deset analyzovaných zástupců afrických plantainů

Číslo vzorku	Název kultivaru	Typ	Země původu
7	ITC.0109; Obino l'Ewai	French	Nigérie
9	ITC.0142; Msisa	French	Burundi
12	ITC.0219; Apem Pa	French	Ghana
16	ITC.0325; Wine Plantain	French	Honduras
30	ITC.0098; Baka	False horn	Gabon
33	ITC.0223; Apantu	False horn	Ghana
38	ITC.0517; Orishele	False horn	Nigérie
42	ITC.0641; Dominico Rojo	False horn	Kolumbie
46	ITC.0185; 3 Hands Planty	Horn	Kamerun
48	ITC.0128; Tshambunu	Horn	Burundi

3.2 Bioinformatická analýza

Před úpravou sekvenačních dat, byla provedena kontrola kvality sekvencí programem FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sekvenační čtení byla na základě informací poskytnutých FastQC dále zpracována softwarem Trimmomatic v0.32 (Bolger *et al.*, 2014). Software zajistil, že byla ponechána pouze čtení o zadané délce (150-nt; viz níže) a bez přítomnosti sekvenačních adaptérů.

Krok byl spuštěn příkazovým řádkem:

```
java -jar trimmomatic-0.39.jar PE -threads 2 -phred33 vstup_R1.fastq vstup_R2.fastq -baseout výstupní_soubory ILLUMINACLIP: TruSeq3-PE.fa:2:30:10 SLIDINGWINDOW:4:15 MINLEN:150,
```

kde se pro vstupní párová čtení `vstup_R1.fastq` a `vstup_R2.fastq` provedlo následující:

1) Odstranění adaptérů (`ILLUMINACLIP: TruSeq3-PE.fa:2:30:10`).
2) Odstranění bází, jejichž průměrná kvalita v okně klesla pod prahovou hodnotu (`SLIDINGWINDOW:4:15`).
3) Odstranění čtení kratších než zadaná délka (`MINLEN:150`).

Z výstupních dat upravených softwarem Trimmomatic byla následně odfiltrována čtení s kvalitou nižší než prahová hodnota pomocí FASTQ Quality Filter, který je součástí FASTX-Toolkit programu (http://hannonlab.cshl.edu/fastx_toolkit/).

Spuštění příkazového řádku ve tvaru:

```
fastq_quality_filter -q 20 -p 90 -i vstupní_soubor.fastq -o výstupní_soubor_Filtr.fastq
```

zajistilo, že ze vstupních souborů (parametr `-i`) byly odstraněny sekvence s kvalitou nižší, než zadaná hodnota parametru `-q` a zároveň hodnota parametru `-p` zajistila, jaké bude minimální procento bází, které musí mít danou kvalitu.

Odfiltrování nekvalitních sekvencí vygenerovalo nesynchronizované výstupní soubory, které bylo potřeba znovu spárovat, protože pro další analýzu se používala pouze párová čtení. Tento krok byl zajištěn využitím softwaru Pairfq (https://github.com/sestaton/Pairfq, version 0.17.0).

Příkazový řádek byl ve tvaru:

`pairfq makepairs -f vstup_R1.fastq -r vstup_R2.fastq -fp výstup_R1_P.fastq -rp výstup_R2_P.fastq -fs výstup_R1_S.fastq -rs výstup_R2_S.fastq –stats`,

kde:

`makepairs` je poziční argument pro párová čtení,

`-f/-r` značí vstupní soubory (forward/reverse),

`-fp/-rp` značí výstupní párová čtení (forward/reverse),

`-fs/-rs` značí výstupní nepárová čtení (forward/reverse),

`--stats` je argument poskytující informace o výsledcích párování.

Párová čtení byla použita pro následnou analýzu pomocí programu VcfHunter (https://github.com/SouthGreenPlatform/vcfHunter), který slučuje několik programů a umožňuje mapování DNA/RNA sekvenačních čtení na referenční genom, k identifikaci jednonukleotidových polymorfismů (SNP) a manipulaci s výslednými vcf soubory, které jsou dále využity pro vizualizaci pozic detekovaných SNP podél referenční genomové sekvence, v našem případě podél jednotlivých chromozomů banánovníku.

První fáze analýzy byla spuštěna příkazem:

`python3 process_reseq_1.0.py -c DNAseq.conf -t 5 -p Plantains -s abcefgh`, kde:

`-c/--conf` je konfigurační soubor obsahující cesty k referenčnímu genomu a párovým čtením,

`-t/--threads` udává počet jader,

`-p/--prefix` je název prefixu pro výstupní vcf soubor a soubor obsahující statistické informace,

`-s/--steps` značí řetězec obsahující kroky k provedení:

`a`: mapování sekvencí DNA na referenční genomovou sekvenci algoritmem BW-MEM, který je součástí softwarového balíčku Burrows-Wheeler Aligner (Li a Durbin, 2010), odstranění nenamapovaných čtení softwarem SAMtools (https://github.com/samtools/samtools),

`b`: identifikace a odstranění duplicitních čtení nástrojem PicardTools (https://broadinstitute.github.io/picard/command-line-overview.html#MarkDuplicates),

`c`: opětovné lokální zarovnání kolem oblastí indelů nástrojem GATK (https://github.com/broadinstitute/gatk/),

`e`: výpočet alel pro každou pozici ve všech vzorcích napříč všemi chromozomy pomocí bam-readcount (https://github.com/genome/bam-readcount),

`f`: určování genotypu (GATK),

`g`: sloučení souborů generovaných krokem `f`,

`h`: statistika mapování.

V rámci analýzy byl tedy zpracován multifasta soubor referenčních sekvencí a několik fastq souborů obsahujících illumina sekvenační čtení plantainů, a pro každého zástupce (tj. zástupci jednotlivých morfologických skupin afrických plantainů) byl vrácen bam soubor a konečný vcf soubor obsahující informace o alelách na každé variabilní pozici, mající alespoň jednu variantní alelu obsaženou alespoň v jednom čtení. Protože genotypy nalezené ve výstupu vcf nemusí odrážet správný genotyp, byl dále využit program VcfPreFilter, který provádí určování variant na základě specifikace uživatele.

Krok byl spuštěn příkazovým řádkem:

```
python3 VcfPreFilter.1.0.py -v Plantains_all_allele_count.vcf -m 10 -M 10000 -f 0.05 -c 3 -o Plantains_prefiltered.vcf -d y
```

a výstupním souborem (parametr `-o`) je vcf soubor, ve kterém byly variantní pozice před-filtrovány následovně: 1) byla zahrnuta pouze data s požadovaným pokrytím (tzv. coverage) `-m` až `-M` čteními; 2) poté byly jako varianty vybrány pouze alely podporované alespoň třemi čteními (parametr `-c`) a frekvencí ≥ 0,05 (parametr `-f`); 3) nakonec byly zachovány pouze pozice, u kterých se jednalo o alespoň jednu alelu odlišnou od reference.

Poslední krok filtrování SNP variant byl proveden programem vcfFilter. V rámci analýzy byly ponechány pouze bialelické pozice (parametr `--RmAlAlt`). Parametry `--MinCov` a `--MaxCov` zajistily převod datových bodů, které byly příliš

pokryty (pravděpodobně vyplívají z repetitivních sekvencí) a které nebyly pokryty dostatečně, na chybějící data. Dále bylo zajištěno, že každá alela byla pokryta alespoň třemi čteními a pokud ne, byla převedena na chybějící data (parametr `--MinAl`). Zadáním parametru `--nMiss` bylo zajištěno, že počet chybějících dat na řádek bude maximálně 1 a pomocí parametru `-g`, byl finální vcf soubor zkomprimován.

Analýza byla spuštěna příkazovým řádkem:

```
python3 vcfFilter.1.0.py --vcf Plantains_prefiltered.vcf --names all_names.tab --MinCov 10 --MaxCov 300 --MinAl 3 --nMiss 1 --RmAlAlt 1:3:4:5:6 --prefix Plantains_Filtered -g y.
```

V další části byl finální vcf soubor obsahující filtrované jednonukleotidové polymorfismy použit jako vstup pro program vcf2allPropAndCov, který na základě vcf souboru provedl: 1) vykreslení alelového pokrytí podél chromozomů u jednotlivých zástupců; a 2) identifikaci alel specifických pro každou skupinu na základě známých předků (AA, BB) ve vcf souboru a vykreslení poměru alel na dané pozici podél všech chromozomů u všech zástupců.

Pro analýzu diploidních předků byl příkazový řádek ve tvaru:

```
python3 vcf2allPropAndCov.py --conf Plantain_Vcf.conf --origin Origin.tab --acc Předek --ploidy 2 --NoMiss n --all y.
```

Analýza triploidních zástupců afrických plantainů byla spuštěna příkazovým řádkem:

```
python3 vcf2allPropAndCov.py --conf Plantain_Vcf.conf --origin Origin.tab --acc Zástupce --ploidy 3 --NoMiss n --all y,
```

kde:

`--conf` je soubor obsahující cesty k vcf souborům,

`--origin` je tvořen dvěma sloupci (první sloupec obsahuje názvy předků a druhý sloupec obsahuje jejich skupiny/origin),

`--acc` značí název analyzovaného zástupce,

`--ploidy` udává ploidii analyzovaných zástupců,

`--NoMiss` značí, že pro přiřazení alel ke skupinám nejsou povolena chybějící data,

`--all` udává, že by alela měla být přítomna u všech zástupců skupiny.

Finální vcf soubor byl také použit k analýze programem vcf2allPropAndCovByChr, který porovnává vždy jeden konkrétní chromozom u všech zástupců a vykresluje alelické pokrytí chromozomů. Parametry byly zadávány podobně jako u předchozí analýzy.

Příkazový řádek byl ve tvaru:

```
python3 vcf2allPropAndCovByChr.py --conf Plantain_Vcf.conf --origin Origin.tab --ploidy 3 --NoMiss n --all y --acc názvy_všech_zástupců
```

Na základě vykreslení poměru pokrytí alel podél chromozomů a výpočtů normalizovaného pokrytí pozic podél chromozomů poskytnutých programem vcf2allPropAndCov, byla provedena identifikace statisticky významných jednonukleotidových polymorfismů sloužících k analýze zastoupení subgenomů A a B v allotriploidních plantainech. Výstupní soubory také obsahovaly informace o počtech SNP specifických pro subgenom A, subgenom B a SNP odpovídajících oběma subgenomům (AA:BB).

Ve vcf souborech byly ponechány pouze SNP stoprocentně přiřazené subgenomu A nebo subgenomu B (odpovídající poměr byl roven 1.0) a dále SNP s duplicitními pozicemi odpovídajícími oběma subgenomům.

K odhalení chromozomové konstituce AAB nebo ABB bylo potřeba ponechat hlavně SNP, jejichž poměr pokrytí odpovídal přibližně 33 % pro alely specifické subgenomu B ku 66 % pro subgenom A (tj. dvojnásobné pokrytí subgenomu A - AAB) či naopak (ABB). Filtrace proto pokračovala náhodným výběrem SNP pozic, jejichž hodnoty poměru pokrytí pro alely specifické pro subgenomy A a B, se pohybovaly v rozpětí mezi 0.15 až 0.35 a 0.65 až 0.85. Ponechány byly také SNP, jejichž poměr pokrytí byl roven číselné hodnotě s maximálně třemi desetinnými místy. Posledním krokem bylo provedení filtrace programem vcfFilter, tentokrát s hodnotami parametrů `--MinCov 30 --MaxCov 100000` pro SNP s poměrem 1.0 a `--MinCov 15 --MaxCov 100000` pro zbývající SNP, čímž se počet SNP výrazně zredukoval.

4 VÝSLEDKY A DISKUZE

4.1 Kontrola kvality a úprava sekvenačních dat

Kontrola kvality provedená programem FastQC poskytla informace o sekvenačních datech. Průměrná délka sekvencí byla stanovena na 35–151 bp se 40 % obsahem GC bází. Dále byly poskytnuty informace o počtu sekvenačních čtení jednotlivých zástupců (Tab. 4).

Sekvenační čtení byla dále upravena softwarem Trimmomatic, který ponechal pouze čtení o délce 150-nt, dále softwarem FASTQ Quality Filter, kterým byly odfiltrovány sekvence s kvalitou nižší než prahová hodnota 20 ve více než 10 % bází, a nakonec byla sekvenační čtení spárována softwarem Pairfq. Upravením sekvenačních dat se počet párových čtení jednotlivých zástupců snížil v průměru o 30 % (Tab. 4).

Tab. 4: Počty čtení před a po úpravě, R1 a R2 značí párová čtení (forward a reverse).

Název kultivaru		**Počet sekvenačních čtení**	
		před úpravou	**po úpravě**
Obino l'Ewai	7_L001_R1	39 686 210	27 879 023
	7_L001_R2	39 686 210	27 879 023
Msisa	9_L001_R1	34 747 996	24 353 076
	9_L001_R2	34 747 996	24 353 076
Apem Pa	12_L001_R1	40 223 263	28 318 181
	12_L001_R2	40 223 263	28 318 181
Wine Plantain	16_L001_R1	47 439 237	33 408 938
	16_L001_R2	47 439 237	33 408 938
Baka	30_L001_R1	30 195 157	21 371 060
	30_L001_R2	30 195 157	21 371 060
Apantu	33_L001_R1	44 491 713	30 928 066
	33_L001_R2	44 491 713	30 928 066
Orishele	38_L001_R1	46 514 563	32 877 333
	38_L001_R2	46 514 563	32 877 333
Dominico Rojo	42_L001_R1	27 361 621	19 359 103
	42_L001_R2	27 361 621	19 359 103
3 Hands Planty	46_L001_R1	35 061 362	25 013 328
	46_L001_R2	35 061 362	25 013 328
Tshambunu	48_L001_R1	56 774 303	40 683 436
	48_L001_R2	56 774 303	40 683 436

4.2 Identifikace SNP a analýza genomové struktury

Trimovaná sekvenační data vybraných deseti zástupců afrických plantainů a dvou diploidní planých druhů, které sloužili v rámci analýzy pro identifikaci A- a B-genom specifických jednonukleotidových polymorfismů, byly mapovány na referenční genomovou sekvenci *Musa acuminata* ssp. *malaccensis,* klon DH Pahang. Z výsledných souborů byla odstraněna čtení, která se na referenci nenamapovala. V následném kroku analýzy, který zahrnoval odstranění duplicitních čtení, žádné duplicity identifikovány nebyly. V další části bylo provedeno opětovné lokální zarovnání (tzv. multiple alignment) kolem oblastí indelů.

