**Jihočeská univerzita v Českých Budějovicích**

**Přírodovědecká fakulta**

# Sestavení genomu cichlidy *Crenicichla semifasciata* kombinací sekvencí z různých sekvenačních platforem

Diplomová práce

**Bc. Daniela Kotalová**

Školitel: Piálek Lubomír, RNDr. Ing. Ph.D.

České Budějovice 2022

# Bibliografické údaje

Kotalová D., 2022: Sestavení genomu cichlidy *Crenicichla semifasciata* kombinací sekvencí z různých sekvenačních platforem. [Genome assembly of the cichlid *Crenicichla semifasciata* in combination of sequences from different sequencing platforms. Mgr. Thesis, in Czech.] - 42 p., Faculty of Science. University of South Bohemia, České Budějovice, Czech Republic.

# Anotace

By a combination of sequencing data from two platforms 10x Genomics and Oxford Nanopore Technologies, this study creates a de novo reference genome of *Crenicichla semifasciata*. Data were processed by both individual and hybrid assembly. Contiguity and genome completeness evaluations were determined for all assemblies.

# Prohlášení

Prohlašuji, že jsem autorem této kvalifikační práce a že jsem ji vypracovala pouze s použitím pramenů a literatury uvedených v seznamu použitých zdrojů.

V Českých Budějovicích dne 8. 12. 2022

# Poděkování

Tímto bych chtěla poděkovat svému školiteli RNDr. Ing. Lubomírovi Piálkovi, Ph.D. za jeho velmi cenné rady a odborné vedení, ale také za jeho vstřícnost, trpělivost a za čas, který mi věnoval. Děkuji také svým blízkým a přátelům, kteří mě při studiu podporovali a byli mi vždy oporou.

Dále bych chtěla poděkovat Metacentru, které jsem mohla bezplatně využívat. Výpočetní zdroje byly dodány v rámci projektu "e-Infrastruktura CZ" (e-INFRA CZ M2018140) podpořeného Ministerstvem školství, mládeže a tělovýchovy ČR.

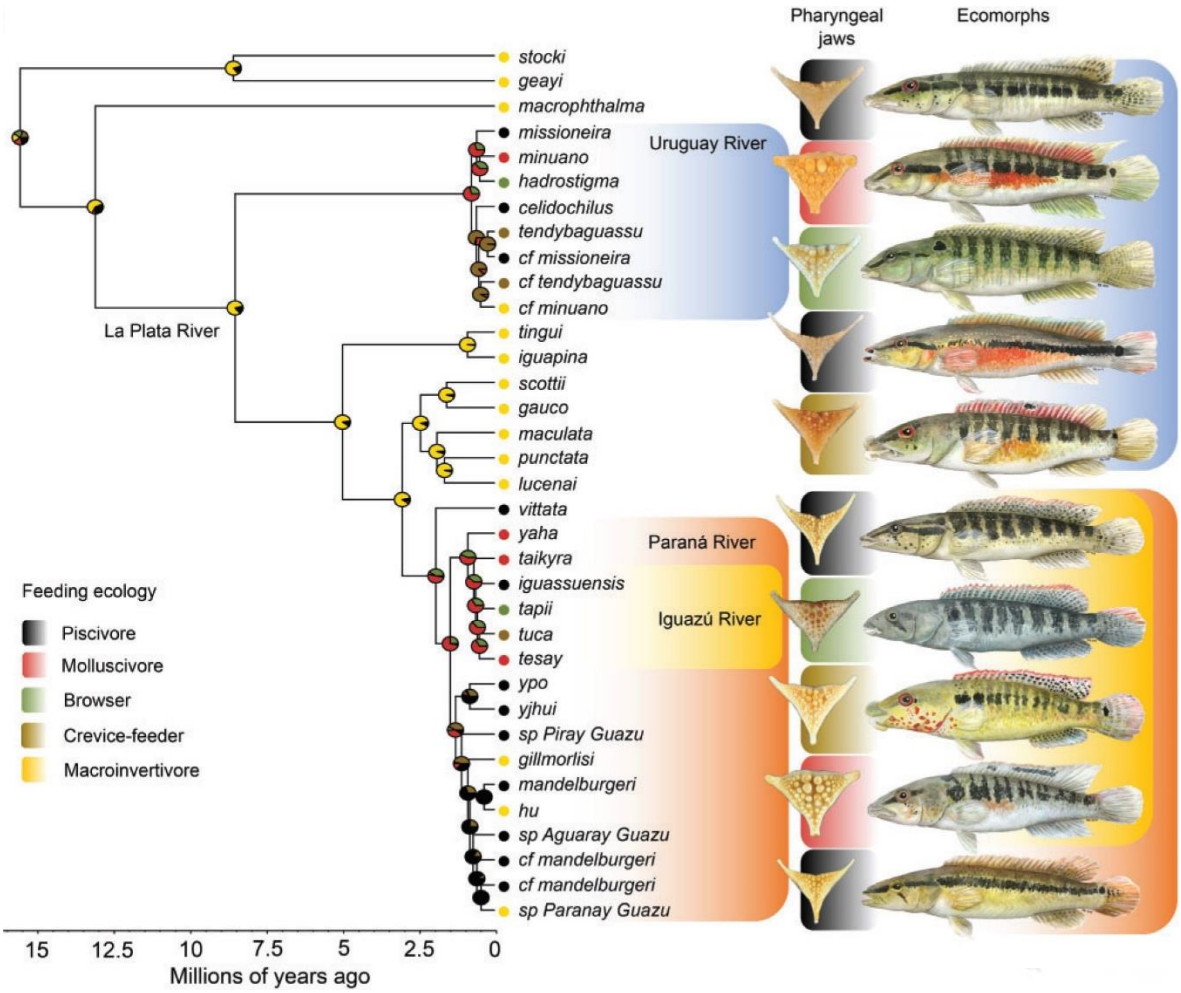
# Obsah

# 1. Úvod

## 1.1. Studium paralelní diverzifikace u cichlid

Jedním z tradičních modelů pro studium nejrůznějších evolučních jevů jsou cichlidy velkých afrických jezer. Díky obrovské druhové rozmanitosti vzniklé v historicky krátkém čase v obdobných habitatech poskytují africké cichlidy jedinečný materiál např. pro studium sympatrické evoluce a paralelní diverzifikace. Jak nedávno prokázaly práce našeho týmu, mechanismy paralelní evoluce se ale dají studovat v menším měřítku i na jihoamerických cichlidách a v říčních systémech. Konkrétně jde o evoluci dvou nepříbuzných druhových hejn rodu *Crenicichla* vyskytujících se v povodí La Plata (Burress et al., 2018; Piálek et al., 2012, 2019). Druhové hejno je skupina blízce příbuzných druhů, které vznikly z jednoho předka, žijí v sympatrii a liší se svou ekomorfologií. Změna morfologie těla proběhla na základě různé potravy, kterou ryby konzumují. Studovaná druhová hejna pocházejí ze dvou různých povodí, a to řeky Uruguay a střední Paraná. Bylo zjištěno, že i když druhová hejna nejsou geneticky příbuzná, tak se v obou řekách vyskytují stejné morfologické typy dle typu přijímané potravy (Obr. 1).

Při analýze mechanismů evoluce uvnitř druhových hejn bylo třeba použít dostatečně citlivé genetické markery, neboť v důsledku nedávné evoluce existuje mezi druhy malá genetická variabilita. Studium fylogeneze také znesnadňují jevy jako např. ancestrální polymorfismus, hybridizace a introgrese nebo disperze. Tradiční mitochondriální markery (cytochrom *b* a ND2) bohužel neposkytly věrohodnou evoluční hypotézu. Dalším krokem proto bylo genotypování pomocí techniky ddRAD (Double digest restriction-site associated DNA), která díky sekvenování náhodných homologických markerů napříč genomem (v objemu ca. 1 % celého genomu) práci významně posunula (Burress et al., 2022). Pro další pokročilé analýzy zahrnující např. definici oblastí genomu, které hrají významnou úlohu při selekci sledovaných znaků jsme se rozhodli sekvenovat u reprezentativního vzorku jedinců celý genom.




Obr. 1: Morfologické typy vyskytující se ve druhových hejnech rodu *Crenicichla* v povodí Río de la Plata (Burress et al., 2022)

Zpracování celogenomových (WGS; „whole genome sequencing“) dat vyžaduje evolučně blízký referenční genom, na který by bylo možné namapovat genomy jedinců z druhových hejn. Nejbližším publikovaným genomem v databázi Genbank byl ale pouze genom neotropické cichlidy rodu *Amphilopus*, na který by bylo možné spolehlivě namapovat pouze 50 % sekvencí. Proto bylo nutné nejprve osekvenovat a sestavit kvalitní a blízký genom ze stejného rodu, který se bude využívat, jako referenční. Pro tyto účely jsme využili druh *Crenicichla semifasciata* (Heckel, 1840) z důvodu podobné evoluční vzdálenosti od obou druhových hejn.

## 1.2. Historie sekvenování DNA

Od roku 1953, kdy James Watson a Francis Crick objevili základní dvoušroubovicovou strukturu DNA (Watson & Crick, 1953), vynaložili vědci spoustu úsilí k vytvoření metod jejichž prostřednictvím by se sekvence jejích molekul daly rutinně číst. První přečtenou sekvencí byl však v roce 1965 úsek alaninové tRNA kvasinky pivní (*Saccharomyces cerevisiae*; Holley et al., 1965). Bylo mnohem jednodušší začít se sekvenováním RNA, jelikož se vyskytuje ve formě jednoho řetězce a je podstatně kratší než dvoušroubovitá DNA. První sekvence DNA byla přečtena až roku 1972 (Min-Jou et al., 1972). Jednalo se o gen kódující "coat protein" (protein tvořící kapsidu) bakteriofága M2.

Kolem roku 1977 byly vyvinuty dvě sekvenační metody - Maxamovo–Gilbertovo (Maxam & Gilbert, 1977) a Sangerovo sekvenování (Sanger et al., 1977), též nazývány jako sekvenování první generace. Sangerova metoda byla mnohem úspěšnější a používá se až do dnešní doby. Je to technika modifikované replikace DNA, která využívá DNA polymerázu, primery, nukleotidy a dideoxyribonukleotidy (dNTPs; Atkinson et al., 1969). Při syntéze DNA se do řetězce začleňují nejen nukleotidy, ale náhodně také dideoxynukleotidy, které syntézu zastaví a vytvoří tak různě dlouhé fragmenty. Vzniklé fragmenty se pak přenesou na gelovou elektroforézu a díky odlišné délce fragmentů je přesně zaznamenána sekvence DNA. Dnes jsou dNTPs značeny fluorescenční sondou, kdy každý dNTPs má svou barvu. Lze tak fragmenty osekvenovat kapilární elektroforézou a díky zaznamenaným amplitudám signálu lze určit sekvenci DNA. Touto metodou se dají osekvenovat celkem malé úseky DNA (maximálně 1000 bazí; Morozova & Marra, 2008) kvůli omezené funkčnosti DNA polymerázy. Sangerova metoda je sice velmi přesná, ale na druhou stranu je časově i finančně náročná. Z těchto důvodů se vymýšlely další postupy, jak zrychlit a zlevnit sekvenování a jak zlepšit jeho produktivitu.

Během 90. let 20. století se začaly vyvíjet metody sekvenování nové generace (next generation sequencing, NGS, sekvenování druhé generace, kam patří například Illumina, Iont Torrent nebo SOLiD. Tyto techniky snížily náklady, urychlily proces sekvenace, a hlavně umožnily masivně paralelní sekvenování. Sekvenování může probíhat buď díky syntéze nebo ligaci, kdy je přítomna DNA polymeráza nebo ligáza (Ju et al., 2006; Pereira et al., 2015). Nejznámější typ NGS je sekvenování syntézou (sequencing by synthesis) prováděný firmou Illumina (Illumina, 2010). Postup je postavený na můstkové PCR

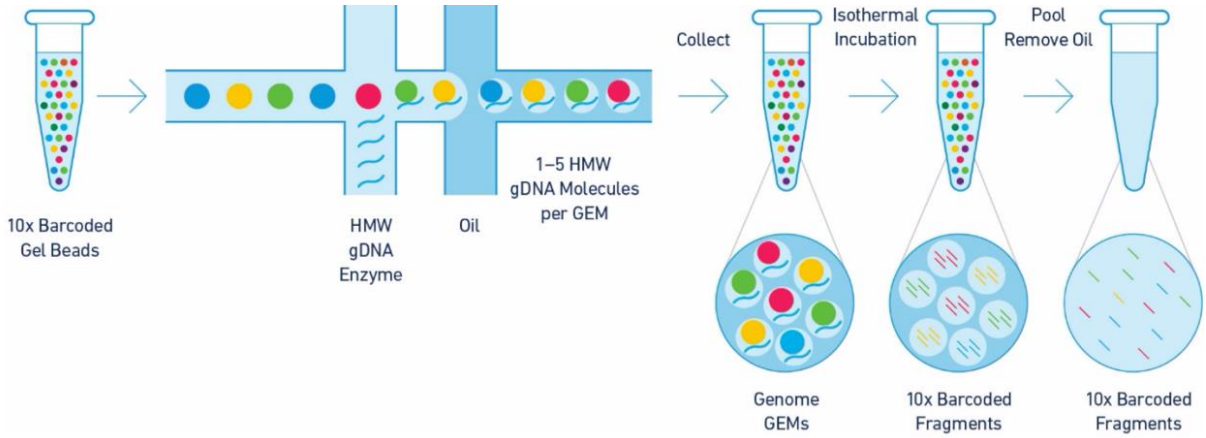
amplifikaci, která probíhá na destičce („flowcell“), následovaná sekvenováním při syntéze DNA. Bohužel metody NGS založené na syntéze jsou limitovány délkou sekvenovaných úseků, kdy maximální délka jednotlivých sekvencí dosahuje 300 bp (páry bází; Illumina, 2022). Velikou výhodou NGS je ale přesnost výsledných krátkých sekvencí. Jejich chybovost dosahuje pouze maximálně 1 % (Glenn, 2011). Hlavní úskalí metod druhé generace spočívá v tom, že je třeba nejdříve amplifikovat nukleovou kyselinu před samotným sekvenováním. Výsledná sekvence je tedy průměrem z mnoha kopií sekvenované molekuly DNA. Bylo tedy žádoucí, aby byla vytvořena metoda, která krok amplifikace obchází.

