Jihočeská univerzita v Českých Budějovicích
Přírodovědecká fakulta

De novo sekvenování genomu *Crenicichla semifasciata* technologií Oxford Nanopore

Bakalářská práce

Adéla Mikešová

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

České Budějovice 2022

Mikešová, A., 2022: *De novo* sekvenování genomu *Crenicichla semifasciata* technologií Oxford Nanopore. [The *De Novo* Genome Sequencing of *Crenicichla semifasciata* using Oxford Nanopore Technology. Bc. Thesis, in Czech.] – 40 p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic.

Annotation

The submitted bachelor thesis deals with *de novo* sequencing of genome of the cichlid species *Crenicichla semifasciata* on the Oxford Nanopore platform and its assembly with the Flye algorithm. The study further investigates the influence of sequencing depth on the BUSCO completeness assessment and N50 metrics of the obtained assembly.

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 13.04.2022

Adéla Mikešová

Poděkování

Nejprve bych chtěla moc poděkovat mému školiteli doktoru Piálkovi za trpělivost, cenné rady a odborné vedení mé práce. Neméně důležitý je i čas, který mi věnoval. Dále vděčím za podporu také mé rodině a blízkým, kteří mi byli po celou dobu vzniku práce oporou.

Klíčová slova

MinION, kompletnost BUSCO, adaptivní radiace, dlouhé sekvence

Key words

MinION, BUSCO completness, adaptive radiation, long reads

Obsah

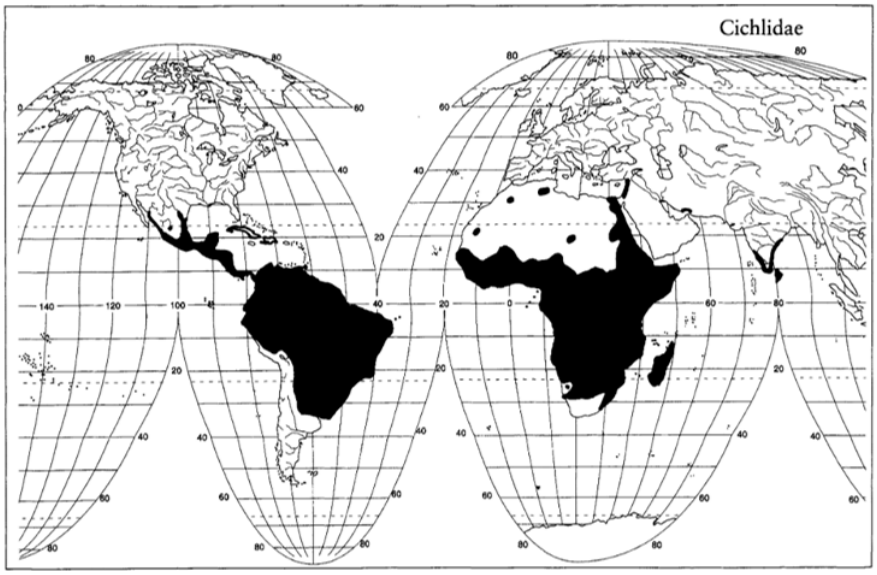
1 Úvod

1.1 Adaptivní radiace u neotropických cichlid

Hlavní náplní mé práce je osekvenování genomu jihoamerické cichlidy *Crenicichla semifasciata* technologií Oxford Nanopore. Získané asembly (sestava genomu) poslouží jako referenční genom při zpracování sekvenačních dat v rámci projektu mého školitele, který se zabývá studiem evolučních mechanismů paralelní speciace dvou druhových hejn cichlid rodu *Crenicichla* z povodí Río de la Plata. Sestavený genom bude dále zpracováván mojí kolegyní Bc. Danielou Kotalovou, která se zaměří na jeho vylepšení pomocí kombinace datasetů z dalších sekvenačních platforem.