Pro každou pozici na všech 11 chromozomech referenčního genotypu byl u všech analyzovaných druhů stanoven počet alel. Tímto krokem byly získány primární vcf soubory, které byly následně upravovány a filtrovány s cílem identifikace strukturních změn v rámci studovaných afrických plantainů.

Kompletní filtrované vcf soubory obsahující informace o jednonukleotidových polymorfismech (SNP), byly použity k identifikaci A- a B-genomově specifických SNP. Za tímto účelem byla použita data diploidních ancestrálních genomů – *M. acuminata* spp. *banksii* (A genom) a *M. balbisiana* (B genom), která byla rovněž mapována na referenční sekvenci banánovníku. Díky přítomnosti specifických SNP mezi rodičovskými genomy (A a B) tak bylo možné identifikovat jednonukleotidové polymorfismy specifické pro A- a B-subgenomy triploidních plantainů (AAB), čehož bylo využito pro identifikaci zastoupení (vykreslení poměru) A- a B-specifických alel podél 11 základních chromozomů banánovníku Byly tak získány informace o celkovém počtu A a B specifických SNP identifikovaných pro jednotlivé zástupce plantainů (Tab. 5).

Tab. 5: Počet SNP specifických pro subgenom A (AA) a subgenom B (BB)

Název kultivaru	AA	BB
ITC.0109; Obino l'Ewai	2 682 142	2 592 290
ITC.0142; Msisa	2 673 581	2 564 842
ITC.0219; Apem Pa	2 713 721	2 613 816
ITC.0325; Wine Plantain	2 734 545	2 616 527
ITC.0098; Baka	2 596 963	2 476 348
ITC.0223; Apantu	2 721 900	2 623 942
ITC.0517; Orishele	2 743 519	2 613 218
ITC.0641; Dominico Rojo	2 462 478	2 393 294
ITC.0185; 3 Hands Planty	2 602 624	2 519 263
ITC.0128; Tshambunu	2 704 033	2 608 780

Z tohoto vysokého počtu A- a B-genom specifických jednonukleotidových polymorfismů bylo potřeba identifikovat jen statisticky významné SNP, které byly dále využity pro vizualizaci zastoupení jednotlivých subgenomů (poměru subgenomů) v jaderném genomu triplodních zástupců plantainů. Ponechány byly pouze SNP, které stoprocentně odpovídaly subgenomům A nebo B a zároveň tzv. duplicitní pozice (SNP specifické pro oba subgenomy), kterým odpovídal alelický poměr pokrytí v rozpětí 15-35 % a 65-85 %. Tato filtrace výrazně snížila počet jednonukleotidových polymorfismů pokrývajících 11 chromozomů (Tab. 6). Počet SNP specifických pro subgenom A/B klesl přibližně o 95 %.

Tab. 6: Počty SNP specifických pro subgenomy A/B po filtrování.

Název kultivaru	AA	BB
ITC.0109; Obino l'Ewai	60 236	55 926
ITC.0142; Msisa	55 065	50 820
ITC.0219; Apem Pa	61 259	57 076
ITC.0325; Wine Plantain	60 203	56 104
ITC.0098; Baka	49 495	44 972
ITC.0223; Apantu	60 770	56 681
ITC.0517; Orishele	60 543	56 434
ITC.0641; Dominico Rojo	43 144	38 735
ITC.0185; 3 Hands Planty	54 640	50 312
ITC.0128; Tshambunu	54 998	51 169

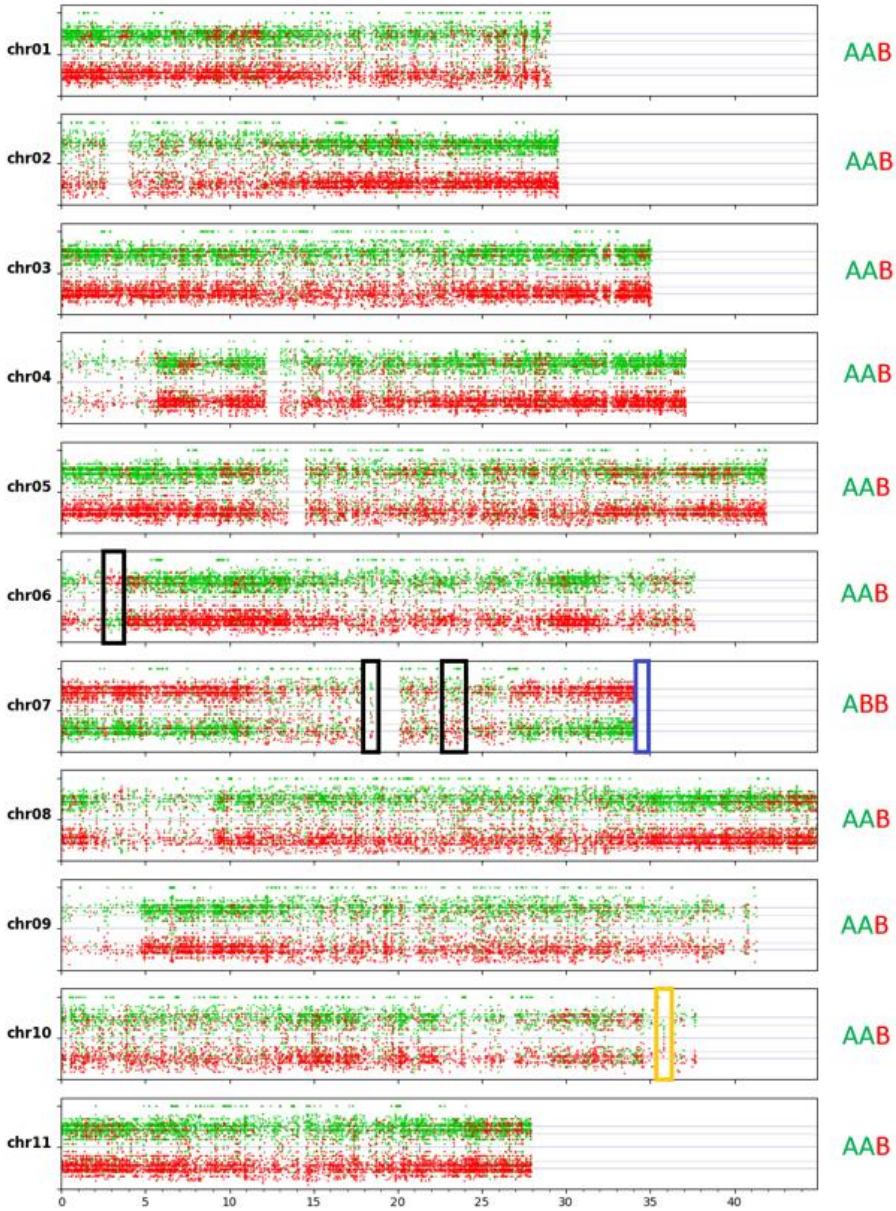
4.2.1 Vizualizace SNP pokrytí chromozomů

Vykreslení pokrytí jednonukleotidových polymorfismů specifických pro subgenomy A a B u allotriploidních zástupců tří hlavních morfologických skupin afrických plantainů prokázalo jejich očekávanou téměř jednotnou euploidní chromozomovou konstituci (příloha 1 – 3).

U všech kultivarů, které jsou běžně klasifikovány jako plantainy s genomovou strukturou AAB, byly výsledky v souladu s touto globální strukturou chromozomů. Bylo u nich také ovšem zjištěno několik významných chromozomálních přestaveb.

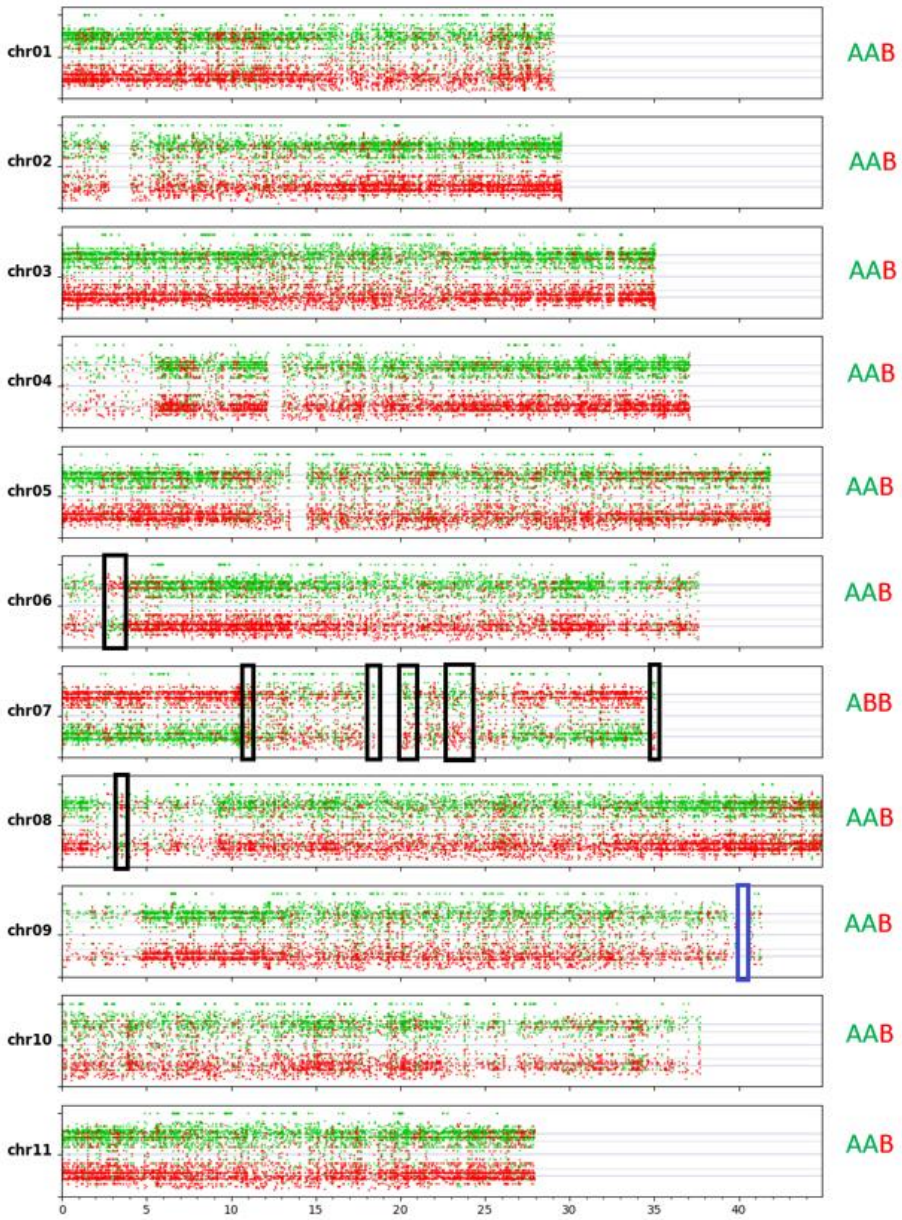
U všech analyzovaných zástupců plantainů byla zjištěna specifická struktura chromozomu 7, kdy je tento chromozom v genomu plantainů zastoupen dvěma kopiemi pocházejícími ze subgenomu B a jednou kopií pocházející ze subgenomu A (Obrázek 14 – 24). Analýza složení a zastoupení rodičovských subgenomů u analyzovaných zástupců triploidních plantainů odhalila i další chromozomální přestavby a potvrdila tak mozaikovou strukturu allotriploidních genomů jedlých banánovníků.

U kultivaru Obino l'Ewai, patřícím mezi morfologickou skupinu French, byl u šestého chromozomu pozorován segment vykazující ABB konstituci. Chromozom 7, byl zastoupen jednou kopií subgenomu A a dvěma kopiemi subgenomu B a vykazoval tak strukturu ABB téměř po celé délce, s výjimkou dvou segmentů očekávané AAB struktury. Dále byla na konci sedmého chromozomu nalezena delece o délce přibližně 1Mb. Další výjimkou byl fragment vykazující diploidní AB konstituci na desátém chromozomu (Obr. 14).