Nejnovější a nejinovativnější metody na poli sekvenování se nazývají sekvenování třetí generace. Metody používající sekvenování třetí generace jsou například PacBio, SMRT (Single-molecule real-time sequencing) nebo Oxford Nanopore. Oproti NGS je vynechán postup amplifikace a sekvenuje se pouze jediná molekula v reálném čase. Sekvenování trvá zpravidla pouze několik hodin. Kvůli zrychlenému procesu se tyto metody staly o trochu méně přesné, jelikož vzniklé sekvence mají vyšší chybovost („error rate“) až 38,2 % (Laver et al., 2015). Tento problém lze předejít díky vyšší míře prosekvenování. Chyby se vyskytují v sekvencích náhodně a při vícenásobném sekvenování téhož úseku se proto dají z velké části eliminovat. Díky absenci kroku PCR mohou metody pracovat s molekulami obrovských délek a teoreticky není specifikován limit. Dosavadní nejdelší zaznamenaná přečtená sekvence dosahuje délky až 4 Mb (Oxford Nanopore technologies, 2022). K dosažení dlouhých sekvencí je nutná precizní a opatrná práce v laboratoři, jelikož při izolaci se molekuly DNA můžou rozlámat na malé fragmenty kvůli pipetování, centrifugaci nebo i kvůli pouhým trhaným pohybům se zkumavkou. Postup izolace je tedy nutné upravit tak, aby se předešlo přílišnému rozlámání nukleové kyseliny.

## 1.3. 10x Genomics

Roku 2012 přišla společnost 10x Genomics s novou metodou přípravy knihovny pro sekvenování na platformě Illumina. Metoda je založená na emulzní PCR, kdy v olejové kapičce probíhá amplifikace DNA. Klíčovým komponentem této technologie jsou GEMs (Gel bead in EMulsion = gelové kuličky v emulzi). Gelové kuličky se nachází po jedné uvnitř olejových kapiček a obsahují desítky milionů různých oligonukleotidů. Oligonukleotidy kromě technických sekvencí obsahují také adaptéry, primery a hlavně

barkód (čárový kód, indexová sekvence) o délce 14 bazí (Zheng et al., 2017). Každá kulička má jedinečný barkód, takže je snadné pak identifikovat sekvence, které patří k sobě.
V procesu vytváření emulze (Obr. 2) se do kapiček s gelovou kuličkou rozdělují dlouhé kusy DNA. Ihned po uzavření kuličky v olejové kapičce se gelová kulička rozpadá a uvolňuje oligonukleotidy s navázanou DNA do kapičky. V kapičce pak probíhá PCR. Je tak zajištěno, že se různé fragmenty mezi sebou nepromíchají a že se tak seskupí fragmenty původem ze stejného fragmentu DNA. Po proběhlé emulzní PCR se amplifikované fragmenty sekvenují na platformě Illumina. Díky stejným barkódům ze stejného fragmentu DNA se zjednoduší následné zrekonstruování původní molekuly DNA.




Obr. 2: Princip 10x Genomics (UC Davis Bioinformatics, 2018).

## 1.4. Oxford Nanopore

Významnou metodou sekvenování třetí generace je metoda vyvinutá firmou Oxford Nanopore Technologies (ONT), která byla založena roku 2005 (Deamer et al., 2016). Technologie dokáže sekvenovat celý genom, určitý úsek DNA, RNA, miRNA, metagenomiku, a dokonce i proteiny. Dále je možné sekvenovat i epigenetické značky a modifikace bází (Schreiber et al., 2013). Díky tomu, že ONT dokáže sekvenovat jednu molekulu v reálném čase, může produkovat ultradlouhé sekvence. Záleží pouze na délce a kvalitě vstupního materiálu.

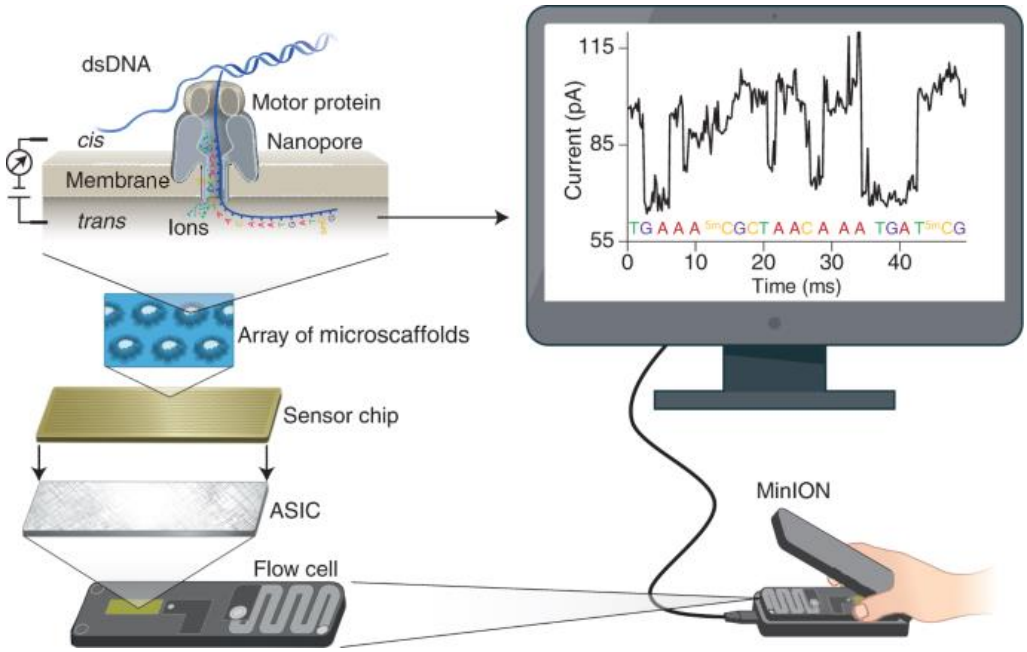
Nejdříve se na sekvenovaný materiál (např. DNA – ideálně s co největšími fragmenty) musí naligovat adaptéry. Adaptéry se nacházejí na obou koncích vyizolované DNA. Na 5´ konec jednoho vlákna nasedne enzym (motorový protein). Ten zajišťuje jednosměrný a přesný chod sekvenování. (Jain et al., 2016).

Princip sekvenování funguje na nanopóru. Je to protein vytvořený ze syntetických polymerů, který je upevněný v elektricky rezistentní membráně. Aplikovaný potenciál pohání iontový proud skrz nanopór (Kasianowicz et al., 1996). Díky motor proteinu, který má helikázovou aktivitu, prochází jedno-vláknová DNA skrz pór. Uvnitř nanopóru je uměle vytvořené vazebné místo, kam se jednotlivé báze přechodně váží a tím se změní i ionový proud. Rozdíly ionového proudu lze detekovat senzorem a zaznamenat pomocí křivky („squiggle“). Každá báze má stanovený čas translokace a tvar křivky - amplitudu a rozptyl (Wang et al., 2004). Průchodem fragmentu nukleové kyseliny nanopórem jsou postupně zaznamenávány báze a tvoří se mnoho sekvencí, ze kterých lze následně rekonstruovat celá molekula DNA („genome assembly“ = sestava genomu).

### 1.4.1. MinION

V roce 2014 vytvořila firma ONT přenosnou verzi sekvenátoru – MinION (Obr. 3.; Deamer et al., 2016). Přístroj obsahuje průtokovou buňku („flowcell“), která obsahuje 512 kanálů. Každý kanál obsahuje 4 póry. Na celé destičce se tedy vyskytuje 2048 jednotlivých pórů, které se používají k sekvenování. V každém kanálu je vždy aktivní pouze jeden nanopór, což umožňuje souběžné sekvenování až 512 molekul. Sekvenování probíhá díky mikročipu ASIC (application specific integrated circuit), který dokáže číst nepatrné změny elektrického proudu uvnitř póru. Rychlost čtení sekvence v reálném čase je obrovská, dosahuje až 420 bází/s a přístroj dokáže během jednoho sekvenačního procesu osekvenovat až 50 Gb (Oxford Nanopore Technologies, 2022b).

Sekvenátor MinION velikosti telefonu váží 90 gramů a tím se stává lehce přenositelným. Díky těmto malým rozměrům je možné sekvenování provádět skoro všude. Je možné sekvenovat DNA nejen v jednotlivých laboratořích, ale i přímo v terénu, a dokonce i ve vesmíru. Díky dostupné ceně si mohou dovolit sekvenování prezentovat i školy v rámci praktických cvičení (Jain et al., 2016).




Obr. 3: Proces Oxford Nanopore technologie a MinION (Y. Wang et al., 2021).

Nejnovější produktem ONT je MinION Mk1C. Je to výkonný dotykový tablet s vysokým rozlišením se zabudovaným sekvenátorem. Má nainstalovány programy pro sekvenování, které zaznamenává v reálném čase a software pro analýzy a vizualizaci výsledků. Je vytvořen hlavně pro práci v terénu, kdy není třeba s sebou brát počítač, bez kterého se jednoduchý MinION neobejde. Navíc váží pouze 420 g, takže je snadno přenositelný. Je možné se k přístroji připojit i bezdrátově pomocí Wi-Fi a nasdílet tak výsledky kdykoliv a kdekoliv (Oxford Nanopore Technologies, 2022b).

## 1.5. Bioinformatické analýzy

### 1.5.1. Zpracování primárních dat

S vyvíjením metod NGS se navyšuje i množství dat, které je třeba zpracovávat pomocí bioinformatických postupů. Po sekvenování vzniknou primární data („raw data“), která je třeba analyzovat a následně upravit („preprocessing“). Rozmanitost úprav se odvíjejí od toho, pro jaké analýzy budou výsledná data použita.

Při sekvenování na platformě Oxford Nanopore je nejprve potřeba zaznamenané elektrické křivky převést na sekvenci příslušných nukleotidů prostřednictvím algoritmu pro určení bází („basecalling“). Pro data z ONT se používá program Guppy, který je vyvinut přímo firmou Oxford Nanopore. Jedná se o program využívající obousměrnou rekurentní neuronovou síť

(Wick et al., 2019), která umožňuje proces „učení se“ správnému určení báze na základě již přečtených sekvencí. Díky tomu se algoritmus překladu s novými verzemi programu neustále zpřesňuje. Nejnovější verze programu Guppy (v.5+) využívá typ modelu „super accuracy“; tato verze je přibližně třikrát přesnější v určování bází než starší mód „high-accuracy“ a chybovost se údajně pohybuje v řádu nízkých jednotek procent (Community - Protocol (nanoporetech.com)). Každý přeložený nukleotid má přiřazenou hodnotu kvality (pravděpodobnosti chyby překladu) a výsledné sekvence se ukládají nejčastěji ve formátu FASTQ, který umožňuje tuto hodnotu pro každou bázi uchovat pro pozdější zpracování.

Dalším krokem úpravy dat je analýza kvality. Ze získaných datových souborů je třeba odfiltrovat nekvalitní nukleotidy/sekvence (tzv. „trimming“). Data mohou obsahovat nekvalitní části, které mohou narušit signál těch správných, a proto je třeba tyto rušivé elementy odstranit pomocí různých algoritmů. Je nutné si ale zachovat určité množství dat, které jsou potřeba k dalším analýzám. Proto by parametry určující kvalitu sekvencí neměly být příliš přísné.