Cichlidy jsou obecně známé jako jedna z druhově nejbohatších čeledí nejen sladkovodních ryb, ale obratlovců vůbec; v současnosti se jejich celková diverzita odhaduje na více než 2200 druhů (Berra 2001; Eschmeyer et al. 2020). Koblmüller (2011) dokonce uvádí, že včetně dosud nepopsaných taxonů může čeleď zahrnovat až 3000 druhů. Nejen díky zmíněné druhové pestrosti, ale též jejich morfologické diverzitě (obvykle související s potravní specializací) a variabilním vzorcům chování (hlavně epigamického – dvoření, páření, péče o potomstvo) jsou cichlidy častým předmětem evolučních i etologických studií. Zvláště cichlidy ve velkých afrických jezerech jsou považovány za přesvědčivý důkaz sympatrické speciace u zvířat ve spojení s jejich adaptivní radiací (Kocher 2004; Koblmüller et al. 2011; Salzburger and Meyer 2004; Salzburger et al. 2014).

Z hlediska geografie se cichlidy vyskytují především v tropických a subtropických regionech Afriky včetně Madagaskaru, Indie a Střední a Jižní Ameriky; jejich rozšíření lze tedy označit jako gondwanské (Smith and Chakrabarty 2004; obr. 1). Obývají nejen jezera a pomalu tekoucí řeky, ale existují i druhy, které se adaptovaly na život v rychleji tekoucích vodách (Kullander 1999; Nelson et al. 2016). Většina druhů cichlid představuje zástupce o velikosti do 20 cm, avšak některé neotropické druhy rodu *Cichla* mohou měřit i jeden metr (Nelson et al. 2016). Tvar těla je velmi různorodý v závislosti na habitatu daného druhu a zahrnuje nejrůznější formy od protáhle štikovité až po diskovitou. Jedním ze znaků odlišujících cichlidy od ostatních čeledí je faryngeální čelistní aparát (přeměněný čtvrtý žaberní oblouk) obvykle nesoucí ozubení napomáhající zpracování potravy (Rice and Lobel 2003; Mabuchi et al. 2007).

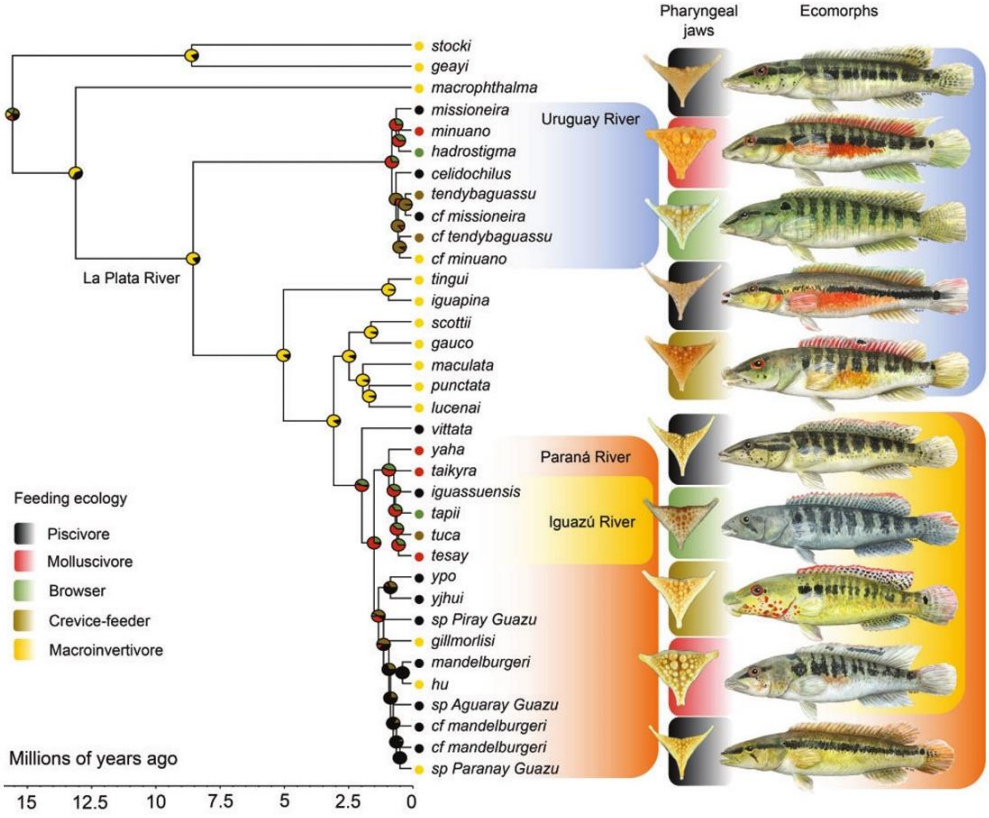


Obrázek 1: Geografické rozšíření čeledi Cichlidae; převzato z Berra (2001)