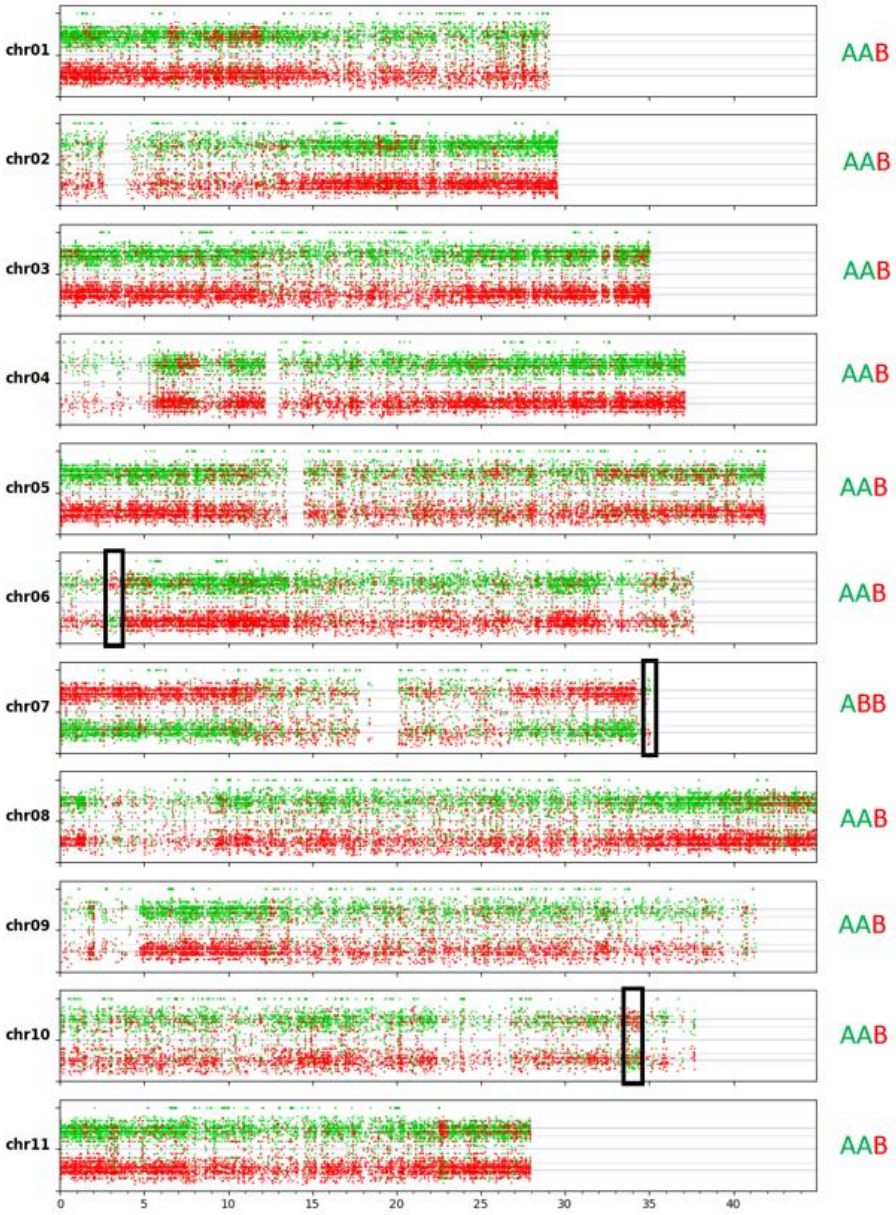
Obr. 14: Mozaiková genomová struktura klonu Obino l'Ewai (French plantain). Zelené body značí poměr pokrytí alel specifických pro subgenom A, červené body značí alely specifické pro subgenom B, vyneseného podél 11 chromozomů referenčního genomu *M. acuminata* spp. *malaccensis*. Černé a žluté rámečky označují chromozomové segmenty lišící se od očekávané genomové konstituce AAB (černě: konstitutce ABB; žlutě: konstituce AB). Modrý rámeček označuje deleci.

Další zástupce morfologické skupiny French, klon Msisa, vykazoval očekávanou genomovou strukturu AAB opět u téměř všech chromozomů, s již zmíněnou výjimkou chromozomu 7 a chromozomu 6. Další změny konstituce genomu byly zaznamenány u chromozomu 8 (Obr. 15) a v koncové částí chromozomu 9, byla nalezena delece o délce přibližně 1 Mb (Obr. 15).



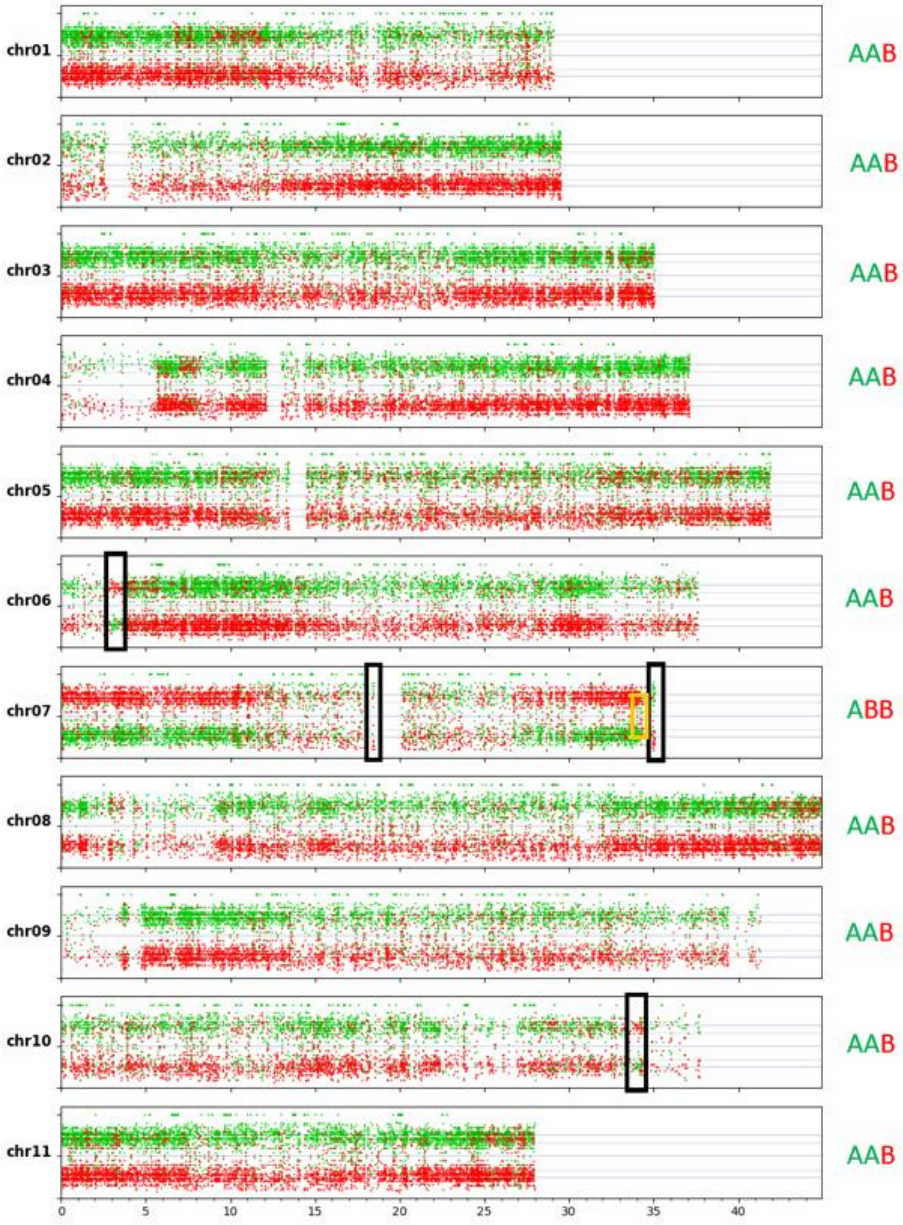
Obr. 15: Mozaiková genomová struktura klonu Msisa (French plantain). Zelené body značí poměr pokrytí alel specifických pro subgenom A, červené body značí alely specifické pro subgenom B, vyneseného podél 11 chromozomů referenčního genomu *M. acuminata*. Černé rámečky označují chromozomové segmenty lišící se od očekávané genomové konstituce AAB a změnu na konstituci ABB. Modrý rámeček označuje deleci.

Klon Apem Pa, který je řazen mezi africké plantainy typu French, opět vykazoval po délce sedmého chromozomu ABB konstituci, kromě koncové oblasti, kde byl nalezen úsek s běžnou AAB konstitucí. Chromozomy 6 a 10 se lišili oproti očekávané konstituci pouze v rámci krátkých fragmentů, které tvořila ABB struktura (Obr. 16).



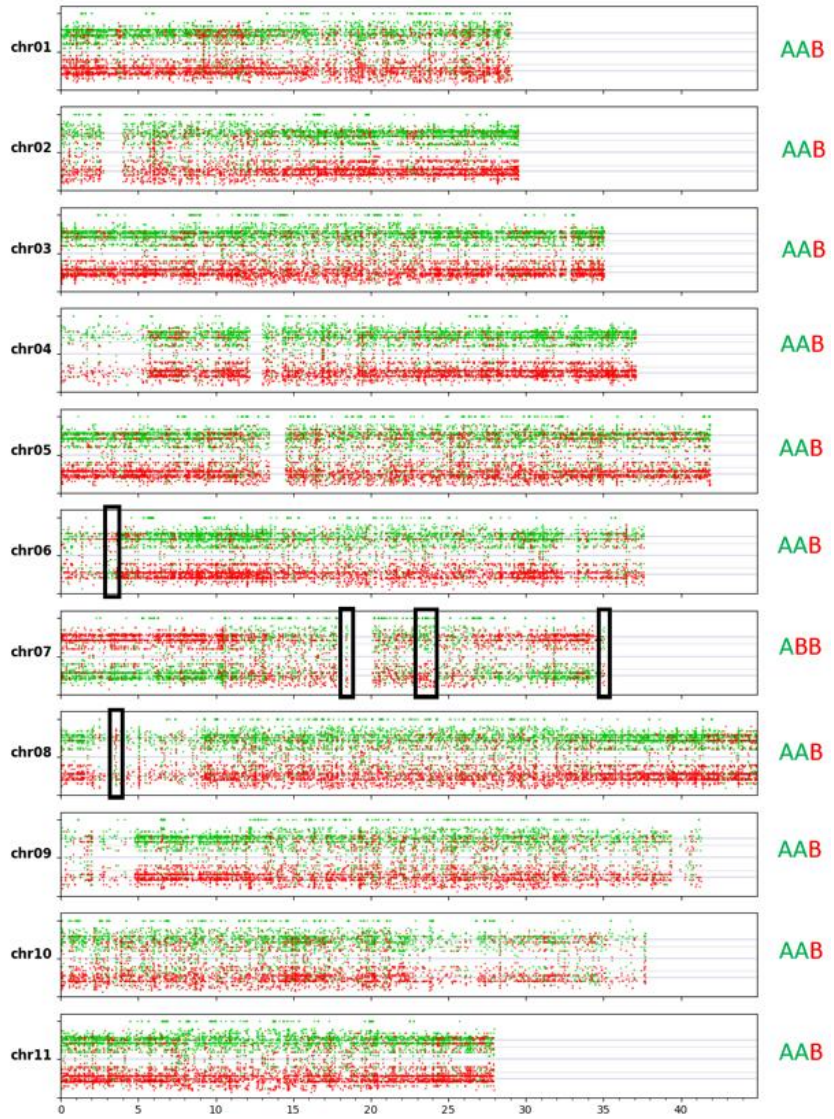
Obr. 16: Mozaiková genomová struktura klonu Apem Pa (French plantain). Zelené body značí poměr pokrytí alel specifických pro subgenom A, červené body značí alely specifické pro subgenom B, vyneseného podél 11 chromozomů referenčního genomu *M. acuminata*. Černé rámečky označují chromozomové segmenty lišící se od očekávané genomové konstituce (AAB → ABB).

Posledním analyzovaným zástupcem morfologické skupiny French, byl Wine Plantain, který se oproti jiným kultivarům skupiny French lišil fragmentem s diploidní AB konstitucí, nacházející se v koncové oblasti chromozomu 7. Chromozom 7 také napříč téměř celé své délky vykazoval opět ABB strukturu, s výjimkou dvou fragmentů v koncovém a pericentromerickém regionu. Další výjimku tvořily fragmenty s ABB konstitucí na chromozomech 6 a 10, které byly identifikovány i u předchozího kultivaru Apem Pa (Obr. 17).



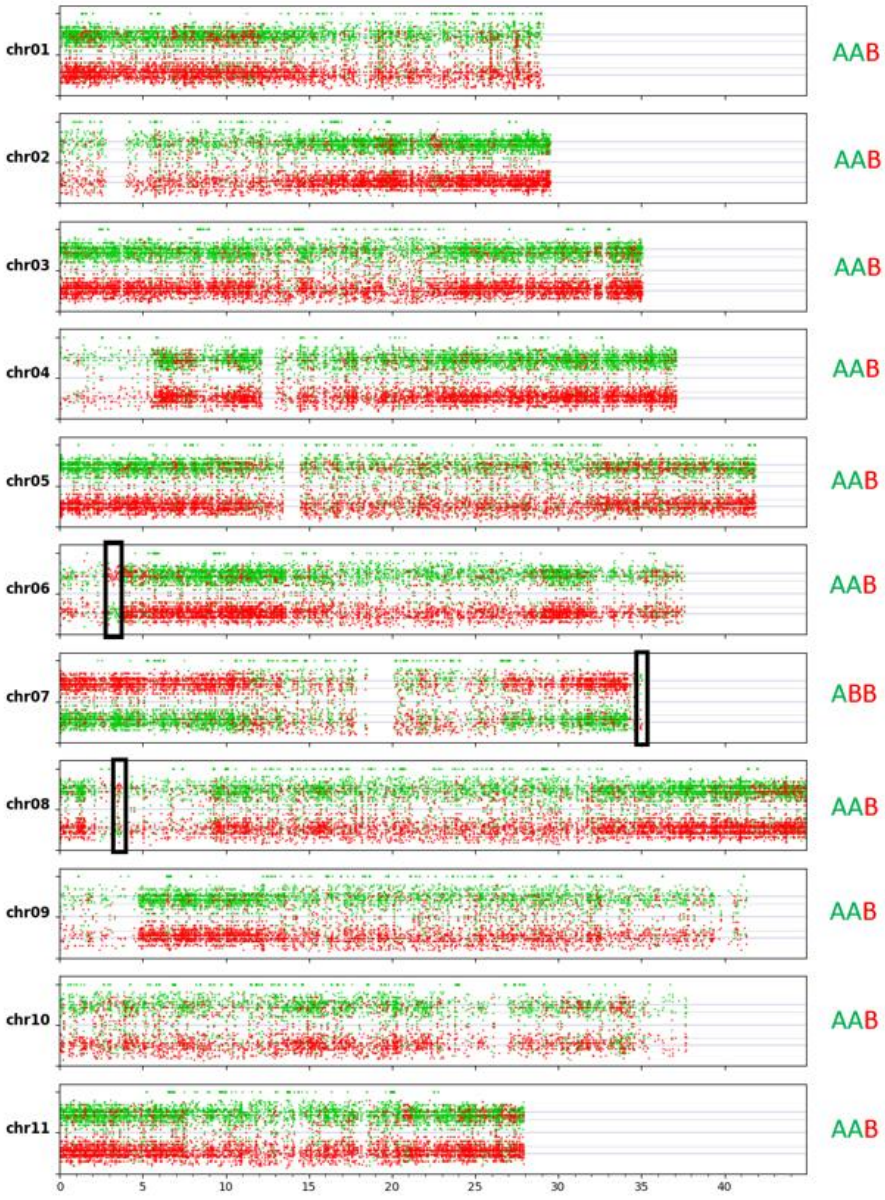
Obr. 17: Mozaiková genomová struktura klonu Wine Plantain (French plantain). Zelené tečky značí poměr pokrytí alel specifických pro subgenom A, červené tečky značí alely specifické pro subgenom B, vyneseného podél 11 chromozomů referenčního genomu *M. acuminata*. Černé (konstituce ABB) a žluté rámečky (konstituce AB) označují chromozomové segmenty lišící se od očekávané genomové konstituce AAB.