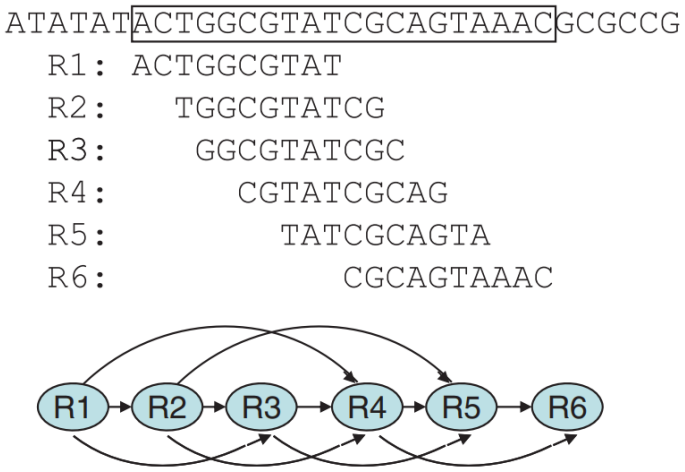
Různé typy sekvenování mají různou chybovost („error rate“) při zaznamenávání sekvencí. Tyto chyby je třeba opravit. Korekce chyb dokáže zvýšit citlivost a kvalitu sekvencí. Algoritmy pro korekci chyb používají různé postupy. Buď se využívají přístupy založené na samokorekci nebo na hybridní korekci, kdy krátké sekvence opravují ty delší (Y. Wang et al., 2021).

### 1.5.2. Sestavení genomu

Bohužel v současné době není možné osekvenovat celý kontinuální chromozóm/genom. Při izolaci DNA se genomová DNA rozláme na různě dlouhé fragmenty, které se následně sekvenují. Když chceme zjistit původní genomovou sekvenci, musíme tyto fragmenty pospojovat procesem tzv. „assembly“. Algoritmy pro sestavování genomu k sobě zarovnávají („align“) sekvence díky jejich překryvům („overlap“) a propojují je. Propojením sekvencí se vytvoří souvislé sekvence překrývajících se fragmentů (kontigy; Green, 2022). Spojením kontigů vzniknou superkontigy („scaffoldy“), které mimo seřazených kontigů obsahují i nepřečtené úseky (gapy) známé délky. Seřazené scaffoldy tvoří výslednou sestavu genomu. Sestava lze vytvořit různými způsoby. Pro rekonstrukci genomu se nejčastěji používají dvě skupiny metod: de Bruijn grafy (DBG) a Overlap layout consensus (OLC).

### 1.5.2.1. Overlap layout consensus

Overlap layout consensus (Obr. 4) je typ skládání genomu, který nejdříve hledá stejné sekvence napříč všemi sekvencemi („all-against-all pair-wise reads aligning“) a následně je přikládá k sobě („overlap“). Dále se všechny sekvence rozloží („layout“), kdy stejné sekvence jsou již seřazeny za sebou na základě návaznosti. Podle tohoto grafu se poté sestrojí nepřetržitá sekvence („consensus“). Graf obsahuje uzly („nodes“) a hrany („edges“). Každé číslo u uzlu reprezentuje počet sekvencí. Hrany zobrazují překryvy mezi stejnými sekvencemi. Překryvy mají podle nastavení parametrů definovaný minimální počet bazí. Počet sekvencí na jednom uzlu se lineárně zvyšuje s hloubkou sekvenování (coverage) a počet hran se zvyšuje o logaritmickou stupnici (Z. Li et al., 2012).

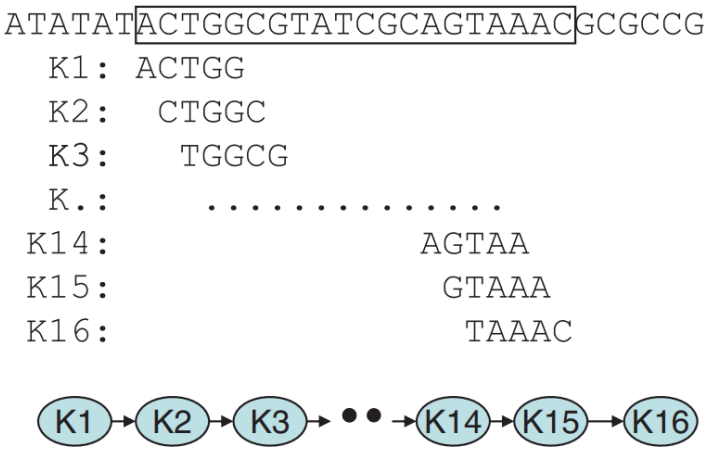



Obr. 4: Overlap Layout Consensus (Z. Li et al., 2012).

První verze algoritmu byla vytvořena již roku 1980 (Staden, 1980). Tato původní verze byla mnoha dalšími vědci rozšířena. V současnosti je mnoho programů, které s OLC pracují. Jsou jimi například Miniasm (H. Li, 2016), Falcon (Chin et al., 2016) a Canu (Koren et al., 2017).

### 1.5.2.2. De Bruijn grafy

De Bruijn grafy (Obr. 5) pracují s k-mery. K-mery jsou krátké sekvence o délce k (jednotky až nízké stovky bazí) do kterých se rozloží původní sekvence. Algoritmus porovnává všechny vytvořené k-mery a určuje jejich sousední vztahy. DBG tvoří také graf, kdy uzly představují k-mery a hrany ukazují vztahy mezi nimi. Při sestavování genomu je nutné každou hranu projít podle Eulerovského tahu (Pevzner et al., 2001) a zaznamenat příslušné k-mery. Takto vzniknou kontigy a následně scaffoldy. K-mery se mohou opakovat (repetice) a mohou tak vzniknout různé varianty sekvencí.




Obr. 5: De Bruijn grafy (Z. Li et al., 2012).

DBG byly poprvé použity k sestavení genomu v roce 2001 (Pevzner et al., 2001), ale algoritmus byl vymyšlen už v roce 1995 (Idury & Waterman, 1995). Programy využívající algoritmy DBG jsou například: SPAdes (Bankevich et al., 2012), ALLPATHS-LG (Gnerre et al., 2011), ABruijn (Lin et al., 2016) a Velvet (Zerbino & Birney, 2008).

#### 1.5.2.3. Sestavení genomu z dlouhých vs. krátkých sekvencí

Jak již bylo zmíněno výše, různé sekvenační metody produkují různě dlouhé sekvence, které při sestavování genomu mají různé výhody. Krátké sekvence, které mají maximální délku 300 bp, jsou produkovány hlavně technikami sekvenování druhé generace (NGS). Hlavním problémem spojování krátkých sekvencí do genomu spočívá v tom, že genom nemusí být sestaven správně; krátké sekvence nemusí vytvořit dostatečné překrytí u OLC nebo DBG algoritmů. Dalším problémem jsou repetice, které krátké sekvence neumí úplně vyřešit. Velkou výhodou krátkých sekvencí je však jejich přesnost (Pearman et al., 2020).

Dlouhé sekvence mají velikost od 300 bp (Midha et al., 2019) a jsou produkovány metodami sekvenování třetí generace. U dlouhých sekvencí je výhodou samotná jejich délka. Díky velkým rozměrům sekvencí je snadnější je spojit dohromady a mizí tím i problém s nedostatečnou mírou překrytí při sestavování genomu. Tím je zajištěno, že výsledný genom je sestaven ve správném pořadí. Na druhou stranu, nevýhodou dlouhých sekvencí se stává náchylnost k chybám (vysoká error rate; Istace et al., 2017). Chyby jsou zde rozmístěny náhodně.

### 1.5.3. Hybridní assembly

Problémy vyplývající z různých vlastností dlouhých i krátkých sekvencí můžeme vyřešit jejich kombinováním, tedy vytvářením tzv. hybridní assembly. Dlouhé, kontinuální a nepřesné sekvence mohou být opraveny krátkými sekvencemi s vysokou mírou přesnosti. Mnoho softwarů založených na hybridním sekvenování kombinuje dlouhé sekvence s krátkými, aby využily silné stránky obou typů délek sekvencí. Délka dlouhých sekvencí se využívá k identifikaci genomové složitosti a vyřešení repetitivních míst, zatímco krátké sekvence jsou užitečné díky jejich vysoké přesnosti ke zlepšení kvality sekvence a pro charakterizaci lokálních detailů s přesností i jednoho nukleotidu (Y. Wang et al., 2021). Výsledkem je tedy dlouhá, přesná a spojitá sekvence, která má vyřešené repetitivní části.

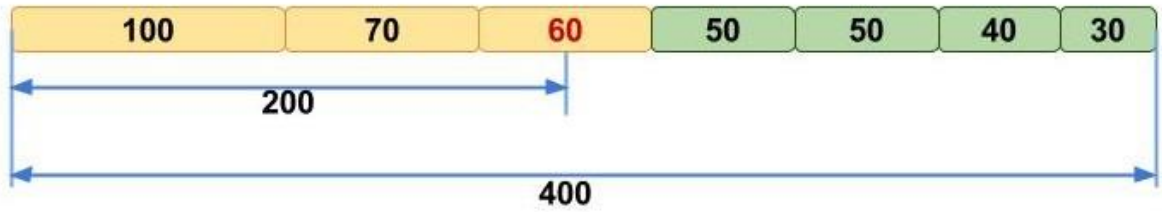
Je vícero možností, jak konstruovat hybridní assembly. Jedna možnost je utvořit lešení (scaffold) z dlouhých sekvencí a přikládat k nim („align“) krátké sekvence (Cao et al., 2017; Madoui et al., 2015), dále je možné vytvořit sestavu z dlouhých sekvencí a výslednou sestavu opravit krátkými sekvencemi (Y. Li et al., 2022) a nakonec se dají zformovat přesné úseky sekvence z krátkých sekvencí zvané „seed-reads“, které se připojí k dlouhým sekvencím (Jansen et al., 2017). Pro každý postup existují různé programy, které lze použít.

Je možné kombinovat například i dlouhé fragmenty z Oxford Nanopore Technologies a krátké 10x Genomics (Y. Li et al., 2022; Ma et al., 2019). Kombinují se dlouhé sekvence z ONT a krátké sekvence s barkódem, který shodně označuje sekvence ze stejného fragmentu. Programy, které lze pro tyto účely využít jsou například kombinace softwarů DBG2OLC s Sparc (Ma et al., 2019) nebo ARKS s LINKS (Coombe et al., 2018; Warren et al., 2015).

### 1.5.4. Vyhodnocení kvality sestav genomu

Výslednou sestavu genomu je potřeba vyhodnotit, aby bylo možné kvalitu jednotlivých sestav mezi sebou porovnat. Pro určení kvality sestavy se standardně používá jednak čistě technické zhodnocení tzv. kontinuity (např. metrika N50, max. a průměrná délka sekvencí, počet sekvencí, počet nepřečtených bází uvnitř sekvencí – tzv. „gapů“, atd.) a jednak biologické zhodnocení a to vyhodnocením kompletnosti genomu na základě porovnání s databází ortologních genů v programu BUSCO (Manchanda et al., 2020).

Metrika N50 vyjadřuje, jak moc je sestava genomu propojená (Bradnam et al., 2013). Podává informace o rozložení délek kontigů. Princip výpočtu N50 (Obr. 6) spočívá v sestupném seřazení všech kontigů podle velikosti, v součtu všech velikostí kontigů a vypočtu 50 % celkové délky. Velikost kontigu, ke kterému sahá 50 % celkové délky je metrika N50 (Videvall, 2017). Číslo N50 je uváděné v jednotkách bp. Čím vyšší je číslo N50, tím vyšší je kontinuita složeného genomu.




Obr.6 : Kontigy seřazené podle délky pro výpočet N50 (Videvall, 2017).

BUSCO (Benchmarking Universal Single-Copy Orthologs; Simão et al., 2015) je program, který odhaduje genovou úplnost sestavy genomu. Je to kvantitativní měřítko, které je založené na evolučně příbuzných genech z různých druhů, ortolozích. Tyto očekávané ortology jsou zaneseny v databázi OrthoDB v9, ze které BUSCO čerpá. BUSCO vyhodnocuje, zda se v určité sestavě ortology vyskytují v jedné kopii, zda jsou duplikované, neúplné anebo jestli úplně chybí. Zjišťuje tak kompletnost zásadních genů jednotlivých sestav genomů.