Většina prací dokladujících adaptivní radiace u cichlid se zaměřuje na prostředí velkých afrických jezer, která díky velké rozloze a hloubkové zonaci nabízejí nespočet relativně stabilních (v krátkodobém měřítku) habitatů podmiňujících druhovou různorodost. Méně pravděpodobnou se donedávna jevila významnější diverzifikace v mělkém prostředí podhorských neotropických řek s výraznou sezónní dynamikou a relativně chudou nabídkou ekologických nik. Recentní práce, na kterých se ve velké míře podílel náš výzkumný tým však prokázaly, že i v tomto prostředí může docházet k rychlé speciaci, což bylo dokumentováno např. u rodu *Herichthys* (Pérez-Miranda et al. 2018) a hlavně pak *Crenicichla* (Piálek et al. 2012; 2019a; Říčan et al. 2021).

Dvě výše zmíněná druhová hejna rodu *Crenicichla* z povodí La Plata byla objevena teprve nedávno v Río Iguazú, levostranného přítoku Paraná (Piálek et al. 2010; 2012; 2015; 2019a; 2019b) a Río Uruguay (de Lucena 2007; Lucenda and Kullander 1992). V obou řekách se vyskytuje několik syntopicky žijících ekologických forem (popsaných v současnosti jako druhy), které se v závislosti na přijímané potravě výrazně odlišují svou morfologií a jsou si navzájem blízce příbuzné (v rámci každé řeky). Pro zjednodušení můžeme čtyři hlavní ekomorfy popsat např. jako rybožravá, šnekožravá, všežravá a pyskatá. Obě druhové skupiny žijící v hydrologicky oddělených částech povodí La Plata si navzájem ale geneticky příbuzné nejsou (Piálek et al. 2012). Podobnost analogických morfologických druhů z obou skupin je přesto velmi výrazná (obr. 2), a proto začala být naším týmem studována jako další z možných modelových příkladů paralelní diverzifikace (Burress et al. 2018; 2022).

Pro odhalení detailních evolučních mechanismů byly v rámci projektu mého školitele osekvenovány celé genomy více než 400 jedinců obou druhových skupin. Pro zpracování těchto sekvenačních dat je ale potřeba mít k dispozici dostatečně blízký (kongenerický) genom, jaký ale zatím nebyl publikován. Jeho osekvenování se proto stalo hlavním cílem mojí bakalářské práce. Pro sekvenaci jsme vybrali druh *Crenicichla semifasciata*, který je z fylogenetického hlediska podobně blízký oběma druhovým skupinám, vyskytuje se ve stejném povodí a podařilo se od něj získat živý exemplář pocházející z odchytu v přírodě.



Obrázek 2: Morfologické typy vyskytující se ve druhových hejnech rodu *Crenicichla* v povodí Río de la Plata; převzato z Burress et al. (2022)

1.2 DNA – od objevu struktury až k jejímu sekvenování

Uvažovat o sekvenování poměrně velkého genomu v prostředí malého výzkumného týmu a časovém rámci bakalářského studia by bylo donedávna považováno za pouhé sci-fi. Překotný vývoj sekvenačních technologií v uplynulých letech ale odkrývá možnosti dříve netušené, a je proto na místě si historii sekvenování DNA trochu přiblížit, včetně vysvětlení základních pojmů.

Molekula DNA představuje jeden ze základních stavebních genetických kamenů a je v ní uložena veškerá genetická informace u buněčných organismů. Poprvé byla izolována z bílých krvinek J. F. Miescherem roku 1869. Jedná se o deoxyribonukleovou kyselinu, makromolekulu složenou z jednotek zvaných nukleotidy tvořených dusíkatou bází, jednoduchým sacharidem a fosfátovou skupinou. Dusíkaté báze mohou být dvojího druhu, a to pyrimidinové (C-cytosin, T-thymin) a purinové (A-adenin, G-guanin). Výsledky prvních analýz konformace DNA ukázaly, že se jedná o dvoušroubovici (tzv. duplex) – dvě dlouhá polynukleotidová vlákna (primární struktura DNA), navzájem spojená vodíkovými můstky mezi jednotlivými bázemi nukleotidů. Záhadu řetězců DNA vyřešila roku 1953 dvojice z Cavendishovy laboratoře experimentální fyziky University of Cambridge – James Watson a Francis Crick (Watson and Crick 1953).