V genomové struktuře afrického plantainu Baka, patřícího mezi zástupce morfologické skupiny False horn, nebyly pozorovány žádné výrazné odlišnosti oproti chromozomálním strukturám zástupců typu French. U kultivaru byly opět nalezeny rozdíly v šestém a osmém chromozomu. Fragmenty těchto chromozomů nebyly v souladu s očekávanou strukturou a vykazovaly konstituci ABB. Struktura ABB tvořila téměř celý chromozom 7, s výjimkou krátkých fragmentů v oblasti pericentromery a na konci chromozomu (Obr. 18).



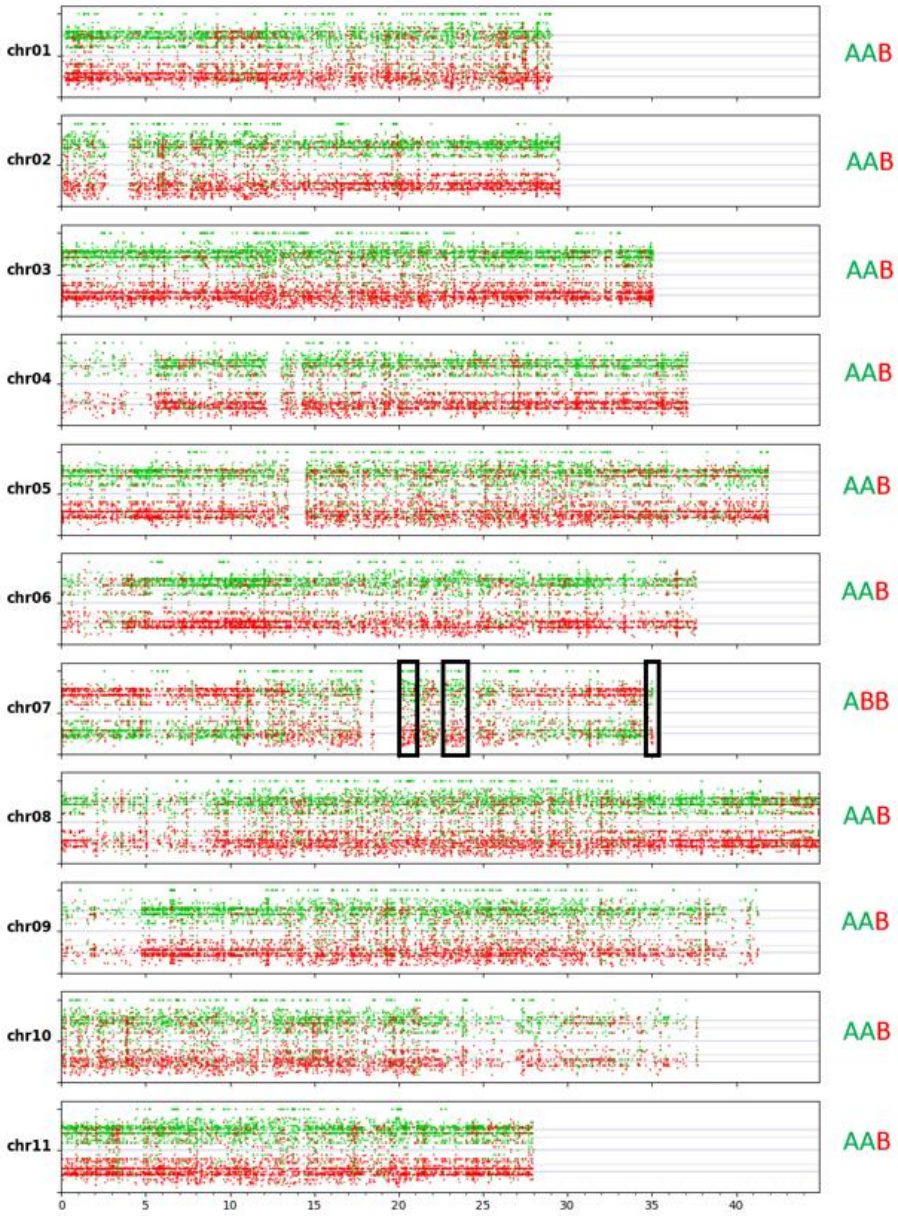
Obr. 18: Mozaiková genomová struktura klonu Baka (False horn plantain). Zelené body značí poměr pokrytí alel specifických pro subgenom A, červené body značí alely specifické pro subgenom B, vyneseného podél 11 chromozomů referenčního genomu *M. acuminata*. Černé rámečky označují chromozomové segmenty lišící se od očekávané genomové konstituce (AAB → ABB).

Sedmý chromozom klonu Orishele (False horn) s výjimkou koncového fragmentu vykazoval ABB strukturu, která byla pozorována i v krátkých úsecích chromozomů 6 a 8 (Obr. 19).



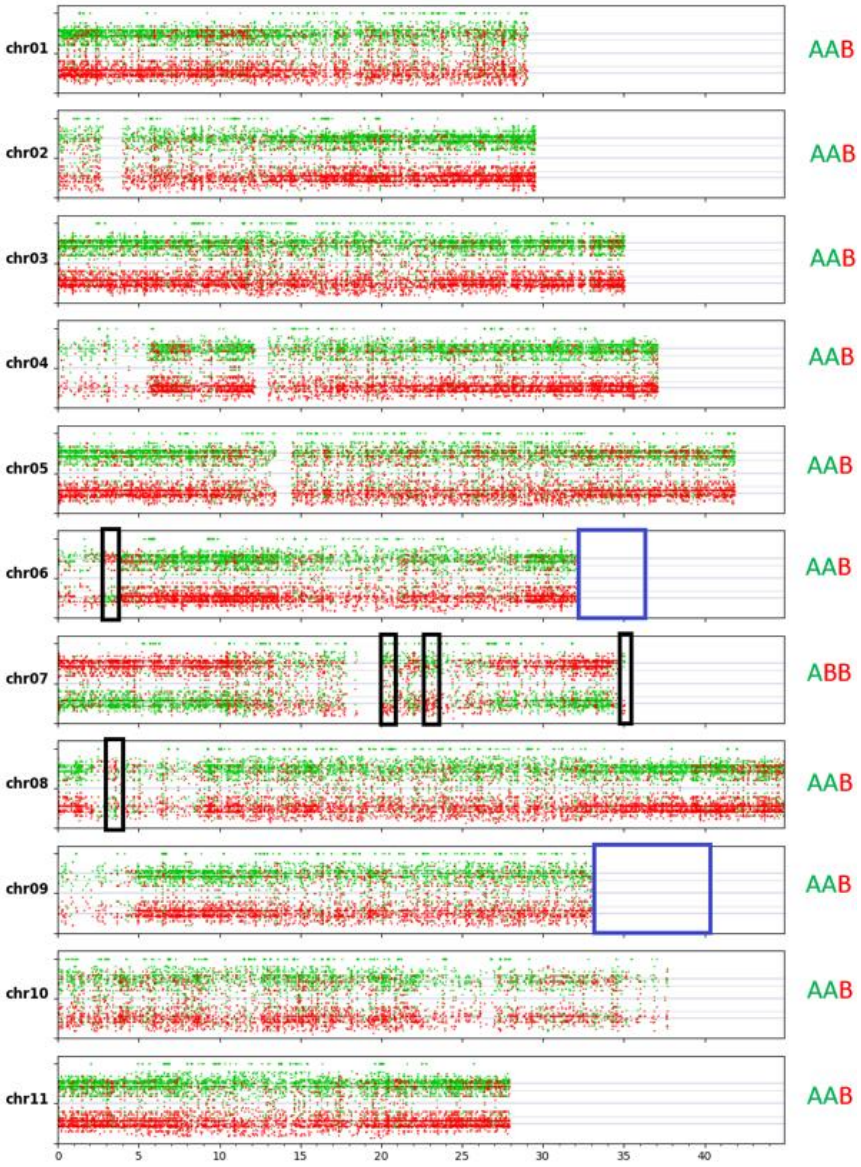
Obr. 19: Mozaiková genomová struktura klonu Orishele (False horn plantain). Zelené body značí poměr pokrytí alel specifických pro subgenom A, červené body značí alely specifické pro subgenom B, vyneseného podél 11 chromozomů referenčního genomu *M. acuminata*. Černé rámečky označují chromozomové segmenty lišící se od očekávané genomové konstituce (AAB → ABB).

Poslední zástupce morfologické skupiny False horn, klon Dominico Rojo, vykazoval očekávanou chromozomovou konstituci AAB u všech chromozomů, kromě sedmého chromozomu, u kterého byla opět pozorována ABB konstituce s fragmenty AAB (Obr. 20).



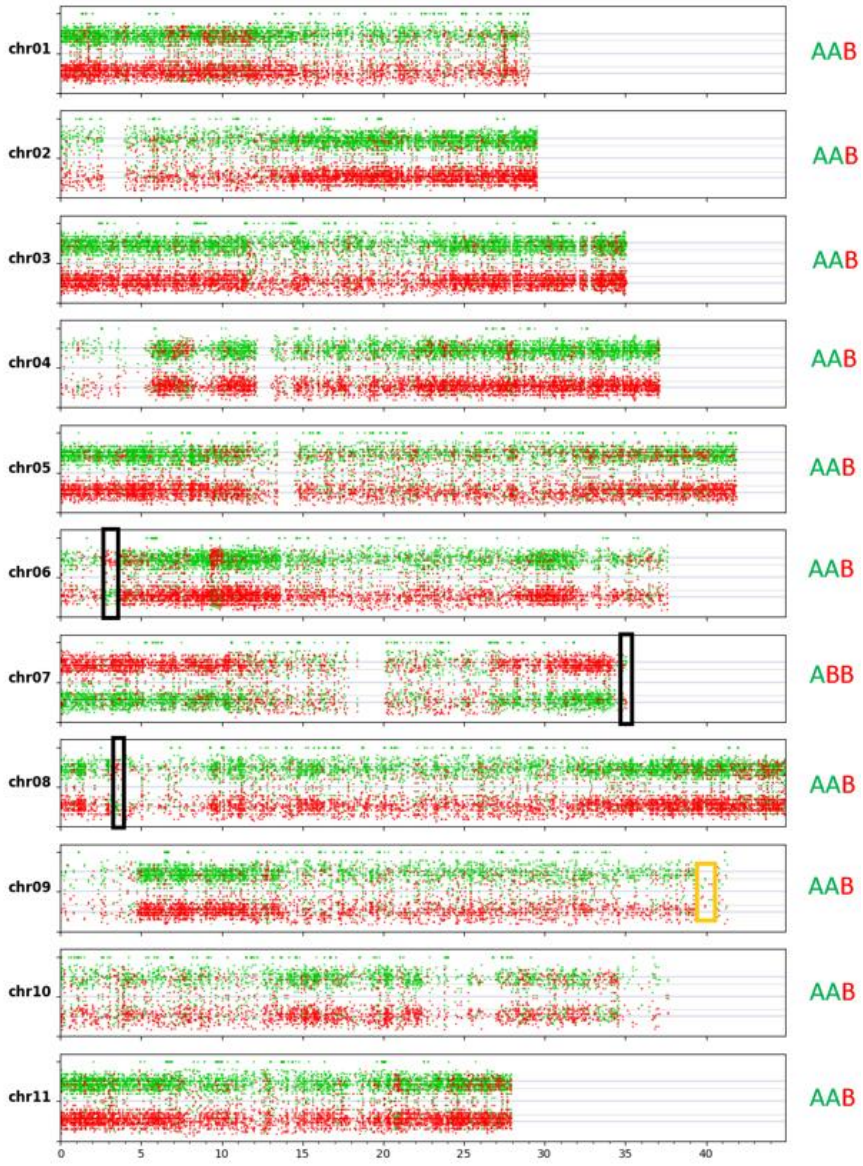
Obr. 20: Mozaiková genomová struktura klonu Dominico Rojo (False horn plantain). Zelené body značí poměr pokrytí alel specifických pro subgenom A, červené body značí alely specifické pro subgenom B, vyneseného podél 11 chromozomů referenčního genomu *M. acuminata*. Černé rámečky označují chromozomové segmenty lišící se od očekávané genomové konstituce (AAB→ ABB).