# 2. Cíle práce

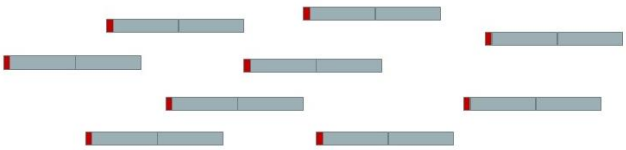
- Sestavit genom cichlidy *Crenicichla semifasciata* různými postupy včetně hybridních
- Porovnat parametry výsledných sestav genomů
- Zhodnotit přínos kombinování sekvenačních technologií na kvalitu výsledku

# 3. Metody

## 3.1. Příprava dat

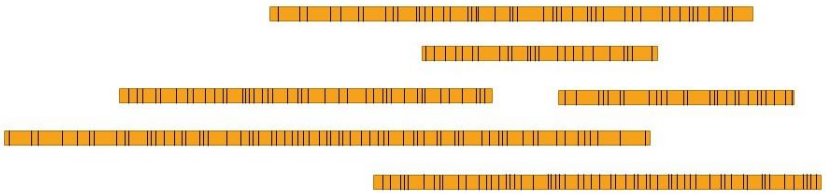
Pro sestavení genomu cichlidy *Crenicichla semifasciata* byl použit jeden exemplář tohoto druhu. Genomová DNA byla extrahovaná z krve odebírané z kaudální cévy s pomocí izolačního kitu Qiagen MagAttract HMW DNA dle standardního protokolu. Data pro zpracování byla získána ze dvou sekvenačních platforem, z technologie 10x Genomics a Oxford Nanopore.

Příprava knihovny 10x Genomics z extrahované DNA a její následná sekvenace na platformě Illumina byla provedena firmou SeqME (Dobříš, ČR). Výsledkem sekvenování byly párové sekvence 2x150 bází s barkódem (Obr. 7), díky kterému je snazší sestavit fragmenty ze stejné molekuly dohromady.



Obr. 7: Data z 10x Genomics. Krátké párové sekvence (šedé) s barkódy (červené).

Pro data z Oxford Nanopore (Obr. 8) se nejdříve musela vyizolovat vysoce kvalitní genomová DNA. Dále se připravila sekvenační knihovna, kde se k DNA připojily adaptéry. Takto připravená DNA se sekvenovala na přístroji MinION. Celý postup je popsán v bakalářské práci Adély Mikešové (Mikešová, 2022). Díky sekvenování na přístroji MinION vznikla primární data (raw data). Celkově byla použita data z 15 sekvenačních běhů. Data, které sekvenovala Mikešová, byla využita pro další analýzy v této práci.



Obr. 8: ONT data. Dlouhé sekvence (oranžové) s chybami (černé).

Pro všechny následující bioinformatické výpočty byly využity zdroje výpočetních center Metacentrum (https://www.metacentrum.cz/cs/). Jedná se o virtuální organizaci, která umožňuje akademickým pracovníkům, zaměstnancům a studentům vědeckovýzkumných institucí v České republice bezplatně využívat výpočetní zdroje, datová uložiště a mnoho dalšího.

## 3.2. Basecalling

Při sekvenování technologií Oxford Nanopore se prostřednictvím ovládacího softwaru MinKNOW zaznamenají průběhy elektrických signálů z detekčního čipu sekvenátoru. Tyto signály reprezentují změny elektrického proudu při průchodu nukleotidu detekčním prostorem nanopóru. Křivka, která znázorňuje změny elektrického proudu se nazývá "squiggle". Primární data ve formátu FAST5, která obsahují záznam elektrických signálů je třeba převést na sekvence bází ve formátu FASTQ. Dekódování křivek na báze se provádí procesem basecalling (Obr. 9).


Obr. 9: Přiřazování bazí ke křivkám squiggle procesem Basecalling. Obrázek vytvořen v BioRender.com

V této práci jsem pro basecalling použila program Guppy (Oxford Nanopore Technologies, 2022a). Kvůli velikosti dat a náročnosti procesu jsem basecalling provedla pro každý sekvenačním běh zvlášť. Použita jsem verzi guppy-6.0.1-gpu s následujícími parametry:

**Pozn**.: Všechny uváděné skripty jsou zobrazeny v rámečcích. Znakem # jsou označeny poznámky.

```
guppy_basecaller -i raw/${run} -s basecalled_reads_sup/${run}\
-c [dna_r9.4.1_450bps_hac.cfg | dna_r10.3_450bps_sup.cfg] --min_qscore 10 \
-x 'auto'  -q 0 --compress_fastq --chunks_per_caller 1000 --gpu_runners_per_device 4 \
--calib_detect --chunk_size 1000 --num_callers 2 --chunks_per_runner 300

#-i,-s = soubory vstupních a výstupních dat včetně cesty
#-c = konfigurační soubor definující typ sekvenační buňky, kit pro přípravu knihovny a model pro basecalling (hac = model High Accuracy, sup = super accuracy)
#--min_qscore 10 = minimální požadované skóre
#většina ostatních parametrů se týká nastavení grafické karty optimalizovaného pro analýzu na výpočetních strojích v Metacentru
```

## 3.3. Úprava a filtrace dat na minimální délku

Po basecallingu jsem sloučila výstupy každého sekvenačního běhu do jednoho souboru a tyto soubory zkomprimovala a přesunula do stejné složky:

```
for i in $(seq -f "%02g" 1 15)          # seq -f "%02g" formátuje číslo i z 4 na 04 atd.
do
  cd dir${i}/pass
  cat *.gz > file${i}
  mv file${i} ../../file${i}.fastq.gz
  cd ../..
done
```

Dále jsem na jednotlivých souborech každého sekvenačního běhu provedla základní statistiky. Pro tyto účely byl použit program seqkit:

```
./seqkit stats -w 0 -G "N" -a *.fastq.gz > all_basecalling_stats.txt


#-w šířka řádku při výstupu formátu FASTA (0 pro žádné zalamování)
#-G definování gapu
#-a všechna statistika včetně kvartilů, délky sekvence, sum_gap, N50
```

Následovalo sloučení všech výsledných souborů do jediného souboru „Adel_all.fastq.gz“. Pro tento soubor jsem opět spočítala statistiky.

Data z platformy Oxford Nanopore obsahují technické sekvence (adaptéry), které bylo potřeba odstranit („trimm“). Adaptéry se nacházejí hlavně na koncích sekvencí, ale mohou se nacházet i uprostřed a indikovat tak artificiálně propojené (chimérické) sekvence. Programy určené na odstranění těchto technických sekvencí je možné přesně nastavit a definovat tak délku a parametry adaptéru. Tyto programy dokáží odstranit adaptéry i uvnitř chimérických sekvencí a rozdělit je do jednotlivých původních sekvencí. Pro odstranění adaptérů v této práci jsem použila program Porechop (https://github.com/rrwick/Porechop) s těmito parametry:

```
Porechop/porechop-runner.py -t 2 -v 1 \
  -i Adel_all.fastq.gz -o Adel_trimmed.fastq.gz --format fastq.gz \
  --check_reads 1000 --end_size 200 \
  --adapter_threshold 90 --middle_threshold 90 --end_threshold 75

#-i,-o = název vstupního a výstupního souboru
#--format = používaný datový formát
#--end_size = délka úseku (v bázích), ve kterém bude na začátku a konci sekvence adaptér hledán
#--check_reads = tento počet readů bude zarovnán se všemi možnými adaptéry, aby bylo možné určit, které sady adaptérů jsou přítomny
#--end_size = počet párů bází na každém konci čtení, ve kterých budou hledány sekvence adaptérů
#--adapter_threshold = adaptéry musí mít alespoň toto procento shody, aby mohla být označena jako přítomná a oříznutá
#--middle_threshold = adaptéry uprostřed readů musí mít alespoň toto procento shody, aby mohly být nalezeny
#--end_threshold = min. percentuální podobnost sekvence, aby byla považována za adaptér odstranění adaptérů uprostřed sekvence a její rozdělení na dvě je zajištěno díky výchozímu nastavení
```

Dalším krokem bylo znáhodnění pořadí sekvencí, jelikož každý sekvenační běh produkoval jinak kvalitní sekvence. Znáhodnění jsem provedla pomocí programu Seqkit a následně soubor zkomprimovala:

```
./seqkit shuffle -j 4 -w 0 Adel_trimmed.fastq.gz > sup10trimshuf.fastq  \
gzip sup10trimshuf.fastq

#-j = počet CPUs
#-w = šířka řádku při výstupu formátu FASTA (0 pro žádné zalamování)
```

Po znáhodnění sekvencí jsem utvořila tři datasety o různých minimálních délkách sekvencí. Minimální délky byly nastaveny na 3000, 5000 a 10000 bazí programem seqkit.

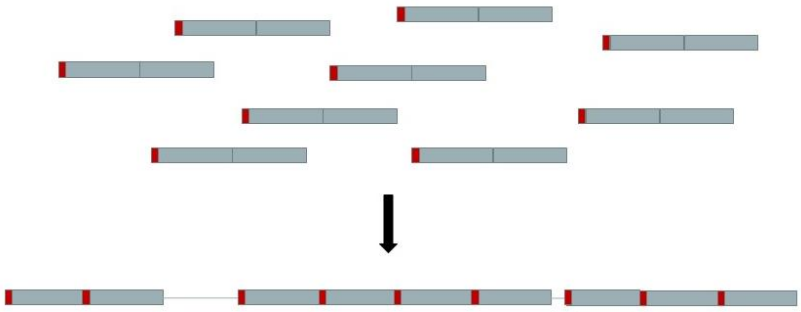
```
./seqkit seq -j 4 -w 0 -m 3000  \
sup10trimshuf.fastq.gz | gzip > sup10trimshuf_len3000.fastq.gz
./seqkit seq -j 4 -w 0 -m 5000  \
sup10trimshuf_len3000.fastq.gz | gzip > sup10trimshuf_len5000.fastq.gz
./seqkit seq -j 4 -w 0 -m 10000  \
sup10trimshuf_len5000.fastq.gz | gzip > sup10trimshuf_len10000.fastq.gz

#-j = počet CPUs
#-w = šířka řádku při výstupu formátu FASTA (0 pro žádné zalamování)
#-m = ponechá pouze sekvence delší než tato minimální délka
```

## 3.4. Sestavení genomu

Genom jsem sestavila s využitím dvou sekvenačních platforem (10x Genomics a ONT) a to jak samostatně, tak i v kombinaci.

### 3.4.1. 10x Genomics

Pro sestavu z krátkých sekvencí (Obr. 10) jsem využila program Supernova (Weisenfeld et al., 2017). Supernova pracuje s párovými sekvencemi.



Obr. 10: Vytvoření sestavy genomu krátkými sekvencemi s poziční informací programem Supernova.

Nejdříve bylo nutné dataset s hloubkou pokrytí 127x odfiltrovat na hloubku pokrytí přibližně 56x. Dále byl spuštěn program Supernova verze supernova-2.1.1, který proběhl ve dvou krocích. Nejprve byla vytvořena de novo sestava celého genomu:

```
supernova run --id semifasciata_01 --fastqs ./raw_zipped \
--maxreads 340000000 --localcores 32 --localmem 256

#--id = unikátní ID běhu
#--fastqs = cesta k vstupním souborům
#--maxreads = cílí přibližně na tento počet readů
#--localcores = limit počtu jader
#--localmem = limit velikosti paměti
```

Druhým krokem bylo generování různých stylů výstupu FASTA pro sestavu genomu.

```
supernova mkoutput --style= pseudohap --asmdir=./semifasciata_01/outs/assembly \
--outprefix=./semifasciata_01/semifasciata_01

#--style = typ využívaného grafu
#--asmdir = cesta výstupních souborů
#--outprefix = označení každého výstupu
```

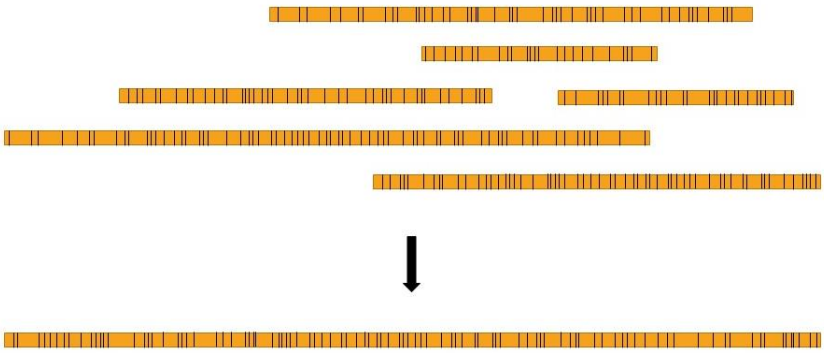
Pro hybridní assembly nebyla tato verze sestavy genomu využívána.