Při spojování vláken DNA do duplexu platí princip komplementarity, podle kterého se v důsledku zaujmutí energeticky nejvýhodnější konformace adenin páruje s thyminem a cytosin spolu s guaninem (Chargaff 1950). Komplementarita je uplatňována při replikaci (zkopírování) molekuly DNA v buňce před rozdělením na dvě buňky dceřinné a analogický princip se uplatňuje *in vitro* i v některých platformách sekvenování DNA (tzv. sekvenování syntézou; Hu et al. 2021).

V současnosti rozlišujeme tři generace sekvenování. Dvě nejstarší klasické, ale tehdy enormně průlomové metody můžeme zařadit do skupiny sekvenování první generace (FGS, First Generation Sequencing). Vznikly téměř současně v 70. letech minulého století. O jednu z nich se zasloužil Maxam-Gilbert (Maxam and Gilbert 1977), kdy podle Munshi (2012) je tato metoda nazývána jako chemické sekvenování či jako analýza putovních bodů. Jednovláknová molekula je značena pouze na jednom svém konci radioaktivním izotopem fosforu, reakce se uskutečňuje ve čtyřech zkumavkách, kdy v každé zvlášť dochází ke štěpení jen v místě daných bází (Maxam and Gilbert 1977). Dle Heather a Chain (2016) se jednalo o první rozšířenou metodu, tudíž s její publikací můžeme datovat i vznik první generace sekvenování. Výhodou této metody byla poměrně velká přesnost, možnost sekvenovat fragmenty od jejich prvního nukleotidu nebo také podobná intenzita proužků na gelu nezávislá na jejich délce. Naopak nevýhodou bylo používání toxických a agresivních reagencií (hydrazin) a jejich vysoká teplota při degradaci (Maxam and Gilbert 1977).

O zavedení druhé z metod první generace se ve Velké Británii zasloužil Frederick Sanger a jeho tým (Sanger et al. 1977) objevem sekvenování s využitím terminace řetězců značenými dideoxynukleotidy. Tato metoda, dle objevitele nazývána Sangerova (ale též

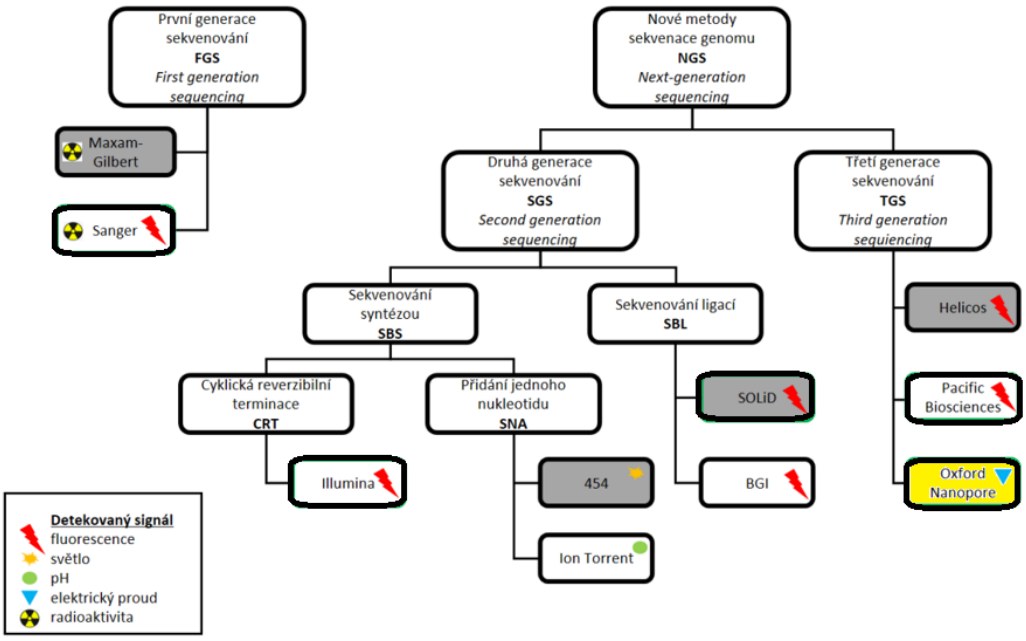
dideoxynukleotidová nebo enzymová), umožnila čtení řetězců DNA o délce až 1000 bází. Díky své jednoduchosti a přesnosti se stala na mnoho let víceméně jedinou používanou metodou sekvenování. Je používaná dodnes, i když lehce modifikovaná, coby zlatý standard na poli molekulární biologie. Sangerova technika se uplatnila v rámci průlomového sekvenování genomu bakteriofága phiX174 (zhruba 5000 bází DNA), bakterií (velikost genomu několik Mb), až po nejrůznější eukaryotické genomy (Heather and Chain 2016), včetně legendárního procesu sekvenování lidského genomu – Human Genome Project (Lander et al. 2001) dokončeného roku 2003.