Kultivar 3 Hands Planty, který patří k plantainům typu Horn, obsahoval na koncích šestého a devátého chromozomu deletované oblasti, dlouhé přibližně 6 Mb a 8 Mb. U šestého chromozomu byl pozorován také krátký fragment s ABB strukturou, která byla také nalezena na začátku osmého chromozomu a po téměř celé délce chromozomu 7 (Obr. 21).



Obr. 21: Mozaiková genomová struktura klonu 3 Hands Planty (Horn). Zelené body značí poměr pokrytí alel specifických pro subgenom A, červené body značí alely specifické pro subgenom B, vyneseného podél 11 chromozomů referenčního genomu *M. acuminata*. Černé rámečky označují chromozomové segmenty lišící se od očekávané genomové konstituce (AAB→ ABB). Modré rámečky označují deleci.

Posledním zástupcem afrických plantainů typu Horn a zároveň posledním analyzovaným klonem byl Tsahambunu, jehož sedmý chromozom a krátké úseky na počátcích chromozomů 6 a 8, byly opět zastoupeny jednou kopií subgenomu A a dvěma kopiemi subgenomu B, obsahoval také krátký fragment s diploidní AB konstitucí na koncové části devátého chromozomu (Obr. 22).



Obr. 22: Mozaiková genomová struktura klonu Tshambunu (Horn). Zelené body značí poměr pokrytí alel specifických pro subgenom A, červené body značí alely specifické pro subgenom B, vyneseného podél 11 chromozomů referenčního genomu *M. acuminata*. Černé rámečky (konstituce ABB) a žluté (konstituce AB) označují chromozomové segmenty lišící se od očekávané genomové konstituce AAB.

Jak je patrné ze získaných výsledků, všech deset analyzovaných genotypů jedlých plantainů má ve svém genomu chromozom 7 zastoupený dvěmi kopiemi pocházející ze subgenomu B a jednu kopii pocházející ze subgenomu A, což je rozporu s předpokládaným genomovým složením těchto jedlých typů banánovníku. Stejná genomová konstituce charakteristická pro chromozom 7 u plantainů byla identifikována také ve studii Baurens *et al.*, (2019), kde autoři poprvé aplikovali program vcfHunter pro analýzu struktury mezidruhových banánovníků, včetně jednoho klonu charakterizujícího plantainy. Tato původní studie také ukázala, že většina mezidruhových hybridů banánovníku má mozaikovou strukturu genomů, kdy v rámci jednotlivých chromozomů dochází k rozličným strukturním změnám, zahrnujícím nejen rozdílné zastoupení jednotlivých subgenomů odlišných od předpokládaného genomového složení daného mezidruhového hybrida, ale naznačuje také přítomnost delecí koncových částí chromozomů.

Jak již bylo zmíněno, triploidní plantainy s genomem AAB vznikly dvěma kroky mezidruhové hybridizace, kdy nejdříve došlo ke zkřížení dvou diploidních planě rostoucích druhů *Musa acuminata* ssp. *banksii* (genom A) a *Musa balbisiana* (genom B), za vzniku mezidruhového diploidního hybridního jedince s genomem AB a jeho následné hybridizaci (neredukované 2n gamety) s haploidní gametou druhu *Musa acuminatou* ssp. *banksii*, což vedlo k výslednému triploidnímu potomstvu AAB. Práce De Langhe *et al.* (2010) naznačila, že proces vzniku jedlých banánovníků byl pravděpodobně mnohem komplikovanější, tím že obsahoval několik kol zpětných křížení. Další komplikaci vzniku jedlých banánovníků mohla představovat skutečnost, kdy se na křížení mohly podílet hybridní diploidní genotypy či strukturní heterozygoti spíše než homozygotní diploidní druhy *M. acuminata* spp. *banksii* a *M. balbisiana*. Tato skutečnost tak mohla mít za následek vytvoření mozaikové struktury jedlých hybridních banánovníků, což bylo naznačeno jak v práci Baurens *et al.* (2019), tak i ve výsledcích předkládané diplomové práce.

Oproti studii Baurens *et al.* (2019), kdy autoři pro identifikaci struktury složení genomů mezidruhových hybridů využili sekvenační data získaná přístupem založeným na redukci komplexity, RADseq data, v naší práci bylo využito sekvencí získaných resekvenováním celkové genomové DNA. Byly tak získány kvalitní datasety, které nejsou zatíženy chybovostí (přítomností tzv. missing dat) ke které dochází při aplikaci redukce komplexity. Celogenomovým sekvenováním bylo získáno větší množství specifických jednonukleotidových polymorfismů, které budou v následující

práci využity také pro podrobnou analýzu genetické variability této skupiny jedlých triploidních banánovníků a případnou identifikaci SNP markerů charakteristických pro jednotlivé morfotypy plantainů. Dosud byla genetická diverzita jedlých banánovníku analyzována pomocí různých molekulárních markerů, např. SSR markerů, DarT markerů nebo pomocí GBS, které ovšem nevedly k jednoznačné charakterizaci jednotlivých genotypů či jednotlivých skupin jedlých banánovníků (např. Christelová *et al.*, 2017, Ruas *et al.*, 2017, Sardos *et al.*, 2016a, Sardos *et al.*, 2016b). Také proto, je dnes brán v úvahu jejich rozdílný epigenetický profil, který tak může mít velký vliv na jejich diverzitě. Komplexní resekvenační data a SNP identifikované v rámci předkládané diplomové práce budou v budoucnu také sloužit jako reference pro analýzu epigenetického profilu vybraných plantainů.

5 ZÁVĚR

Diplomová práce pojednává o aplikaci NGS sekvenování pro analýzu složení a struktury genomů jedlých škrobových banánovníků, zvaných plantainy, které patří mezi nejrozšířenější jedlé typy banánovníků. Úvod diplomové práce se věnuje nejen problematice vzniku plantainů, charakteristik jejich jaderných genomů, ale tvoří také úvod do problematiky sekvenování nové a třetí generace, a *in silico* analýze Next-gen sekvenačních dat s cílem analýzy organizace a struktury jaderných genomů rostlin.

Cílem experimentální části práce byla analýza celogenomových resekvenačních dat deseti vybraných druhů afrických plantainů, za účelem zjištění strukturní variability jejich jaderných genomů.

Sekvenační data byla nejprve upravena a filtrována programy dostupnými na serverech Centra strukturní a funkční genomiky rostlin, které jsou součástí MetaCentra. Data byla následně použita pro mapování na referenční genomovou sekvenci *M. acuminata* ssp. *malaccensis* (klon DH Pahang, verze 2). Výsledná namapovaná čtení byla dále zpracovávána VcfHunterem slučujícím několik programů dohromady a jejichž výsledkem je vykreslení poměru a pokrytí alel specifických pro subgenomy A a B.

Pomocí programů VcfHunteru bylo identifikováno průměrně více než 2 miliony jednonukleotidových polymorfismů specifických pro subgenom A, a více než 2 miliony SNP odpovídajících subgenomu B. Následnou filtrací sekvenačních dat se počet SNP zredukoval přibližně o 95 %.

Identifikace SNP jednoznačně charakterizujících A a B subgenomy plantainu umožnilo vykreslení pokrytí A/B specifických alel podél 11 chromozomů referenčního genomu banánovníku a zkoumání chromozomových struktur jednotlivých zástupců afrických plantainů.

U analyzovaných kultivarů nebylo, podle očekávání, pozorováno výrazné odlišení od běžné konstituce AAB. Byly ale ovšem objeveny také krátké oblasti několika chromozomů s konstitucemi, které nebyly v souladu s globální strukturou chromozomů, jednalo se o ABB struktury, a oblasti s diploidní AB strukturou, kde s největší pravděpodobností došlo k deleci v rámci jednoho chromozomu pocházejícího ze subgenomu A. U všech analyzovaných plantainů bylo navíc zjištěno, že chromozom 7 je téměř po celé délce zastoupen jednou kopií subgenomu A a dvěma

kopiemi subgenomu B. Struktura sedmého chromozomu odpovídá i výsledkům práce Baurens *et al.* (2019) a je tedy možné, že by daná struktura mohla charakterizovat všechny plantainy.

Chromozomy kultivarů Obino l'Ewai, Msisa a 3 Hands Planty obsahovaly také různě dlouhé delece koncových oblastí chromozomů. Delece u klonu Obino l'Ewai se nacházela na sedmém chromozomu, u klonu Msisa na devátém chromozomu a u klonu 3 Hands Planty, byly nalezeny dlouhé delece na koncích chromozomů 6 a 9.

Objasnění složení a strukturních změn jedlých klonů banánovníku, stejně tak jejich pravděpodobných rodičovských genomů je velmi důležité pro výběr vhodných genotypů pro tradiční šlechtění banánovníku. Škrobové banánovníky typu plantain představují jednu z nejvýznamnějších hospodářských plodin světa, a to hlavně v tropických a subtropických oblastech Afriky, kde jsou jednou ze základních potravin. Objasnění složení a strukturních změn jedlých klonů banánovníku, stejně tak jejich pravděpodobných rodičovských genomů je velmi důležité pro výběr vhodných genotypů pro tradiční šlechtění banánovníku.

Všechny stanovené cíle této diplomové práce byly splněny.

6 LITERATURA

Adey A., Morrison H., Asan Xun X., Kitzman J., Turner E. (2010): Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* **11**(12), 119.

Adheka J. G., Dhed'a D. B., Karamura D., Blomme G., Swennen R., De Langhe E. (2018): The morphological diversity of plantain in the Democratic Republic of Congo. *Scientia Horticulturae* **234**,126 - 133.

Adheka J. G. (2014): Contribution to the characterization and classification of the Congo basin African plantains (*Mus*a AAB) in the Democratic Republic of Congo. PhD thesis. Department of Biotechnological Sciences, Faculty of Sciences, University of Kisangani.**114**.

Agarwal M., Shrivastava N., Padh H. (2008): Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Rep*. **27**, 617–631.

Aggeli D., Karas V. O., Sinnott-Armstrong N. A., Varghese V., Shafer R. W., Greenleaf W. J., Sherlock G. (2018): Diff-seq: A high throughput sequencing-based mismatch detection assay for DNA variant enrichment and discovery. *Nucleic Acids Res*. **46**(7), e42.

Albers C. A., Lunter G., MacArthur D. G., McVean G., Ouwehand W. H., Durbin R. (): Dindel: accurate indel calls from short-read data. *Genome Res*. **21**(6), 961-973.

Asif, M. J., Mak, C., and Othman, R. Y. (2001): Characterization of indigenous *Musa* species based on flow cytometric analysis of ploidy and nuclear DNA content. *Caryologia*. **54** (2), 161–168. doi: 10.1080/00087114.2001.10589223.

Baird N. A., Etter P. D., Atwood T. S., Currey M. C., Shiver A. L., Lewis Z. A., Selker E. U., Cresko W. A., Johnson E. A. (2008): Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**, e3376.

Bansal V. (2017): A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. *BMC bioinformatics* **18**(Suppl 3), 43.

Bartoš, J., Alkhimova, O., Doleželová, M., De Langhe, E., and Doležel, J. (2005): Nuclear genome size and genomic distribution of ribosomal DNA in *Musa* and *Ensete* (*Musaceae*): taxonomic implications. Cytogenet. *Genome Res*. **109**, 50–57. doi: 10.1159/000082381.

Baurens F. C., 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. *Mol Biol Evol*. **36**(1): 97-111. doi: 10.1093/molbev/msy199.

Boeva V., Zinovyev A., Bleakley K., Vert J. P., Janoueix-Lerosey I., Delattre O., Barillot E. (2011): Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27**, 268-269.

Bolger A. M., Lohse M., Usadel B. (2014): Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

Brown C. G., Clarke J. (2016): Nanopore development at Oxford Nanopore. *Nature Biotechnology* **34**, 810 - 811.

Buermans H. P., den Dunnen J. T. (2014): Next generation sequencing technology: Advances and applications. *Biochim Biophys Acta*. **1842**(10), 1932-1941.

Čížková, J., Hřibová, E., Christelová, P., Van den Houwe, I., Häkkinen, M., Roux, N., et al. (2015): Molecular and cytogenetic characterization of wild *Musa* species. *PLoS One* **10**:e0134096. doi: 10.1371/journal.pone.0134096.

D'Hont, A., Denoeud, F., Aury J. M., Baurens F. C., Carreel F., Garsmeur O., Noel B., Bocs S., Droc G., Rouard M., Da Silva C., Jabbari K., Cardi C., Poulain J., Souquet M., Labadie K., Jourda C., Lengelle J., Rodier-Goud M., Alberti A., Bernard M., Correa M., Ayyampalayam S., Mckain M. R., Leebens-Mack J., Burgess D., Freeling M., Chabannes M., Wicker T., Panaud O., Barbosa J., Hřibová E., Heslop-Harrison P., Rivallan R., Francois P., Poiron C., Kilian A., Burthia D., Jenny Ch., Bakry F., Guignon V., Kema G., Dita M., Waalwijk C.,

Joseph S., Dievart A., Jaillon O., Leclercq J., Argout X., Jeridi M., Doležel J., Roux N., Risterucci A. M., Weissenbach J., Ruiz M., Glaszmann J. Ch., Wincker P. (2012): The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213–217.

Daniells, J., Jenny Ch., Karamura D., Tomekpe K., (2001): Diversity in the Genus *Musa*. INIBAP.

Davey, M. W., Gudimella R., Harikrishna J. A., Wan Sin L., Khalid N., Neulemans J. (2013): A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter-and intra-specific *Musa* hybrids. *BMC Genom*. **14**, 683.

Deamer D., Akeson M., Branton D. (2016): Three decades of nanopore sequencing. *Nature Biotechnology* **34**, 518 - 524.

De Langhe E., Hřibová E., Carpentier S., Doležel J., Swennen R. (2010): Did backcrossing contribute to the origin of hybrid edible bananas? *Ann Bot*. **106**(6): 849–857.