### 3.4.2. Oxford Nanopore

Sestavení genomu z dlouhých sekvencí jsem nejdříve provedla prostřednictvím programu Canu (Koren et al., 2017). Tento postup byl ale pro tuto práci zamítnut kvůli extrémní výpočetní a časové náročnosti (1058 h) a chabým výsledkům.

```
canu -d all_data -p file_all genomeSize=1g \
-nanopore-raw allruns_q7_trim_len10000.fastq.gz -useGrid=false \
-hapMemory=255g -hapThreads=20 -maxMemory=256g \
-maxThreads=20 -stopOnLowCoverage=8

#-d, -p = vstupní soubor a označení assembly
#další parametry se týkají nastavení grafické karty optimalizovaného pro analýzu na výpočetních strojích v Metacentru
```

Pro sestavení genomu z dlouhých sekvencí produkovaných technologií Oxford Nanopore (Obr. 11) jsem tedy použila program Flye (Kolmogorov et al., 2019). Je to de-novo assembler, který pracuje se zobecněnými DBG, opakujícími se grafy („repeat graphs“), jako základní datovou strukturu (core data structure). Grafy jsou sestaveny s použitím přibližných shod sekvencí a mohou tolerovat vyšší úrovně šumu spojeného se čtením sekvenování jedné molekuly a dokážou pracovat i s repeticemi (https://github.com/fenderglass/Flye). Program Flye dokáže zlepšit kvalitu výsledné sestava genomu díky leštění (polishing), které je možné definovat při spouštění programu.



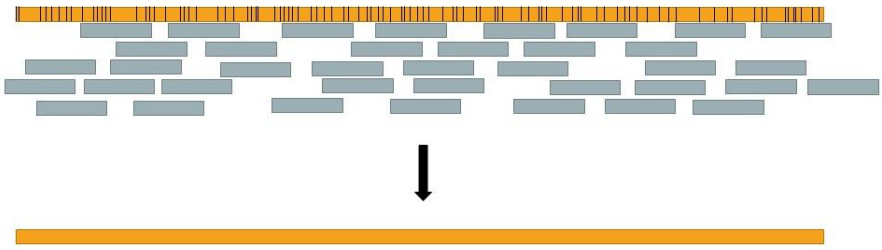
Obr. 11: Sestavení genomu z dlouhých sekvencí technologie Oxford Nanopore programem Flye.

Dlouhé sekvence jsem zpracovala a sestavila od genomové sestavy verzí flye-2.9 s těmito parametry:

```
flye --nano-hq sup10trimshuf_len3000.fastq.gz --out-dir flye29_sup10_polish1 \
-t 16 --iterations 1 –scaffold

#--nano-hq = vysoce kvalitní ONT sekvence: Guppy5+
#--out-dir = Výstupní adresář
#-t = počet paralelních uzlů
#--iterations = počet iterací leštění
#–scaffold = povolit scaffolding pomocí grafu
```

### 3.4.3. Mapování krátkých sekvencí na sestavu Flye

Po sestavení genomu z dat technologie ONT jsem mapovala krátké sekvence z technologie 10x k hotové sestavě (Obr. 12). Použitím programu Flye se sestavil genom, který je nepřesný, jelikož je sestaven jen z dlouhých a nepřesných sekvencí ONT. Mapováním krátkých a přesných sekvencí můžeme přetisknout správnou informaci na sestavu. Díky tomu, že krátké sekvence dosahují vysokou hloubku pokrytí („coverage“ = 127x), je každý nukleotid v sestavě průměrně pokryt 127krát nukleotidem z krátké sekvence. Je tedy možné s velkou jistotou opravit chybně zanesené báze v sestavě a zlepšit tak její kvalitu a kompletnost. Tento proces vytvoření konsenzu jsem prováděla pomocí software BWA (https://bio-bwa.sourceforge.net/bwa.shtml) s využitím algoritmu BWA-MEM. Tento program dokáže pracovat s krátkými sekvencemi o velikosti od 70 bp do 1 Mbp a je rychlý a přesný. Obsahuje funkce, které dokáží zpracovat dlouhé sekvence, a dokonce i neurčitý (rozdělený) alignment a je doporučován pro vysoce kvalitní zpracování sekvencí.



Obr. 12: Krátké sekvence (bez poziční informace) namapované na genomovou sestavu zkonstruovanou v programu Flye opravují chybně osekvenované báze.

Prvním krokem bylo upravení krátkých sekvencí programem LongRanger (Marks et al., 2019). Tento program odstraní barkódy z vlastní sekvence a přemístí je do jejího popisu (hlavičky), zároveň vygeneruje data v tzv. prokládaném („interlaced“ nebo též „interleaved“) formátu (spojí všechny „forward“ a „reverse“ sekvence do jediného souboru, s tím že párové sekvence jsou uvedeny vždy bezprostředně po sobě). Použita jsem verzi longranger-2.2.2:

```
longranger basic --id longranger_basic --fastqs ./raw_data_10x_zipped \
--localmem=64 --localcores=8
```

Krátké sekvence (z technologie 10x Genomics) jsem namapovala na genom sestavený ve Flye prostřednictvím programu bwa v0.7.17. Algoritmus BWA-MEM namapuje maximální počet přesných shod (maximal exact matches, MEMs) a poté je rozšíří pomocí Smith-Watermanovo algoritmu; před vlastním mapováním bylo potřeba pro genom vytvořit pomocné indexové soubory:

```
bwa index assembly_flye_3000.fasta
bwa mem -t 24 -M -T 30 -R "@RG\tID:sem5\tSM:sem5\tPL:ILLUMINA\tLB:sem5" \
-p ../../nanopore/assemblies/${assembly}.fasta \
tenex/longranger_basic/outs/barcoded.fastq.gz > ${assembly}_shortreads.sam

#bwa-mem = Nejnovější typ, obecně doporučovaný pro vysoce kvalitní analýzy, protože je rychlejší a přesnější.
#-t = počet uzlů
#-M = označuje kratší dělené zásahy jako sekundární
#-T = nezobrazovat zarovnání se skóre nižším, než je číslo. Tento parametr ovlivňuje pouze výstup.
#-R = řádek záhlaví dokončené skupiny readů.
#-p = předpokládá, že první vstupní soubor dotazu je prokládaný párovaný FASTA/Q.
```

Z výsledného „alignmentu“ jsem odfiltrovala nekvalitní a nejednoznačně namapované sekvence v programu samtools (Danecek et al., 2021):

```
samtools view -h -q 30 -F 256 -f 2 \
${assembly}_shortreads.sam > ${assembly}_sr_filtered.sam

#-h = ponechává záhlaví ve formátu SAM
#-q = minimální kvalita mapování
#-F = žádné sekundární zarovnání
#-f = pouze správně zarovnané páry
```

Po filtraci jsem ve stejném programu převedla soubory z formátu SAM do binárního formátu BAM:

```
samtools view -S -b ${assembly}_sr_filtered.sam > ${assembly}_sr_filtered.bam

#-S = možnost zkráceného vzorkování
#-b = výstup ve formátu BAM
```

a dále jednotlivé záznamy v souboru BAM setřídila:

```
samtools sort -m 64G -@ 4 -T tmp_ -O bam \
-o ${assembly}_sr_filtered_sorted.bam ${assembly}_sr_filtered.bam

#-m = přibližně maximální požadovaná paměť na vlákno
#-@ = nastavuje počet třídicích a kompresních uzlů
#-T = do této složky se zapisují dočasné soubory
#-O = definovaný typ výstupu
#-o = zapisuje konečný seřazený výstup do souboru, nikoli do standardního výstupu
```

Duplicitní sekvence jsem identifikovala a odstranila s využitím programu Picard v2.8.1. Následně jsem výsledný soubor oindexovala:

```
java -Xms1g -Xmx8g -XX:ParallelGCThreads=5 -jar $PICARD281 MarkDuplicates \
REMOVE_DUPLICATES=true INPUT=${assembly}_sr_filtered_sorted.bam \
OUTPUT=${assembly}_sr_filtered_sorted_dedup.bam\
METRICS_FILE=${assembly}_dedup_metrics.txt

samtools index ${assembly}_sr_filtered_sorted_dedup.bam
```

Genom sestavený ve Flye jsem po indexaci rozdělila kvůli komplexnosti výpočtů do 20 intervalů a následné kroky prováděla paralelně pro každý interval zvlášť:

```
samtools faidx ${assembly}.fa
java -Xmx1g -jar $PICARD281 CreateSequenceDictionary R=${assembly}.fa \
O=${assembly}.dict

mkdir -p intervals_${assembly}
gatk --java-options "-Xmx500m -XX:+UseSerialGC" SplitIntervals \
-R /storage/brno2/home/dada/assemblies/${assembly}.fa \
--scatter-count 20 -O intervals_${assembly}
```

Dále jsem pokračovala s identifikací variant v alignmentu:

```
java -Xmx 8g -Djava.io.tmpdir=tmp-${assembly}-${i}/ -XX:+UseSerialGC \
-jar $GATK/GenomeAnalysisTK.jar -T HaplotypeCaller \
  --output_mode EMIT_VARIANTS_ONLY \
  --variant_index_type LINEAR \
  --variant_index_parameter 128000 \
  --filter_reads_with_N_cigar \
  -L ../../assemblies/intervals_${assembly}/${i}-scattered.interval_list \
  -R /storage/…/assemblies/${assembly}.fa \
  -I ${assembly}_sr_filtered_sorted_dedup.bam \
  -o ${assembly}_sr_called_variants_${i}.vcf.gz

rm -r tmp-${assembly}-${i}
```

Výsledné soubory z jednotlivých genomových intervalů jsem spojila dohromady příkazem gatk:

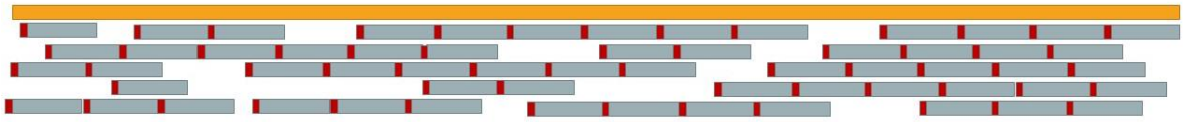
```
gatk --java-options "-Xmx60g -XX:+UseSerialGC" GatherVcfs \
-I ${assembly}_sr_called_variants_0000.vcf.gz \
-I ${assembly}_sr_called_variants_0001.vcf.gz \
-I ${assembly}_sr_called_variants_0002.vcf.gz \
-I ${assembly}_sr_called_variants_0003.vcf.gz \
  ⋮
-I ${assembly}_sr_called_variants_0019.vcf.gz \
-O ${assembly}_sr_called_variants.vcf.gz

gatk --java-options "-Xmx60g -XX:+UseSerialGC" IndexFeatureFile \
-F ${assembly}_sr_called_variants.vcf.gz
```

Posledním krokem bylo vytvoření konsenzu ve formátu FASTA:

```
cat ../../assemblies/${assembly}.fa \
| bcftools consensus ${assembly}_sr_called_variants.vcf.gz > ${assembly}_sr.fas
```

### 3.4.4. Hybridní assembly

Pro hybridní assembly jsem použila postup podle práce Molitor et al., 2021, který pracuje s krátkými sekvencemi z 10x Genomics a jejich barkódy a mapuje je na hotovou sestavu z Flye, která je sestrojena z dlouhých sekvencí (Obr. 13). Dokáže se tak zlepšit kontinuita sestavy díky propojení i dříve nepropojených sekvencí.



Obr. 13: Hybridní assembly kombinuje opravenou sestavu z dlouhých sekvencí z programu Flye s krátkými sekvencemi z metody 10x Genomics obsahující barkód.

Hybridní assembly pracovala s krátkými sekvencemi z technologie 10x Genomics obsahující poziční informaci a se sestavou genomu z Flye opravenou krátkými sekvencemi. Postupovalo se podle práce (Molitor et al., 2021), kde hlavními programy byly Arcs (Coombe et al., 2018) a Links (Warren et al., 2015). Program Arks dokáže zpracovat poziční informaci z barkódu; LINKS je program pro spojování kontigů vytvořených programem Arcs.