1.3 Sekvenování dalších generací – NGS

Při sekvenování větších genomů se začala projevovat technická omezení sekvenátorů daná především jejich malou kapacitou, tj. časová náročnost a vysoká cena sekvenované báze. Tyto limitace měly za důsledek vývoj sekvenačních technik nové generace (zkratka NGS, Next Generation Sequencing), jejichž cílem bylo sekvenování co nejvíce zrychlit a zlevnit, a to prostřednictvím simultánního sekvenování obrovského množství (např. desítky mil.) různých fragmentů DNA zachycených současně na jediném sekvenačním elementu (čipu, destičce, buňce). Právě proto je možná vhodnějším označením nových generací sekvenátorů alternativně používaná zkratka MPS, masivně paralelní sekvenování (Massive Parallel Sequencing). K nejbouřlivějšímu vývoji nových technologií pravděpodobně došlo v letech 2004–2006. Jednotlivé platformy založené na velmi různorodých chemických i fyzikálních principech (Hu et al. 2021; Mardis 2013; přehled na obr. 3), což se odráží v rozdílné ceně, délce a chybovosti produkovaných sekvencí (Metzker 2009).

V rámci MPS technologií můžeme dále rozlišit dvě různé generace sekvenačních platforem (z hlediska historie tedy generaci druhou a třetí). Druhé generaci (SGS, Second Generation Sequencing) se také přezdívá technologie „krátkých čtení“ (angl. short reads). Produkované sekvence jsou totiž poměrně krátké, jejich délka se pohybuje v rozmezí 50-800 bází (v současnosti vzhledem k dominanci platformy Illumina spíše 100-450 bází), na druhé straně vykazují sekvence velmi malou chybovost, většinou < 1 % (Hu et al. 2021). Významným omezením této generace platforem je nemožnost určení sekvenované báze na základě signálu z jediné molekuly, a proto sekvenační postupy vždy zahrnují PCR amplifikaci DNA templátu prováděnou buď během přípravy sekvenační knihovny nebo v sekvenátoru samotném (případně obojí). Proces sekvenování je v důsledku tohoto přídavného kroku samozřejmě dražší, časově náročnější a potenciálně více chybový (Kumar et al. 2019; Mccombie et al. 2019).

Nejnověji zaváděná třetí generace sekvenátorů (TGS, Third Generation Sequencing) je charakterizována čtením sekvence DNA jedné jediné molekuly (proto též „single-molecule technologies“), což jednak odstraňuje potřebu předchozí PCR amplifikace, ale hlavně nabízí možnost sekvenování extrémně dlouhých úseků DNA (výjimečně až milióny bází). Detekce příslušné sekvence na základě signálu z jediné molekuly ale s sebou na druhé straně přináší značnou chybovost pohybující se nyní v rozmezí zhruba 5-15 % (Lin et al. 2021). Na sekvenačním trhu v současnosti dominují dva konkurenti této třetí generace: PacBio od společnosti Pacific Biosciences a platforma Oxford Nanopore Technologies (dále jen ONT) od stejnojmenné společnosti (Rhoads and Au 2015; Hu et al. 2021).