De Langhe, E. Vrydaghs L., de Maret P., Denham T. (2009): Why bananas matter: an introduction to the history of banana domestication. *Ethnobot. Res. Appl*. **7**, 165–177.

Denham T. (2004): The roots of agriculture and arboriculture in New Guinea: Looking beyond Austronesian expansion, Neolithic packages and indigenous origins. *World Archaeology* **36**(4), 610-620.

Denham T. (2010): From Domestication Histories to Regional Prehistory: Using Plants to Re-evaluate Early and Mid-Holocene Interaction between New Guinea and Southeast Asia. *Food & History* **8**(1), 3-22.

DePristo M. A., Banks E., Poplin R., Garimella K. V., Maguire J. R., Hartl C., Philippakis A. A., del Angel G., Rivas M. A., Hanna M., McKenna A., Fennell T. J., Kernytsky A. M., Sivachenko A. Y., Cibulskis K., Gabriel S. B., Altshuler D., Daly M. J. (2011): A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. **43**(5), 491-498.

Dodds K. S., Simmonds N. W. (1948): Sterility and parthenocarpy in diploid hybrids of *Musa*. Heredity **2**(Pt 1):101–117.

Doležel, J., Doleželová, M., and Novák, F. J. (1994). Flow cytometric estimation of nuclear DNA amount in diploid bananas (*Musa acuminata* and *Musa balbisiana*). *Biol. Plant* **36**, 351–357. doi: 10.1007/BF02920930.

Dupouy M., Baurens F. C., Derouault P., Hervouet C., Cardi C., Cruaud C., Istace B., Labadie K., Guiougou C., Toubi L., Salmon F., Mournet P., Rouard M., Yahiaoui N., Lemainque A., Martin G., D'Hont A. (2019): Two large reciprocal translocations characterized in the disease resistance-rich burmannica genetic group of *Musa acuminata*. *Ann Bot.* **124**(2), 319 - 329.

Ebbert M. T., Wadsworth M. E., Staley L. A., Hoyt K. L., Pickett B., Miller J., Duce J.; Alzheimer's Disease Neuroimaging Initiative, Kauwe J. S., Ridge P. G. (2016): Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics* **17** Suppl 7, 239.

Eid J., Fehr A., Gray J., Luong K., Lyle J., Otto G., Peluso P., Rank D., Baybayan P., Bettman B., Bibillo A., Bjornson K., Chaudhuri B., Christians F., Cicero R., Clark S., Dalal R., Dewinter A., Dixon J., Foquet M., Gaertner A., Hardenbol P., Heiner C., Hester K., Holden D., Kearns G., Kong X., Kuse R., Lacroix Y., Lin S., Lundquist P., Ma C., Marks P., Maxham M., Murphy D., Park I., Pham T., Phillips M., Roy J., Sebra R., Shen G., Sorenson J., Tomaney A., Travers K., Trulson M., Vieceli J., Wegener J., Wu D., Yang A., Zaccarin D., Zhao P., Zhong F., Korlach J., Turner S. (2009): Real-time DNA sequencing from single polymerase molecules. *Science* **323**(5910), 133 – 138.

Elshire R. J., Glaubitz J. C., Sun Q., Poland J. A., Kawamoto K., Buckler E. S., Mitchell S. E. (2011): A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**(5), e19379.

Escaramís G., Docampo E., Rabionet R. (2015): A decade of structural variants: description, history and methods to detect structural variation. *Briefings in Functional Genomics* **14**(5), 305–314.

Feuillet C., Leach J. E., Rogers J., Schnable P. S., Eversole K. (2011): Crop genome sequencing: lessons and rationales. *Trends Plant Sci*. **16**, 77 - 88.

Gao G., Smith D. I. (2015): Mate-Pair Sequencing as a Powerful Clinical Tool for the Characterization of Cancers with a DNA Viral Etiology. *Viruses* **7**, 4507 - 4528.

Gargis A. S., Kalman L., Berry M. W., Bick D. P., Dimmock D. P., Hambuch T., Lu F., Lyon E., Voelkerding K. V., Zehnbauer B. A., Agarwala R., Bennett S. F., Chen B., Chin E. L., Compton J. G., Das S., Farkas D. H., Ferber M. J., Funke B. H., Furtado M. R., Ganova-Raeva L. M., Geigenmüller U., Gunselman S. J., Hegde M. R., Johnson P. L., Kasarskis A., Kulkarni S., Lenk T., Liu C. S., Manion M., Manolio T. A., Mardis E. R., Merker J. D., Rajeevan M. S., Reese M. G., Rehm H. L., Simen B. B., Yeakley J. M., Zook J. M., Lubin I. M. (2012): Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol.* **30**(11), 1033-1036.

Garsmeur O., Droc G., Antonise R., Grimwood J., Potier B., Aitken K., Jenkins J., Martin G., Charron C., Hervouet C., Costet L., Yahiaoui N., Healey A., Sims D., Cherukuri Y., Sreedasyam A., Kilian A., Chan A., Van Sluys M. A., Swaminathan K., Town C., Bergès H., Simmons B., Glaszmann J. C., van der Vossen E., Henry R., Schmutz J., D'Hont A. (2018): A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nat Commun* **9**(1), 2638.

Goodwin S., McPherson J. D., McCombie W. R. (2016): Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, 333 – 351.

Govindaraj M., Vetriventhan M., Srinivasan M. (2015): Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genet Res Int*. 431487.

Guajardo V., Solís S., Almada R., Saski C., Gasic K., Moreno M. Á. (2020): Genome-wide SNP identification in Prunus rootstocks germplasm collections using Genotyping-by-Sequencing: phylogenetic analysis, distribution of SNPs and prediction of their effect on gene function. *Sci Rep*. **10**(1):1467.

Häkkinen, M. (2013): Reappraisal of sectional taxonomy in *Musa* (*Musaceae*). *Taxon* **62**, 809–813.

Head S. R., Komori H. K., LaMere S. A., Whisenant T., Van Nieuwerburgh F., Salomon D. R., Ordoukhanian P. (2014): Library construction for next-generation sequencing: overviews and challenges. *Biotechniques* **56**(2), 61-64.

Hoadley K. A., Yau C., Wolf D. M., Cherniack A. D., Tamborero D., Ng S., Leiserson M. D., Niu B., McLellan M. D., Uzunangelov V., Zhang J., Kandoth C., Akbani R., Shen H., Omberg L., Chu A., Margolin A. A., van't Veer L. J., LopezBigas N., Laird P. W., Raphael B. J., Ding L., Robertson A. G., Byers L. A., Mills G. B., Weinstein J. N., Waes C. V., Chen Z., Collisson E. A., Benz C. C., Perou C. M., Stuart J. M. (2014): Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**(4), 929.

Huang X., Feng Q., Qian Q., Zhao Q., Wang L., Wang A., et al. (2009): High-throughput genotyping by whole-genome resequencing. *Genome. Res* **19**, 1068 – 1076.

Huang X. F., Wu J., Lv J. N., Zhang X., Jin Z. B. (2015): Identification of false-negative mutations missed by next-generation sequencing in retinitis pigmentosa patients: a complementary approach to clinical genetic diagnostic testing. *Genet Med*. **17**(4), 307-311.

Cheesman, E. E. (1947): Classification of the bananas. *Kew Bull*. **2**, 97 – 117.

Chen G., Qiu Y., Zhuang Q., Wang S., Wang T., Chen J., Wang K. (2018): Next-generation sequencing library preparation method for identification of RNA viruses on the Ion Torrent Sequencing Platform. *Virus Genes* **54**(4), 536-542.

Chen K., Wallis J. W., Mclellan M. D., Larson D. E., Kalicki J. M., Pohl C. S., et al. (2009): BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681.

Christelová P., De Langhe E., Hřibová E., Čížková J., Sardos J., Hušáková M., Van den Houwe I., Sutanto A., Kepler A.K., Swennen R., Roux N., Doležel J. (2017): Molecular and cytological characterization of the global *Musa* germplasmcollection provides insights into the treasure of banana diversity. *Biodivers Conserv*. **26**, 801-824.

Illumina, Inc. (2010): Illumina Sequencing Technology Highest data accuracy, simple workflow, and a broad range of applications. Pub. No. 770-2007-002. https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Illumina, Inc. (2019): NovaSeq™ 6000 Sequencing System. Pub. No. 770-2016-025-L QB. https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/novaseq-6000-system-specification-sheet-770-2016-025.pdf

Jain M., Fiddes I. T., Miga K. H., Olsen H. E., Paten B., Akeson M. (2015): Improved data analysis for the MinION nanopore sequencer. *Nature Methods* **12**, 351 – 356.

Jain M., Koren S., Miga K. H., Quick J., Rand A. C., Sasani T. A., Tyson J. R., Beggs A. D., Dilthey A. T., Fiddes I. T., Malla S., Marriott H., Nieto T., O'Grady J., Olsen H. E., Pedersen B. S., Rhie A., Richardson H., Quinlan A. R., Snutch T. P., Tee L., Paten B., Phillippy A. M., Simpson J. T., Loman N. J., Loose M. (2018): Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. **36**(4):338 -345.

Jain M., Tyson J. R., Loose M., Ip C. L. C., Eccles D. A., O'Grady J., Malla S., Leggett R. M., Wallerman O., Jansen H. J., Zalunin V., Birney E., Brown B. L., Snutch T. P., Olsen H. E. (2017): MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Res*. **6**, 760.

Janevski A., Varadan V., Kamalakaran S., Banerjee N., Dimitrova N. (2012): Effective normalization for copy number variation detection from whole genome sequencing. *BMC Genomics* **13**, S16.

Janssen A., Breuer G. A., Brinkman E. K., van der Meulen A. I., Borden S. V., van Steensel B., Bindra R. S., LaRocque J. R., Karpen G. H. (2016): A single double-strand break system reveals repair dynamics and mechanisms in heterochromatin and euchromatin. *Genes Dev*. **30**(14), 1645-1657.

Jeffries D. L., Copp G. H., Lawson Handley L., Olsén K. H., Sayer C. D., Hänfling B. (2016): Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. *Mol Ecol*. **25**(13), 2997-3018.

Jeger M. J., Eden-Green S., Thresh J. M., Johanson A., Waller J.M., Brown A.E. (1995): Banana diseases. V: Goven S (ed.): Bananas and Plantains. London: *Chapman and Hall*; 317 - 381.

Jeridi M., Perrier X., Rodier-Goud M., Ferchichi A., D'Hont A., Bakry F. (2012): Cytogenetic evidence of mixed disomic and polysomic inheritance in an allotetraploid (AABB) *Musa* genotype. *Ann Bot.* **110**(8):1593–1606.

Jiang Y., Wang Y., Brudno M. (2012): PRISM: pair-read informed splitread mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics* **28**, 2576–2583.

Kagy V., Wong M., Vandenbroucke H., Jenny C., Dubois C., Ollivier A., Cardi C., Mournet P., Tuia V., Roux N., Doležel J., Perrier X. (2016): Traditional Banana Diversity in Oceania: An Endangered Heritage. *PLoS One* **11**(3), e0151208.

Kamura D., Karamura e. B., Blomme G. (2011): General plant morphology of *Musa*. *Banana Breeding*.

Kamaté, K., Brown, S., Durand, P., Bureau, J. M., De Nay, D., and Trinh, T. H. (2001): Nuclear DNA content and base composition in 28 taxa of *Musa*. *Genome*. **44**, 622–627. doi: 10.1139/g01-058.

Kitavi M., Downing T., Lorenzen J., Spillane Ch., (2016): The triploid East African Highland Banana (EAHB) genepool is genetically uniform arising from a single ancestral clone that underwent population expansion by vegetative propagation January. *Theoretical and Applied Genetics* **129** (3) DOI: 10.1007/s00122-015-2647-1.

Korbel J. O., Abyzov A., Mu X. J., Carriero N., Cayting P., Zhang Z., Snyder M., Gerstein M. B. (2009): PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol*. **10**(2), R23.

Korbel J. O., Urban A. E., Affourtit J. P., Godwin B., Grubert F., Simons J. F., Kim P. M., Palejev D., Carriero N.J., Du L., Taillon B. E., Chen Z., Tanzer A., Saunders A. C., Chi J., Yang F., Carter N. P., Hurles M. E., Weissman S. M., Harkins T. T., Gerstein M. B., Egholm M., Snyder M. (2007): Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426.

Langmead B., Trapnell C., Pop M., Salzberg S. L. (2009): Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. **10**(3), R25.

Leache A., Oaks J. R. (2017): The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. *Annu. Rev. Ecol. Evol. Syst*. **48**, 69–84.

Lebl M., Buermann D., Reed M. T., Heiner D. L., Triener A. (2012): Flow cells and manifolds having an electroosmotic [Google Patents].

Lee H. C, Lai K., Lorenc M. T., Imelfort M., Duran C., Edwards D. (2011): Bioinformatics tools and databases for analysis of next-generation sequence data. *Brief Funct Genomics* **11**, 12 - 24.

Li H. (2009): The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079

Li H., Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**:1754-60.

Li H., Ruan J., Durbin R. (2008): Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851–1858.

Li Q., Zhao X., Zhang W., Wang L., Wang J., Xu D., Mei Z., Liu Q., Du S., Li Z., Liang X., Wang X., Wei H., Liu P., Zou J., Shen H., Chen A., Drmanac S., Liu J. S., Li L., Jiang H., Zhang Y., Wang J., Yang H., Xu X., Drmanac R., Jiang Y. (2019): Reliable multiplex sequencing with rare index mis-assignment on DNB-based NGS platform. *BMC Genomics* **20**(1), 215.

Li Z., Wang Y., Wang F. (2018): A study on fast calling variants from next-generation sequencing data using decision tree. *BMC Bioinformatics* **19**, 145.

Lian S., Li Q., Dai Z., Xiang Q., Dai X. (2014): A de novo genome assembly algorithm for repeats and nonrepeats. *Biomed Res Int*, 736473.