Prvním krokem byla technická úprava vstupního souboru (separace barkódu v hlavičce všech sekvencí) v programu bioperl v1.6.1. Dále jsem opět namapovala krátké sekvence na vytvořený genom v programu bwa a záznamy ve výsledném souboru seřadila:

```
gunzip -c  tenex_barcoded.fastq.gz \
| perl -ne 'chomp;$ct++;$ct=1 if($ct>4);if($ct==1){if(/(\@\S+)\sBX\:Z\:(\S{16})/)\
{$flag=1;$head=$1."_".$2;print "$head\n";}  else{$flag=0;}} \
else{print "$_\n" if($flag);}' > ../chromium_barcoded/CHROMIUM_interleaved.fastq


bwa mem -t12  assembly_flye_3000.fasta -p CHROMIUM_interleaved.fastq \
| samtools view -bS - | samtools sort -n - -T sorted > alignment-sorted.bam
```

Poziční informace z barkódů byla zpracována v programu Arcs:

```
arcs -f ../ont_assembly/assembly_flye_3000.fasta -a alignments.fof

mv ../ont_assembly/*main.tsv ./
mv ../ont_assembly/*original.gv ./
mv ../ont_assembly/*dist.gv ./
```

Byl vytvořen pomocný soubor checkpoint.tsv s použitím programu makeTSVfile.py (https://github.com/bcgsc/arcs/blob/binomialx2/Examples/makeTSVfile.py).

```
python2 ../makeTSVfile.py \
$assembly.fas.scaff_arcs_s98_c5_l0_d0_e30000_r0.05_original.gv \
${assembly}_s98_c5_l0_d0_e30000_r0.05.tigpair_checkpoint.tsv \
../../assemblies/$assembly.fas
```

a nakonec vygenerován výsledný genom v programu LINKS v1.8.7:

```
touch empty.fof

LINKS -f ../../assemblies/$assembly.fas -s empty.fof -b \
${assembly}_s98_c5_l0_d0_e30000_r0.05 -a 0.3
```

## 3.5. Zhodnocení kvality

Pro porovnání jednotlivých sestav genomů z hlediska kompletnosti a kontinuity jsem využívala dva programy: Seqkit pro zjištění technických statistik a BUSCO (Manni et al., 2021) pro analýzu kompletnosti genomů na základě porovnání s databází Actinopterygii, která obsahovala 3640 ortologů. Základní statistiky byly vypočítány programem Seqkit na vstupních datech, na sestavách genomu dlouhých i krátkých sekvencí, sestavách Flye s opravou krátkými sekvencemi, a nakonec na hybridní assembly. BUSCO proběhlo na sestavách s následujícími parametry:

```
busco -i ${assembly}.fasta -l actinopterygii_odb10 -o busco_${assembly} -m genome

#-i = vstupní soubor
#-l = dataset rodové linie
#-o = výstupní soubor
#-m = typ (genom, proteiny, transkriptom)
```

Celý proces sestavování genomu od sestavy Flye přes opravu krátkými sekvencemi po hybridní assembly se všemi potřebnými zhodnoceními kvality jsem zopakovala v druhém a třetím opakování. Bylo to z důvodu ovlivnění sestavování genomu náhodnými faktory, kdy algoritmy volí pořadí sekvencí a výsledné sestavy se tak mohou trochu lišit.

# 4. Výsledky

## 4.1. Vstupní data

### 4.1.1. 10x Genomics

DNA knihovna připravená technologií 10x Genomics a sekvenovaná na platformě Illumina poskytla celkem 424,2 mil. párů sekvencí o délce 2x150b s celkovým počtem 117,5 miliard bází Při odhadované délce genomu 921,51 Mb (podle programu Supernova), data dosahují hloubku pokrytí přibližně 127x.

### 4.1.2. Oxford Nanopore

Pro data z ONT bylo dle bakalářské práce Adély Mikešové (Mikešová, 2022) provedeno 10 izolací DNA s průměrnou délkou fragmentů mezi 20 a 40kb, ze kterých bylo následně sestaveno 12 sekvenačních knihoven s použitím dvou typů ligačních kitů: SQK-LSK109 a SQK-LSK110. Dále A. Mikešová provedla celkem 16 sekvenačních běhů na přístroji MinION. Tři experimenty proběhly na sekvenační destičce (flowcell) R10.4.1 a zbytek na R9.4.1. Některé destičky byly znovu použity, jelikož jejich kapacita nebyla po jednom sekvenování spotřebována. Sekvenováním bylo celkově přečteno 42,6 mil. fragmentů DNA, které obsahovaly 102,9 mld. bází (Mikešová, 2022). Dva datasety (Pilot00, Adel15) nebyly do následných analýz zařazeny z důvodu nízké kvality. K těmto datům se přidala data vytvořená firmou SeqMe. Typ destičky následně ovlivnil nastavení parametrů basecallingu.

„Basecalling“ jsem provedla v programu Guppy s využitím algoritmu „super accuracy“ a po sloučení dat vznikl soubor all.fastq.gz s 23,2 mil. sekvencí a počet bazí čítal 79,2 mld (Tab. I). Hloubka pokrytí ONT dat činila přibližně 86x. Po odstranění technických sekvencí (adaptérů) a filtrování dat na minimální délku fragmentů 3000 bp, 5000 bp a 10000 bp jsem spočítala základní statistiky pro každý soubor zvlášť (Tab. II). Filtrováním na minimální délku fragmentů 3000, 5000 a 10000 bp se odfiltrovalo 19 %, 26,6 % a 41,1 % dat a hloubka pokrytí jednotlivých datasetů činila 70x, 63x a 50x (podle pořadí).

Tab. I: Přehled statistik pro jednotlivé sekvenační běhy ONT procesu basecalling jak jednotlivě, tak i sloučené (all).

| Soubor | Počet sekvencí | Počet bází | Min. délka | Průměrná délka | Max délka | N50 |
|---|---|---|---|---|---|---|
| **adel01.fastq.gz** | 423 664 | 3 367 111 114 | 11 | 7 947,6 | 266 525 | 18 922 |
| **adel02.fastq.gz** | 1 111 703 | 8 457 449 224 | 96 | 7 607,7 | 340 429 | 19 725 |
| **adel03.fastq.gz** | 10 056 980 | 15 231 497 285 | 48 | 1 514,5 | 177 245 | 6 982 |
| **adel04.fastq.gz** | 3 530 710 | 5 374 626 398 | 9 | 1 522,3 | 225 170 | 4 167 |
| **adel05.fastq.gz** | 373 785 | 3 006 830 330 | 114 | 8 044,3 | 185 912 | 16 397 |
| **adel06.fastq.gz** | 153 451 | 1 188 400 454 | 106 | 7 744,5 | 264 578 | 15 099 |
| **adel07.fastq.gz** | 1 108 273 | 8 507 815 754 | 39 | 7 676,6 | 231 861 | 17 568 |
| **adel08.fastq.gz** | 359 915 | 2 469 735 385 | 70 | 6 862 | 189 078 | 15 015 |
| **adel09.fastq.gz** | 3 321 058 | 8 414 361 044 | 28 | 2 533,6 | 200 505 | 4 509 |
| **adel10.fastq.gz** | 372 116 | 2 699 640 213 | 88 | 7 254,8 | 388 362 | 28 504 |
| **adel11.fastq.gz** | 723 006 | 5 160 532 792 | 110 | 7 137,6 | 180 041 | 17 401 |
| **adel12.fastq.gz** | 558 248 | 4 826 437 089 | 29 | 8 645,7 | 170 958 | 19 295 |
| **adel13.fastq.gz** | 17 618 | 181 585 241 | 167 | 10 306,8 | 115 449 | 20 207 |
| **adel14.fastq.gz** | 40 094 | 511 189 477 | 85 | 12 749,8 | 152 925 | 22 331 |
| **SeqMe.fastq.gz** | 1 042 510 | 9 827 788 260 | 79 | 9 427 | 167 469 | 20 594 |
| **all.fastq.gz** | 23 193 131 | 79 225 000 060 | 9 | 3 415,9 | 388 362 | 13 830 |

Tab. II: Přehled statistik pro data po odstranění adaptérů a filtrování na určitou minimální délku fragmentů.

| Soubor | Počet sekvencí | Počet bází | Min. délka | Průměrná délka | Max délka | N50 | Q20(%) | Q30(%) |
|---|---|---|---|---|---|---|---|---|
| **3000.fas** | 5 261 867 | 64 154 599 644 | 3 000 | 12 192,4 | 388 330 | 18 387 | 65,28 | 20,16 |
| **5000.fas** | 3 706 272 | 58 115 307 713 | 5 000 | 15 680,3 | 388 330 | 20 472 | 65,32 | 20,19 |
| **10000.fas** | 2 086 028 | 46 624 529 107 | 10 000 | 22 350,9 | 388 330 | 24 946 | 65,47 | 20,31 |

## 4.2. Sestavení genomu

### 4.2.1. 10x Genomics

Sestava genomu vytvořená v programu Supernova složená z krátkých sekvencí z metody 10x Genomics dosahovala vysoké hodnoty kompletnosti genomu BUSCO = 97 % (Tab. IV), jelikož původní sekvence jsou velice přesné. Na druhou stranu obsahuje velké množství mezer (skoro 62,5 mil.; Tab. III) a stává se tak málo kontinuální. Z důvodu malé kvality nebyla tato sestava genomu použita v dalším zpracování.

### 4.2.2. Oxford Nanopore

Nejprve byla vytvořena sestava genomu programem Canu. Samotný proces skládání sestavy trval 1058 h (cca 1,5 měsíce). Výsledná sestava byla velmi nekvalitní, měla velmi nízkou hodnotu N50: 2,6 Mb (Tab. III) a BUSCO dosahovalo pouze 86 % (Tab. IV). Z těchto důvodů byl tento postup z této práce vyřazen.

Další sestava genomu byla vytvořena programem Flye s jedním přečištěním (polishing). Tento proces skládání genomu trvalo nejdéle 33 h. Byly vytvořené sestavy, které dosahovaly N50 od 5,17 do 6,51 Mb. Sestavy genomu vznikly z 964 až 1 220 scaffoldů s maximální délkou 21,8 až 26,9 Mb (Tab. III). BUSCO dosahovalo hodnot 98 % – 98,2 % (Tab. IV).

### 4.2.3. Oprava krátkými sekvencemi

Opravením sestavy genomu z Flye krátkými sekvencemi vznikla sestava, která měla o trochu nižší hodnotu N50 než sestava z Flye, ale hodnotu Busco měla lepší o 0,4 – 0,6 % (Tab. IV).

### 4.2.4. Hybridní assembly

Díky hybridní assembly se spojily scaffoldy, které předtím být spojené nemohly. Zvětšila se tak hodnota N50 více než dvojnásobně (Tab. III). U minimální délky vstupních sekvencí 3000 bp to bylo až o 188 %, u min. délky 5000 bp až o 177,6 % a u min. délky 10000 bp až o 127 %. Vzniklé sestavy genomu obsahovaly 754–928 scaffoldů s maximální délkou 31,1 – 37,9 Mb. BUSCO se od verze sestavy opravené krátkými sekvencemi nelišilo, až na min délku 3000, kde BUSCO vzrostlo o 0,1 % (Tab. IV).

## 4.3. Zhodnocení kvality

### 4.3.1. Seqkit

Pro každou sestavu genomu jsem spočítala základní technické statistiky vyjadřující kontinuitu. Vznikla data (Tab. III), kde nejdůležitější parametr je číslo N50. Výsledky prvního, druhého i třetího opakování korelují. Velikost sestav ONT dosahovala průměrně 824,5 Mb, což je přibližně 90 % odhadované délky genomu.

Tab. III: Výsledky statistik všech sestav genomu. Nejlepší (zeleně) a nejhorší (červeně) výsledky jsou vyznačeny barvami.

| Assembly | Počet sekvencí | Počet bází | Min. délka | Průměrná délka | Max délka | Počet gapů | N50 |
|---|---|---|---|---|---|---|---|
| **Supernova** | 13 847 | 842 068 128 | 1 000 | 60 812,3 | 31 290 156 | 62 478 660 | 10 701 586 |
| **Canu_10000** | 2 954 | 1 023 534 364 | 3 555 | 346 491 | 20 771 730 | 0 | 2 704 201 |
| **Flye_3000_1** | 1 193 | 820 382 117 | 513 | 687 663,1 | 21 772 651 | 6 100 | 5 538 283 |
| **Flye_sr_3000_1** | 1 193 | 820 011 260 | 513 | 687 352,3 | 21 766 931 | 6 100 | 5 534 396 |
| **Hybridní_3000_1** | 754 | 830 260 333 | 526 | 1 101 141 | 33 554 068 | 27 300 | 14 557 639 |
| **Flye_3000_2** | 1 220 | 819 665 230 | 514 | 671 856,7 | 21 762 711 | 5 700 | 5 561 506 |
| **Flye_sr_3000_2** | 1 220 | 819 275 751 | 514 | 671 537,5 | 21 756 382 | 5 700 | 5 557 654 |
| **Hybridní_3000_2** | 928 | 819 304 951 | 514 | 882 871,7 | 37 863 710 | 34 900 | 14 698 581 |
| **Flye_3000_3** | 1 214 | 818 418 347 | 517 | 674 150,2 | 21 763 829 | 5 900 | 5 530 780 |
| **Flye_sr_3000_3** | 1 214 | 818 030 235 | 517 | 673 830,5 | 21 757 291 | 5 900 | 5 527 093 |
| **Hybridní_3000_3** | 902 | 818 061 435 | 517 | 906 941,7 | 33 856 498 | 37 100 | 15 918 398 |
| **Flye_5000_1** | 1 158 | 825 111 426 | 526 | 712 531,5 | 26 921 785 | 6 800 | 5 175 022 |
| **Flye_sr_5000_1** | 1 158 | 824 638 346 | 526 | 712 122,9 | 26 912 065 | 6 800 | 5 171 584 |
| **Hybridní_5000_1** | 889 | 820 041 660 | 513 | 922 431,6 | 33 885 359 | 36 500 | 14 355 502 |
| **Flye_5000_2** | 1 162 | 825 056 031 | 511 | 710 031 | 26 921 446 | 6 500 | 5 175 029 |
| **Flye_sr_5000_2** | 1 162 | 824 584 643 | 512 | 709 625,3 | 26 911 417 | 6 500 | 5 171 574 |
| **Hybridní_5000_2** | 847 | 824 616 143 | 512 | 973 572,8 | 26 911 417 | 38 000 | 12 681 628 |
| **Flye_5000_3** | 1 187 | 824 514 050 | 518 | 694 620,1 | 22 124 116 | 6 500 | 5 246 180 |
| **Flye_sr_5000_3** | 1 187 | 824 042 662 | 518 | 694 223 | 22 110 799 | 6 500 | 5 239 669 |
| **Hybridní_5000_3** | 860 | 824 075 362 | 518 | 958 227,2 | 31 989 029 | 39 200 | 13 542 798 |
| **Flye_10000_1** | 964 | 830 894 482 | 526 | 861 923,7 | 26 782 077 | 6 300 | 6 189 181 |
| **Flye_sr_10000_1** | 964 | 830 239 333 | 526 | 861 244 | 26 765 326 | 6 300 | 6 182 375 |
| **Hybridní_10000_1** | 850 | 824 669 146 | 526 | 970 199 | 31 131 802 | 37 600 | 12 549 625 |
| **Flye_10000_2** | 1 000 | 829 758 647 | 513 | 829 758,6 | 26 761 853 | 6 000 | 6 189 403 |
| **Flye_sr_10000_2** | 1 000 | 829 103 116 | 513 | 829 103,1 | 26 745 435 | 6 000 | 6 182 918 |
| **Hybridní_10000_2** | 792 | 829 123 916 | 513 | 1 046 873,6 | 36 606 483 | 26 800 | 13 105 769 |
| **Flye_10000_3** | 976 | 829 789 458 | 519 | 850 194,1 | 26 762 485 | 6 000 | 6 516 845 |
| **Flye_sr_10000_3** | 976 | 829 131 754 | 519 | 849 520,2 | 26 746 097 | 6 000 | 6 511 801 |
| **Hybridní_10000_3** | 770 | 829 152 354 | 519 | 1 076 821,2 | 33 553 078 | 26 600 | 14 815 052 |

### 4.3.2. BUSCO

Pro každou výslednou sestavu genomu jsem vypočítala kompletnost pomocí porovnání s databází ortologních genů v programu BUSCO. Hodnota BUSCO C obsahuje sekvence, které splňují omezení délky a jsou tedy „kompletní“. Pokud jsou příliš krátké, jsou zařazeny k hodnotě F jako „fragmentované“. Hodnota M („missing“) zobrazuje sekvence, které v sestavě chybí. Z výsledků v tabulce (Tab. IV) jde vidět, že díky opravě sestavy z programu Flye krátkými sekvencemi se už tak výborná kvalita sestavy genomu 98 % zvětšila až na 98,7 %.

Tab. IV: Výsledky BUSCO všech sestav genomu.

| **Assembly** | **Busco C** | **Busco F** | **Busco M** |
|---|---|---|---|
| **Supernova** | 97,00 % | 1,20 % | 1,80 % |
| **Canu_10000** | 89,80 % | 3,70 % | 6,50 % |
| **Flye_3000_1** | 98,00 % | 0,70 % | 1,30 % |
| **Flye_sr_3000_1** | 98,60 % | 0,40 % | 1,00 % |
| **Hybridní_3000_1** | 98,70 % | 0,40 % | 0,90 % |
| **Flye_3000_2** | 98,10 % | 0,70 % | 1,20 % |
| **Flye_sr_3000_2** | 98,60 % | 0,40 % | 1,00 % |
| **Hybridní_3000_2** | 98,70 % | 0,40 % | 0,90 % |
| **Flye_3000_3** | 98,10 % | 0,70 % | 1,20 % |
| **Flye_sr_3000_3** | 98,60 % | 0,40 % | 1,00 % |
| **Hybridní_3000_3** | 98,70 % | 0,40 % | 0,90 % |
| **Flye_5000_1** | 98,00 % | 0,70 % | 1,30 % |
| **Flye_sr_5000_1** | 98,60 % | 0,40 % | 1,00 % |
| **Hybridní_5000_1** | 98,60 % | 0,40 % | 1,00 % |
| **Flye_5000_2** | 98,10 % | 0,70 % | 1,20 % |
| **Flye_sr_5000_2** | 98,60 % | 0,40 % | 1,00 % |
| **Hybridní_5000_2** | 98,70 % | 0,40 % | 0,90 % |
| **Flye_5000_3** | 98,10 % | 0,70 % | 1,20 % |
| **Flye_sr_5000_3** | 98,70 % | 0,40 % | 0,90 % |
| **Hybridní_5000_3** | 98,70 % | 0,40 % | 0,90 % |
| **Flye_10000_1** | 98,10 % | 0,70 % | 1,20 % |
| **Flye_sr_10000_1** | 98,60 % | 0,40 % | 1,00 % |
| **Hybridní_10000_1** | 98,60 % | 0,40 % | 1,00 % |
| **Flye_10000_2** | 98,20 % | 0,70 % | 1,10 % |
| **Flye_sr_10000_2** | 98,60 % | 0,40 % | 1,00 % |
| **Hybridní_10000_2** | 98,60 % | 0,40 % | 1,00 % |
| **Flye_10000_3** | 98,20 % | 0,70 % | 1,10 % |
| **Flye_sr_10000_3** | 98,60 % | 0,40 % | 1,00 % |
| **Hybridní_10000_3** | 98,60 % | 0,40 % | 1,00 % |

# 5. Diskuse

Sekvenační technologie dokáží vyprodukovat obrovské množství dat, které je třeba bioinformaticky zpracovat. Je mnoho postupů a druhů analýz, které je možné provést. V této práci bylo tématem hybridní „assembly“, tedy kombinování vstupních dat z různých sekvenačních platforem při sestavování genomu. Hlavní otázkou bylo, zda tento způsob složení genomu přináší lepší výsledky a zda se vyplatí do toho investovat.

Výsledkem assembleru Supernova, který pracoval s krátkými sekvencemi s poziční informací z 10x Genomics je sestava, která se může na první pohled zdát dostačující. Hodnota N50 dosahuje velkých hodnot: 10,7 Mb (Tab. III) a kompletnost genů podle BUSCO je také vysoká: 97 % (Tab. IV). Vysoká hodnota N50 značí, že sestava genomu je složená z dlouhých scaffoldů. V tomto případě je scaffold kromě kontinuálních kontigů vyplněn velkým množstvím mezer. Sekvence jsou spojeny do kontigů podle poziční informace, která určuje, které sekvence patří k sobě. Bohužel se už neví, v jakém pořadí se tyto sekvence vyskytují. Dále je scaffold vyplněn nespecifickými bázemi, které se značí písmenem „N“. Nespecifické báze N jsou definovány jako gapy. V této stavě z assembleru Supernova se nachází 62,5 milionů nespecifických bází (gapů), což je přibližně 7,4 % celkové délky sestavy. To značí, že sestava je složená nejen ze sekvencí o nejasném pořadí, ale také z mnoha nespecifických sekvencí. Takto vypadající sestava genomu je typická pro sestavy z krátkých sekvencí (Peona et al., 2021). Z těchto poznatků nelze určit tuto sestavu z krátkých sekvencí jako vyhovující pro referenční genom.

V případě sestav genomů z dlouhých sekvencí byl program Canu nejpoužívanější assembler z důvodu nejlepších výsledků (Lu et al., 2016). V této práci, kdy se pracovalo s obrovským množstvím dat, však vyšla sestava s nejhoršími výsledky. Canu vytvořilo sestavu s nejmenší hodnotou N50 (2,7 Mb) stejně tak s nejnižší hodnotou BUSCO (89,8 %). Program Canu se stal pro tuto práci nepoužitelný nejen kvůli špatným výsledkům, ale také kvůli obrovskému množství použitých zdrojů a času procesu. Z těchto důvodů byl vybrán assembler Flye, který využíval mnohonásobně menší výpočetní zdroje a přinesl mnohem lepší výsledky.

Pro sestavení genomu programem Flye jsem nejdříve primární data z Oxford Nanopore upravila a provedla basecalling s nejnovějším verzí Guppy 6.0.1. Nový typ algoritmu pro překlad elektrických signálů do sekvencí DNA „super-accuracy“ poskytuje kvalitnější výsledky než jeho starší verze. Mikešová používala verzi programu pro basecalling 4.4.1., filtrovala data na minimální délku sekvencí 1000 bazí a k sestavení genomu používala stejnou verzi Flye (Mikešová, 2022). Vyšly jí sestavy genomu, které u hloubky pokrytí 70x dosahovaly maximální hodnoty N50 12,3 Mb, a kompletnost dosahovala 96,6 %. V této práci jsem používala verzi programu Guppy 6.0.1. a data jsem filtrovala na minimální délku sekvencí 3000, 5000 a 10000 bazí, kdy dataset s minimální délkou 3000 dosahoval hloubky pokrytí přibližně 70x. Z tohoto datasetu vznikly sestavy genomu s maximální N50 5,6 Mb a kompletností 98,1 %. Oproti sestavě Mikešové se N50 snížila a hodnota BUSCO se zvýšila. Lze tedy usuzovat, že s přesnější verzí programu pro basecalling se zkvalitnily výsledné sestavy. Zastaralejší verze „high-accuracy“ mohla vyprodukovat méně přesná data a ta mohla způsobit větší variabilitu při spojování sekvencí v procesu sestavování genomu. Mohlo tak vzniknout nesprávné spojení sekvencí, čímž se zvýšila hodnota N50. Kvůli špatnému spojení sekvencí mohly vzniknout artefakty a jednotlivé geny nemusely být spojeny správně. Z tohoto důvodu má sestava Mikešové nižší hodnotu BUSCO. Pro utvoření věrohodnější sestavení genomu je tedy lepší využít pro basecalling novější verzi algoritmu.

Díky programu Flye a nejnovější verzi programu pro basecalling vznikla velice kvalitní sestava genomu. Tato sestava dosahovala vysoké hodnoty BUSCO (98 - 98,2%), ale hodnota N50 vyšla menší (maximálně 6,18 Mb). Pro zkvalitnění výsledků jsem tedy nejdříve sestavu opravila krátkými sekvencemi a pro zlepšení kontinuity jsem použita hybridní assembly. Díky opravě sestavy krátkými sekvencemi ze sekvenační platformy Illumina se zvýšila hodnota BUSCO přibližně o 0,5 % až na hodnotu 98,6 % (Tab. IV). V takto vysokých hodnotách je významné každé zlepšení. V porovnání se zatím nejkvalitnějším genomem cichlid, tilapie nilské (*Oreochromis niloticus*; https://www.ncbi.nlm.nih.gov/genbank/; txid8128 [Organism:exp]), je genová kompletnost mých genomů nižší jen o jedno procento. Genom *O. niloticus* dosahuje kompletnosti 99,7 %. Dá se tedy říci, že z hlediska kompletnosti jsou mé genomy velice kvalitní. Oproti hodnotě BUSCO se hodnota N50 mírně snížila ze stejných důvodů, jako v případě použití lepší verze Guppy. Sestava v základním stavu mohla být složená z méně přesných sekvencí. Opravením chybných bází tyto nepřesnosti vymizely a mohlo se tak změnit propojení sekvencí. Tímto krokem bylo dokázáno, že oprava krátkými sekvencemi zvyšuje kompletnost genů v sestavě genomu.

Jak už bylo zmíněno, základní sestava z assembleru Flye dosahovala nízkých hodnot N50 a bylo nutné zvýšit kontinuitu sestavy. Po sestavení hybridní assembly, která kombinovala krátké sekvence s barkódy a opravenou sestavu genomu z Flye vznikl genom, který dosahoval hodnoty N50 až 15,9 Mb (Tab. III). Díky poziční informaci z krátkých sekvencí se dokázaly propojit i dříve nepropojené kontigy a vytvořily tak mnohem delší scaffoldy. Počet scaffoldů se tak snížil, zvýšila se průměrná i maximální délka scaffoldů a hodnota N50 stoupla až 2,9x. V některých případech se také zlepšila kompletnost o 0,1 % až na hodnotu 98,7 % (Tab. IV). Je možné, že se spojily úseky s geny, které nebyly dříve propojené. Z těchto údajů je vidět, že kombinací krátkých sekvencí s poziční informací s opravenou sestavou genomu krátkými sekvencemi se zvyšuje kontinuita výsledného sestavení genomu.