Lysák, M. A., Doleželová, M., Horry, J. P., Swennen, R., and Doležel, J. (1999): Flow cytometric analysis of nuclear DNA content in *Musa*. *Theor. Appl. Genet.* **98** (8), 1344–1350. doi: 10.1007/s001220051201.

Malinsky M., Trucchi E., Lawson D. J., Falush D. (2018): RADpainter and fineRADstructure: Population Inference from RADseq Data. *Mol Biol Evol*. **35**(5):1284-1290.

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 - 353.

Martin, G., Baurens F. C., Droc G., Rouard M., Cenci A., Kilian A, Hastie A, Doležel J., Aury J. M., 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 Genom*. **17**, 243.

Martin G., Carreel F., Coriton O., Hervouet C., Cardi C., Derouault P., Roques D., Salmon F., Rouard M., Sardos J., Labadie K., Baurens F. C., D'Hont A. (2017): Evolution of the Banana Genome (*Musa acuminata*) Is Impacted by Large Chromosomal Translocations. *Mol Biol Evol*. **34**(9), 2140 - 2152.

Mbida C. M., Van Neer W., Doutrelepont H., Vrydaghs L. (2000): Evidence for banana cultivation and animal husbandry during the first millennium BC in the forest of southern Cameroon. *Journal of Archaeological Science* **27**, 151-162.

McCouch S. R., Zhao K., Wright M., Tung C. W., Ebana K., Thomson M., Reynolds A., Wang D., DeClerck G., Liakat Ali L., McClung A., Eizenga G., Bustamante C. (2010): Development of genome-wide SNP assays for rice. *Breeding Science* **60**, 524 – 535.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M. A. (2010): The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**(9), 1297-303.

Mills R. E., Walter K., Stewart C., Handsaker R. E., Chen K., Alkan C., Abyzov A., Yoon S. C., Ye K., Cheetham R. K., Chinwalla A., Conrad D. F., Fu Y., Grubert F., Hajirasouliha I., Hormozdiari F., Iakoucheva L. M., Iqbal Z., Kang S., Kidd J. M., Konkel M. K., Korn J., Khurana E., Kural D., Lam H. Y., Leng J., Li R., Li Y., Lin C. Y., Luo R., Mu X. J., Nemesh J., Peckham H. E., Rausch T., Scally A., Shi X., Stromberg M. P., Stütz A. M., Urban A. E., Walker J. A., Wu J., Zhang Y., Zhang Z. D., Batzer M. A., Ding L., Marth G. T., McVean G., Sebat J., Snyder M., Wang J., Ye K., Eichler E. E., Gerstein M. B., Hurles M. E., Lee C., McCarroll S. A., Korbel J. O., 1000 Genomes Project. (2011): Mapping copy number variation by population-scale genome sequencing. *Nature* **470**(7332), 59-65.

Nijkamp J. F., Van Den Broek M. A., Geertman J. M., Reinders M. J., Daran J. M., De Ridder D. (2012): De novo detection of copy number variation by co-assembly. *Bioinformatics* **28**, 3195–3202.

Nyine M., Uwimana B., Blavet N., Hřibová E., Vanrespaille H., Batte M., Akech V., Brown A., Lorenzen J., Swennen R., Doležel J. (2018): Genomic Prediction in a Multiploid Crop: Genotype by Environment Interaction and Allele Dosage Effects on Predictive Ability in Banana. *Plant Genome* **11**(2).

Paul, J. Y., Khanna H., Kleidon J., Hoang P., Geijskes J., Daniells J., Zaplin E., Rosenberg Y., James A., Mlalazi B., Deo P., Arinaitwe G., Namanya P., Becker D., Tindamanyire J., Tushemereirwe W., Harding R., Dale J. ()2017: Golden bananas in the field: elevated fruit pro-vitamin A from the expression of a single banana transgene. *Plant Biotechnol. J*. **15**, 520 – 532.

Perrier, X., De Langhe E., Donohue M., Lentfer C., Vrydaghs L., Bakry F., Carreel F., Hippolyte I., Horry J. P., Jenny Ch., Lebot V., Risterucci A. M., Tomekpe K., Doutrelepont H., Ball T., Manwaring J., de Maret P., Denham T. (2011): Multidisciplinary perspectives on banana (*Musa spp*.) domestication. *Proceedings of the National Academy of Sciences* **108**.28: 11311-11318.

Pirooznia M., Goes F. S., Zandi P. P. (2015): Whole-genome CNV analysis: advances in computational approaches. *Frontiers in genetics* **6**, 138.

Płoski R. (2016): Next Generation Sequencing - General Information about the Technology, Possibilities, and Limitations. *Clinical Applications for Next-Generation Sequencing*. 1-18.

Poplin R., Ruano-Rubio V., DePristo M. A., Fennell T. J., Carneiro M. O., Van der Auwera G. A., Kling D. E., Gauthier L. D., Levy-Moonshine A., Roazen D., Shakir K., Thibault J., Chandran S., Whelan C., Lek M., Gabriel S., Daly M. J., Neale B., MacArthur D. G., Banks E. (2017): Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, doi: https://doi.org/10.1101/201178.

Pushkarev D., Neff N. F., Quake S. R. (2009): Single-molecule sequencing of an individual human genome. *Nature Biotechnology* 27, 847 – 850.

Rimmer A., Phan H., Mathieson I., Iqbal Z., Twigg S. R. F.; WGS Consortium, Wilkie A. O. M., McVean G., Lunter G. (2014): Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet*. **46**, 912-918.

Ruas M., Guignon V., Sempere G., Sardos J., Hueber Y., Duvergey H., Andrieu A., Chase R., Jenny C., Hazekamp T., Irish B., Jelali K., Adeka J., Ayala-Silva T., Chao C.P., Daniells J., Dowiya B., Effa Effa B., Gueco L., Herradura L., Obobondji L., Kempenars E., Kilangi J., Muhangi S., Ngo Xuan P., Paofa J., Pavis C., Tiemele D., Tossou C., Sandovaů J., Sutanto A., Vangu Paka G., Van den Houwe I., Roux N., Rouard M. (2017): MGIS: managing banana (*Musa* spp.) genetic resources information and high-throughput genotyping data. Database (Oxford). doi: 10.1093/database/bax046.

Salzberg S. L., Yorke J. A. (2005): Beware of mis-assembled genomes. *Bioinformatics* **21**, 4320 – 4321.

Sand C. (1989): Petite histoire du peuplement de l'Océanie. *Publications de l'Université du Pacifique* **1**, 39-40.

Sanger F., Coulson A. R. (1975): A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**(3), 441 – 448.

Sardos J., Perrier X., Doležel J., Hřibová E., Christelová P., Van den Houwe I., Kilian A., Roux N. (2016a): DArT whole genome profiling provides insight on the evolution and taxonomy of edible Banana (*Musa* spp.). *Ann Bot*.**118**(7), 1269-1278.

Sardos J., Rouard M., Hueber Y., Cenci A., Hyma K.E., Van den Houwe I., Hřibová E., Courtois B., Roux N. (2016b): A gnome-wide association study on the seedless phenotype in banana (Musa spp.) reveals the potential of a selected panel to detect candidate genes in a vegetatively propagated crop. *PLoS One* **11**(5):e0154448.

Seitz V., Schaper S., Dröge A., Lenze D., Hummel M., Hennig S. (2015): A new method to prevent carry-over contaminations in two-step PCR NGS library preparations. *Nucleic Acids Res.* **43**(20), e135.

Schadt E. E., Turner S., Kasarskis A. (2010): A window into third-generation sequencing. *Human Molecular Genetics* **19**, 227 - 240.

Schneeberger K., Ossowski S., Ott F., Klein J. D., Wang X., Lanz C., Smith L. M., Cao J., Fitz J., Warthmann N., Henz S. R., Huson D. H., Weigel D. (2011): Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proc Natl Acad Sci U S A* **108**(25), 10249-54.

Schuster S. C. (2008): Next-generation sequencing transforms today's biology. *Nat Methods* **5**, 16 – 18.

Shepherd K. (1999): Cytogenetics of the genus *Musa*. International Network for the Improvement of Banana and Plantain, Montpellier, France.

Simmonds, N. W., Shepherd, K. (1955): The taxonomy and origins of the cultivated bananas. *Journal of the Linnean Society of London, Botany*. **55**, 302 – 312.

Sudmant P. H., Rausch T., Gardner E. J., Handsaker R. E., Abyzov A., Huddleston J., Zhang Y., Ye K., Jun G., Fritz M. H., Konkel M. K., Malhotra A., Stütz A. M., Shi X., Casale F. P., Chen J., Hormozdiari F., Dayama G., Chen K., Malig M., Chaisson M. J. P., Walter K., Meiers S., Kashin S., Garrison E., Auton A., Lam H. Y. K., Mu X. J., Alkan C., Antaki D., Bae T., Cerveira E., Chines P., Chong Z., Clarke L., Dal E., Ding L., Emery S., Fan X., Gujral M., Kahveci F., Kidd J. M., Kong Y., Lameijer E. W., McCarthy S., Flicek P., Gibbs R. A., Marth G., Mason C. E., Menelaou A., Muzny D. M., Nelson B. J., Noor A., Parrish N. F., Pendleton M., Quitadamo A., Raeder B., Schadt E. E., Romanovitch M., Schlattl A., Sebra R., Shabalin A. A., Untergasser A., Walker J. A., Wang M., Yu F., Zhang C., Zhang J., Zheng-Bradley X., Zhou W., Zichner T., Sebat J., Batzer M. A., McCarroll S. A., 1000 Genomes Project Consortium, Mills R. E., Gerstein M. B., Bashir A., Stegle O., Devine S. E., Lee C., Eichler E. E., Korbel J. O. (2015): An integrated map of structural variation in 2,504 human genomes. *Nature* **526**(7571), 75-81.

Sun X., Liu D., Zhang X., Li W., Liu H., Hong W., Jiang C., Guan N., Ma C., Zeng H., Xu C., Song J., Huang L., Wang C., Shi J., Wang R., Zheng X., Lu C., Wang X., Zheng H. (2013): SLAF-seq: an efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing. *PLoS One* **8**, e58700.

Swennen R. (1990): Limits of morphotaxonomy: names and synonyms of plantain in Africa and elsewhere. *The identification of genetic diversity in the genus Musa. Proceedings of an International Workshop*. 172 – 210.

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.*). *Front Plant Sci.* **10**, 1503.

Teo S. M., Pawitan Y., Ku C. S., Chia K. S., Salim A. (2012): Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* **28**, 2711–2718.

Tewhey R., Warner J. B., Nakano M., Libby B., Medkova M., David P. H. (2009): Microdroplet-based PCR Enrichment for large-scale targeted sequencing. *Nat Biotechnol* **27**(11), 1025 – 1094.

Tézenas du Montcel H., De Langhe E., Swennen R. (1983): Essai de classification des bananiers plantains (AAB). *Fruits* **38**, 461 – 474.

The 1000 Genomes Project Consortium (2012): An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422), 56.

Trebbi D., Ravi S., Broccanello C., Chiodi C., Francis G., Oliver J., Mulpuri S., Srinivasan S., Stevanato P. (2019): Identification and validation of SNP markers linked to seed toxicity in *Jatropha curcas* L. *Sci Rep.* **9**(1), 10220.

Tsai K. J., Jade Lu M. Y., Yang K. J., Li M., Teng Y., Chen S., Ku M. S. B., Li W. H. (2016): Assembling the Setaria italica L. Beauv. genome into nine chromosomes and insights into regions affecting growth and drought tolerance. *Scientific Reports* **6**, 35076.

Uitdewilligen J. G., Wolters A. M., D'hoop B. B., Borm T. J., Visser R. G., van Eck H. J. (2013): A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS One* **8**(5), e62355.

van Dijk E. L., Jaszczyszyn Y., Naquin D., Thermes C. (2018): The Third Revolution in Sequencing Technology. *Trends Genet*. **34**(9), 666 - 681.

Vilgis S., Deigner P. (2018): Sequencing in Precision Medicine. *Precision Medicine* kapitola **5**.

Voelkerding K. V., Dames S. A., Durtschi J. D. (2009): Next-generation sequencing: from basic research to diagnostics. *Clin Chem.* **55**(4), 641-658.

Voskoboynik A., Neff N. F., Sahoo D., Newman A. M., Pushkarev D., Koh W., Passarelli B., Fan H. C., Mantalas G. L., Palmeri K. J., Ishizuka K. J., Gissi C., Griggio F., Ben-Shlomo R., Corey D. M., Penland L., White R. A., Weissman I. L., Quake S. R. (2013): The genome sequence of the colonial chordate, *Botryllus schlosseri*. *Elife*. **2**: e00569.

Wang Z., Miao H., Liu J., Xu B., Yao X., Xu C., Zhao S., Fang X., Jia C., Wang J., Zhang J., Li J., Xu Y., Wang J., Ma W., Wu Z., Yu L., Yang Y., Liu C., Guo Y., Sun S., Baurens F. C., Martin G., Salmon F., Garsmeur O., Yahiaoui N., Hervouet C., Rouard M., Laboureau N., Habas R., Ricci S., Peng M., Guo A., Xie J., Li Y., Ding Z., Yan Y., Tie W., D'Hont A., Hu W., Jin Z. (2019): Musa balbisiana genome reveals subgenome evolution and functional divergence. *Nature Plants* **5**, 810-821.

Weischenfeldt J., Symmons O., Spitz F., Korbel J. O. (2013): Phenotypic impact of genomic structural variation: insights from and for human disease. Nature Reviews Genetics 14, 125 – 138.

Wondji C. S., Hemingway J., Ranson H. (2007): Identification and analysis of single nucleotide polymorphisms (SNPs) in the mosquito Anopheles funestus, malaria vector. *BMC Genomics* **8**, 5.