# 6. Závěr

Pro skládání genomu daty z technologie Oxford Nanopore se vyplatí data zpracovat nejnovějším programem pro basecalling s nejnovější variantou „super accuracy“. Hybridní assembly přineslo zlepšení jak v oblasti kontinuity (N50), tak v oblasti kompletnosti (BUSCO). Má tedy velký smysl investovat do sekvenování na obou platformách finance a do zpracovávání výsledků drahocenný čas.

# 7. Reference


Atkinson, M. R., Deutscher, M. P., Kornberg, A., Russell, A. F., & Moffatt, J. G. (1969). Enzymatic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2',3'-dideoxyribonucleotide. *Biochemistry*, *8*(12), 4897–4904. https://doi.org/10.1021/BI00840A037

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W. C., Corbeil, J., Fabbro, C. Del, Docking, R. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., … Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1). https://doi.org/10.1186/2047-217X-2-10

Burress, E. D., Piálek, L., Casciotta, J., Almirón, A., & Říčan, O. (2022). Rapid Parallel Morphological and Mechanical Diversification of South American Pike Cichlids ( Crenicichla ) . *Systematic Biology*. https://doi.org/10.1093/sysbio/syac018

Burress, E. D., Piálek, L., Casciotta, J. R., Almirón, A., Tan, M., Armbruster, J. W., & Říčan, O. (2018). Island- and lake-like parallel adaptive radiations replicated in rivers. *Proceedings. Biological sciences*, *285*(1870). https://doi.org/10.1098/RSPB.2017.1762

Cao, M. D., Nguyen, S. H., Ganesamoorthy, D., Elliott, A. G., Cooper, M. A., & Coin, L. J. M. (2017). Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nature communications*, *8*. https://doi.org/10.1038/NCOMMS14515

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., & Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods*, *13*(12), 1050–1054. https://doi.org/10.1038/NMETH.4035

Coombe, L., Zhang, J., Vandervalk, B. P., Chu, J., Jackman, S. D., Birol, I., & Warren, R. L.

(2018). ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC bioinformatics*, *19*(1). https://doi.org/10.1186/S12859-018-2243-X

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/GIGASCIENCE/GIAB008

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. *Nature Biotechnology*, *34*(5), 518–524. https://doi.org/10.1038/nbt.3423

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, *11*(5), 759–769. https://doi.org/10.1111/J.1755-0998.2011.03024.X

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., & Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(4), 1513–1518. https://doi.org/10.1073/PNAS.1017351108/SUPPL_FILE/PNAS.201017351SI.PDF

Green, E. (2022, říjen 28). *Contig*. National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Contig

Heckel, J. J. (1840). Johann Natterer's neue Flussfische Brasilien's: nach den Beobachtungen und Mittheilungen des Entdeckers. In *Johann Natterer's neue Flussfische Brasilien's: nach den beobachtungen und mittheilungen des entdeckers*. Annalen des Wiener Museums der Naturgeschichte . https://doi.org/10.5962/bhl.title.101457

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquiee, M., Mmerrill, S. H., Penswick, J. R., & Zamir, A. (1965). Structure of a Ribonucleic Acid. *Science*, *147*(3664), 1462–1465. https://doi.org/10.1201/9781420012026-6

Idury, R. M., & Waterman, M. S. (1995). A new algorithm for DNA sequence assembly. *Journal of computational biology : a journal of computational molecular cell biology*, *2*(2), 291–306. https://doi.org/10.1089/CMB.1995.2.291

Illumina. (2010). Illumina Sequencing Technology. *Technology Spotlight: Illumina ® Sequencing*. https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Illumina. (2022, červen 20). *Maximum read length for Illumina sequencing platforms*. https://support.illumina.com/bulletins/2020/04/maximum-read-length-for-illumina-sequencing-platforms.html

Istace, B., Friedrich, A., d'Agata, L., Faye, S., Payen, E., Beluche, O., Caradec, C., Davidas, S., Cruaud, C., Liti, G., Lemainque, A., Engelen, S., Wincker, P., Schacherer, J., & Aury, J. M. (2017). de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *GigaScience*, *6*(2). https://doi.org/10.1093/GIGASCIENCE/GIW018

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology 2016* , *17*(1), 1–11. https://doi.org/10.1186/S13059-016-1103-0

Jansen, H. J., Liem, M., Jong-Raadsen, S. A., Dufour, S., Weltzien, F. A., Swinkels, W., Koelewijn, A., Palstra, A. P., Pelster, B., Spaink, H. P., Thillart, G. E. V. Den, Dirks, R. P., & Henkel, C. V. (2017). Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *Scientific Reports 2017 7:1*, *7*(1), 1–13. https://doi.org/10.1038/s41598-017-07650-6

Ju, J., Kim, D. H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M. S., Shi, S., Wu, J., Edwards, J. R., Romu, A., & Turro, N. J. (2006). Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(52), 19635–19640. https://doi.org/10.1073/PNAS.0609513103/SUPPL_FILE/09513FIG8.JPG

Kasianowicz, J. J., Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(24), 13770–13773. https://doi.org/10.1073/PNAS.93.24.13770

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology 2019 37:5*, *37*(5), 540–546.

https://doi.org/10.1038/s41587-019-0072-8

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/GR.215087.116

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

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics (Oxford, England)*, *32*(14), 2103–2110. https://doi.org/10.1093/BIOINFORMATICS/BTW152

Li, Y., Tan, C., Li, Z., Guo, J., Li, S., Chen, X., Wang, C., Dai, X., Yang, H., Song, W., Hou, L., Xu, J., Tong, Z., Xu, A., Yuan, X., Wang, W., Yang, Q., Chen, L., Sun, Z., … Li, J. (2022). The genome of Dioscorea zingiberensis sheds light on the biosynthesis, origin and evolution of the medicinally important diosgenin saponins. *Horticulture Research*, *9*. https://doi.org/10.1093/HR/UHAC165

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

Lin, Y., Yuan, J., Kolmogorov, M., Shen, M. W., Chaisson, M., & Pevzner, P. A. (2016). Assembly of long error-prone reads using de Bruijn graphs. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(52), E8396–E8405. https://doi.org/10.1073/PNAS.1604560113

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, *14*(5), 265–279. https://doi.org/10.1016/J.GPB.2016.05.004

Ma, Z. (Sam), Li, L., Ye, C., Peng, M., & Zhang, Y. P. (2019). Hybrid assembly of ultra-long Nanopore reads augmented with 10x-Genomics contigs: Demonstrated with a human genome. *Genomics*, *111*(6), 1896–1901.

https://doi.org/10.1016/J.YGENO.2018.12.013

Madoui, M. A., Engelen, S., Cruaud, C., Belser, C., Bertrand, L., Alberti, A., Lemainque, A., Wincker, P., & Aury, J. M. (2015). Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC genomics*, *16*(1). https://doi.org/10.1186/S12864-015-1519-Z

Manchanda, N., Portwood, J. L., Woodhouse, M. R., Seetharam, A. S., Lawrence-Dill, C. J., Andorf, C. M., & Hufford, M. B. (2020). GenomeQC: A quality assessment tool for genome assemblies and gene structure annotations. *BMC Genomics*, *21*(1), 1–9. https://doi.org/10.1186/S12864-020-6568-2/FIGURES/3

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/MOLBEV/MSAB199

Marks, P., Garcia, S., Barrio, A. M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., Fiddes, I. T., Galvin, B., Heaton, H., Herschleb, J., Hindson, C., Holt, E., Jabara, C. B., Jett, S., Keivanfar, N., … Church, D. M. (2019). Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Research*, *29*(4), 635–645. https://doi.org/10.1101/GR.234443.118

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, *74*(2), 560–564. https://doi.org/10.1073/PNAS.74.2.560

Midha, M. K., Wu, M., & Chiu, K. P. (2019). Long-read sequencing in deciphering human genetics to a greater depth. *Human genetics*, *138*(11–12), 1201–1215. https://doi.org/10.1007/S00439-019-02064-Y

Mikešová, A. (2022). De novo sekvenování genomu Crenicichla semifasciata technologií Oxford Nanopore. *Bakalářská práce*.

Min-Jou, W., Haegeman, G., Ysebaert, M., & Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, *237*, 82–88. https://doi.org/10.1038/239137a0

Molitor, C., Kurowski, T. J., De Almeida, P. M. F., Eerolla, P., Spindlow, D. J., Kashyap, S. P., Singh, B., Prasanna, H. C., Thompson, A. J., & Mohareb, F. R. (2021). De novo genome assembly of Solanum sitiens reveals structural variation associated with drought and salinity tolerance. *Bioinformatics*, *37*(14), 1941–1945. https://doi.org/10.1093/BIOINFORMATICS/BTAB048

Morozova, O., & Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, *92*(5), 255–264. https://doi.org/10.1016/J.YGENO.2008.07.001

Oxford Nanopore technologies. (2022). *Product specifications*. https://nanoporetech.com/products/specifications#specs[tab]=versatile

Oxford Nanopore Technologies. (2022a). *Guppy Basecalling Software verze guppy-6.0.1-gpu*.

Oxford Nanopore Technologies. (2022b). *MinION*. https://nanoporetech.com/products/minion

Pearman, W. S., Freed, N. E., & Silander, O. K. (2020). Testing the advantages and disadvantages of short- And long- read eukaryotic metagenomics using simulated reads. *BMC Bioinformatics*, *21*(1), 1–15. https://doi.org/10.1186/S12859-020-3528-4/TABLES/1

Peona, V., Blom, M. P. K., Xu, L., Burri, R., Sullivan, S., Bunikis, I., Liachko, I., Haryoko, T., Jønsson, K. A., Zhou, Q., Irestedt, M., & Suh, A. (2021). Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Molecular Ecology Resources*, *21*(1), 263–286. https://doi.org/10.1111/1755-0998.13252

Pereira, D. M., Fernandes, J. C., Valentêo, P., & Andrade, P. B. (2015). Target Identification and Validation: „Omics" Technologies: Promises and Benefits for Molecular Medicine. *Principles of Translational Science in Medicine: From Bench to Bedside: Second Edition*, 25–39. https://doi.org/10.1016/B978-0-12-800687-0.00003-7

Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, *98*(17), 9748–9753. https://doi.org/10.1073/PNAS.171285098

Piálek, L., Burress, E., Dragová, K., Almirón, A., Casciotta, J., & Říčan, O. (2019). Phylogenomics of pike cichlids (Cichlidae: Crenicichla) of the C. mandelburgeri species complex: rapid ecological speciation in the Iguazú River and high endemism in the Middle Paraná basin. *Hydrobiologia*, *832*(1), 355–375. https://doi.org/10.1007/s10750-018-3733-6

Piálek, L., Říčan, O., Casciotta, J., Almirón, A., & Zrzavý, J. (2012). Multilocus phylogeny of Crenicichla (Teleostei: Cichlidae), with biogeography of the C. lacustris group: Species flocks as a model for sympatric speciation in rivers. *Molecular Phylogenetics and Evolution*, *62*(1), 46–61. https://doi.org/10.1016/j.ympev.2011.09.006

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Schreiber, J., Wescoe, Z. L., Abu-Shumays, R., Vivian, J. T., Baatar, B., Karplus, K., & Akeson, M. (2013). Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(47), 18910–18915. https://doi.org/10.1073/PNAS.1310615110/SUPPL_FILE/SM01.WMV

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*, *31*(19), 3210–3212. https://doi.org/10.1093/BIOINFORMATICS/BTV351

Staden, R. (1980). A mew computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Research*, *8*(16), 3673–3694. https://doi.org/10.1093/NAR/8.16.3673

UC Davis Bioinformatics. (2018). *10X genomics data and assembly with Supernova* |. https://ucdavis-bioinformatics-training.github.io/2018-Dec-Genome-Assembly/10x-supernova/10x-supernova.html

Videvall, E. (2017, březen 29). *What's N50?* . The Molecular Ecologist. https://www.molecularecologist.com/2017/03/29/whats-n50/

Wang, H., Dunning, J. E., Huang, A. P. H., Nyamwanda, J. A., & Branton, D. (2004). DNA

heterogeneity and phosphorylation unveiled by single-molecule electrophoresis. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(37), 13472–13477. https://doi.org/10.1073/PNAS.0405568101

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology 2021 39:11*, *39*(11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x

Warren, R. L., Yang, C., Vandervalk, B. P., Behsaz, B., Lagman, A., Jones, S. J. M., & Birol, I. (2015). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*, *4*(1). https://doi.org/10.1186/S13742-015-0076-3

Watson, J. D., & Crick, F. H. C. . (1953). A structure for deoxyribose nucleic acid. *Nature*, *171*, 737–738. https://www.nature.com/scitable/content/Molecular-Structure-of-Nucleic-Acids-16331/

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, *27*(5), 757–767. https://doi.org/10.1101/GR.214874.116

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*, *20*(129), 1–10. https://doi.org/10.1186/s13059-019-1727-y

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/GR.074492.107

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications 2017 8:1*, *8*(1), 1–12. https://doi.org/10.1038/ncomms14049