Wu C. G., Spies M. (2013): Overview: what are helicases? *Adv Exp Med Biol*. **767**, 1 - 16.

Xi R., Hadjipanayis A. G., Luquette L. J., Kim T. M., Lee E., Zhang J., Johnson M. D., Muzny D. M., Wheeler D. A., Gibbs R. A., Kucherlapati R., Park P. J. (2011): Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci U S A* **108**(46), E1128-36.

Xi R., Lee S., Park P. J. (2012): A survey of copy-number variation detection tools based on high-throughput sequencing data. *Curr. Protoc. Hum. Genet. Chap*. **7**, 19.

Xie C., Tammi M. T. (2009): CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinform*. **10**, 80.

Xu C. (2018): A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and structural biotechnology journal* **16**, 15 – 24.

Ye K., Hall G., Ning Z. (2016): Structural Variation Detection from Next Generation Sequencing. *Next Generat Sequenc & Applic* S1:007.

Ye K., Schulz M. H., Long Q., Apweiler R., Ning Z. (2009): Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871.

Zarrei M., MacDonald J. R., Merico D., Scherer S. W. (2015): A copy number variation map of the human genome. *Nat Rev Genet.* **16**(3), 172-183.

Zarrella I., Herten K., Maes G. E., Tai S., Yang M., Seuntjens E., Ritschard E. A., Zach M., Styfhals R., Sanges R., Simakov O., Ponte G., Fiorito G. (2019): The survey and reference assisted assembly of the *Octopus vulgaris* genome. *Sci Data* **6**(1), 13.

Zhang Z. D., Du J., Lam H., Abyzov A., Urban A. E., Snyder M., Gerstein M. (2011): Identification of genomic indels and structural variations using split reads. *BMC Genomics*. **12**, 375.

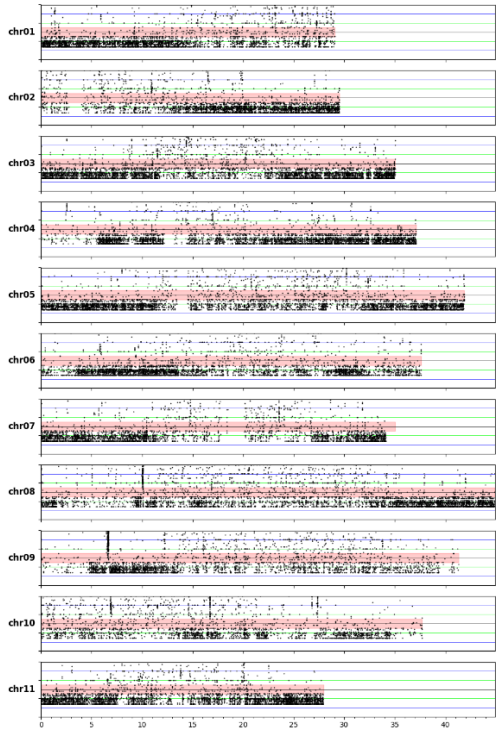
Zhao M., Wang Q., Wang Q., Jia P., Zhao Z. (2013): Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinform*. **14**, S1.

7 PŘÍLOHY

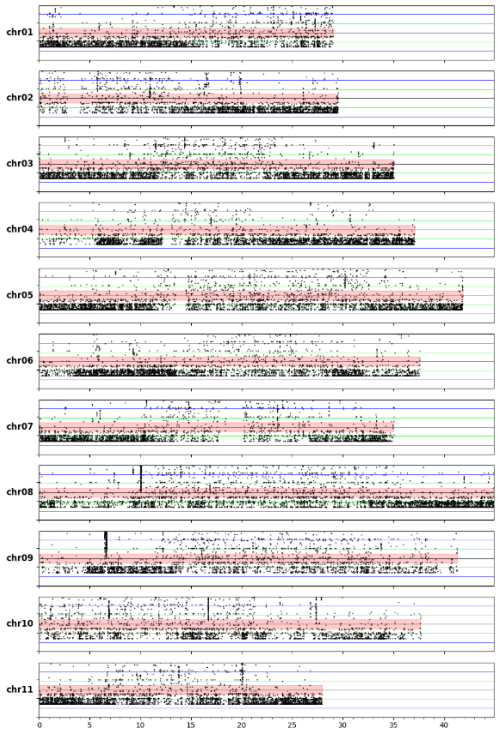
Příloha 1 – vykreslení pokrytí chromozomů, plantainy typu French

Jedná se o podíl: počet čtení nesoucích danou SNP pozici / průměrný počet čtení vypočítaný ze všech pozic; (normalizované pokrytí).

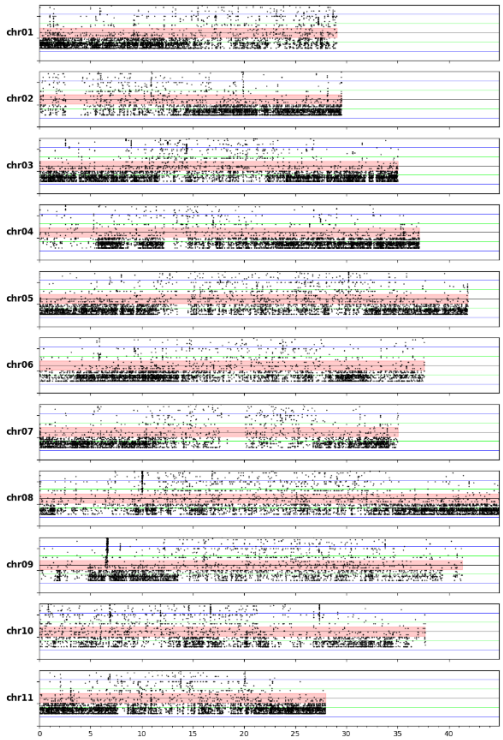
ITC.0109; Obino l'Ewai



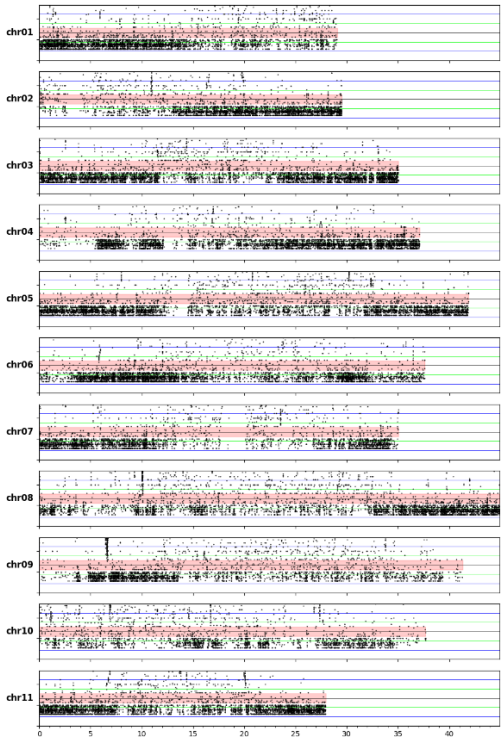
ITC.0142; Msisa



ITC.0219; Apem Pa



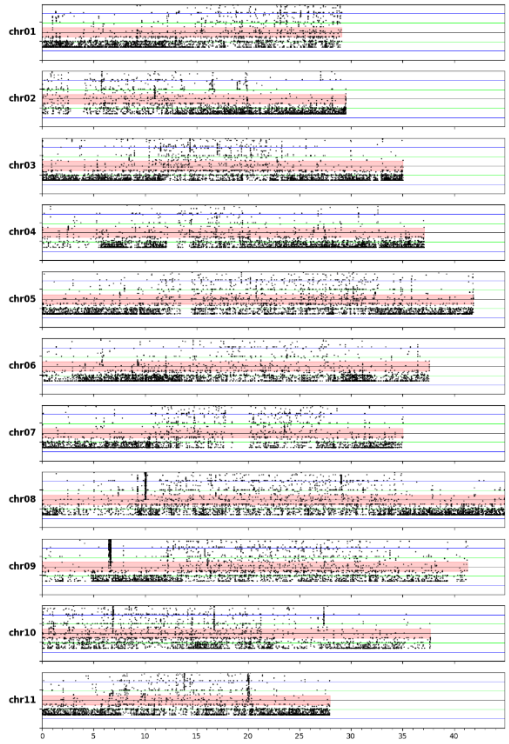
ITC.0325; Wine Plantain



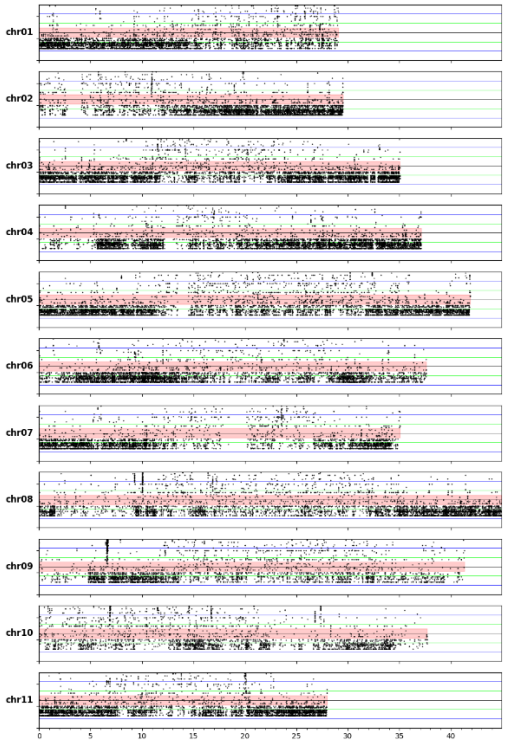
Příloha 2 – vykreslení pokrytí chromozomů, plantainy typu False horn

Jedná se o podíl: počet čtení nesoucích danou SNP pozici / průměrný počet čtení vypočítaný ze všech pozic; (normalizované pokrytí).

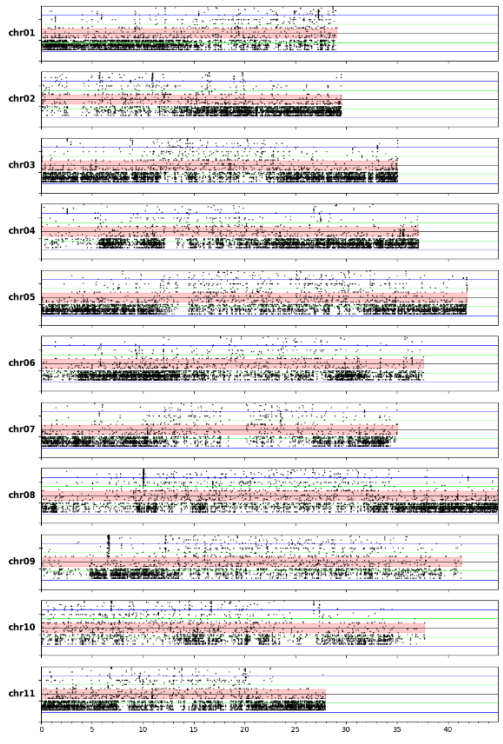
ITC.0098; Baka



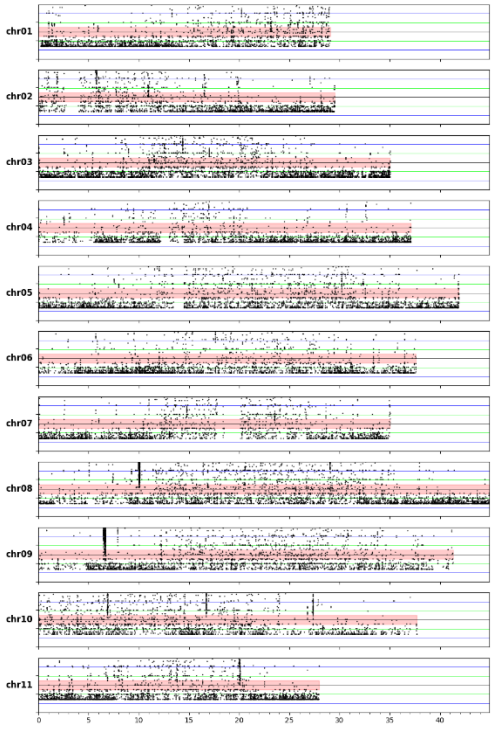
ITC.0223; Apantu



ITC.0517; Orishele



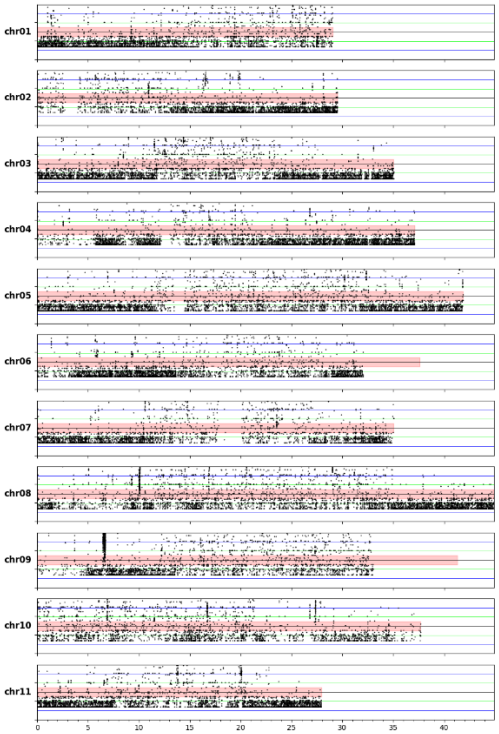
ITC.0641; Dominico Rojo



Příloha 3 – vykreslení pokrytí chromozomů, plantainy typu Horn

Jedná se o podíl: počet čtení nesoucích danou SNP pozici / průměrný počet čtení vypočítaný ze všech pozic; (normalizované pokrytí).

ITC.0185; 3 Hands Planty



ITC.0128; Tshambunu