Obrázek 3: Přehled sekvenačních platforem. Členění podle generace sekvenování: FGS, NGS, SGS, TGS. Žluté zvýraznění – platforma využitá v předkládané práci. Šedá výplň – platformy, které se již nevyužívají a nejsou komerčně dostupné.; převzato a upraveno z Přistoupilová (2020)

1.4 Sekvenování na platformě Oxford Nanopore Technologies (ONT)

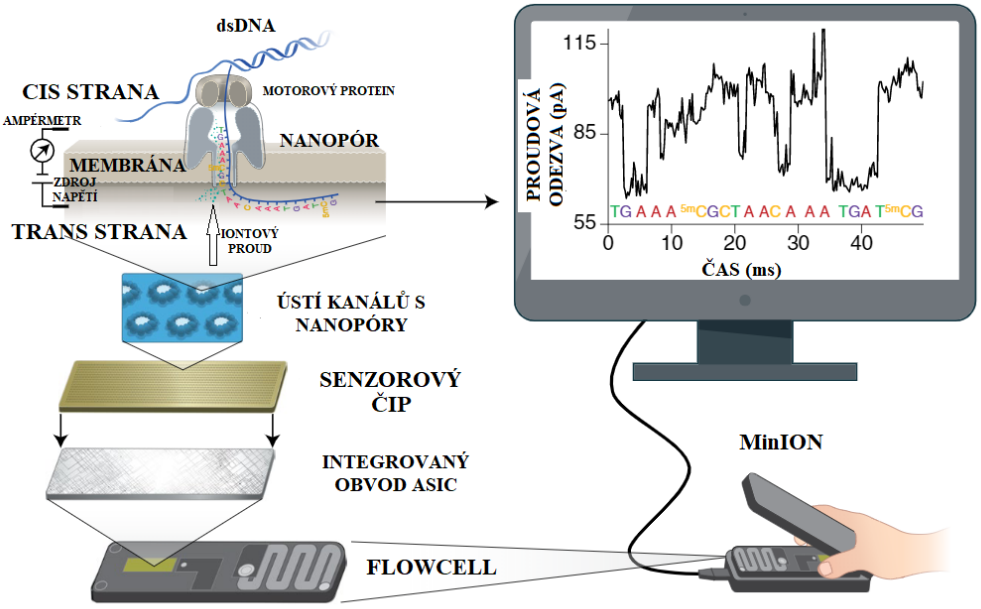
Jak již z názvu vypovídá, ONT a jejich zařízení jsou založeny právě na technologii nanopórů. Ačkoliv byl koncept průchodu molekuly skrz nanopór navržen již v roce 1989 (Deamer et al. 2016) a prvotní experimenty provedeny roku 1993 (Kasianowicz et al. 1996), první komerční produkt na principu nanopórového sekvenování se objevil na trhu až roku 2015 právě od ONT – zařízení MinION umožňující sekvenaci v reálném čase (Bayley 2014). Původní inspirací pro tuto technologii byly iontové kanály biomolekul, které řídí téměř všechny buněčné procesy (Bayley 2014; Lin et al. 2021).

Snahy ONT o převrat na poli vědy a biologické analýzy můžeme datovat k roku 2005, kdy byla společnost založena; popis principu nanopóru jako biosenzoru zmínil profesor Bayley z Oxfordské univerzity (jeden ze zakladatelů ONT) dokonce již o čtyři roky dříve (Howorka et al. 2001). Od svého založení prodělala společnost významný vývoj a vytvořila několik typů zařízení, která jsou schopna sekvenovat v nejrůznějších prostředích, bez vynaložení velkých finančních prostředků na přístroje i reagencie, s rychlým zpracováním vzorků a možností zobrazovat výsledky v reálném čase. Dostupné publikace ukazují, že byly přístroje použity při identifikaci nejen virových patogenů, monitorování eboly, životního prostředí, haplotypování, ale například i při výzkumu týkajícího se virů SARS-Cov2 (https://nanoporetech.com/).

1.4.1.1 Princip nanopóru

Nanopór je otvor velmi malých rozměrů (jak název napovídá v řádu nanometrů) sloužící k průchodu molekul buněčnou membránou. Dle původu rozeznáváme póry biologické nebo uměle připravené, syntetické – založené na bázi silikonu (Lin et al. 2021). Podstatou biologického póru je proteinový transmembránový kanál (vel. 1 nm = 100 000 menší než lidský vlas) zanořený v lipidové dvojvrstvě vyskytující se v každé živé buňce. Patří mezi ně například draslíkové kanály udržující aktivitu v srdečních myocytech nebo proteinový pór tvořený α-hemolysinem (α-HL), který je produkovaný lidským patogenem *Staphylococcus aureus* jako endotoxin (Lin et al. 2021). Právě nanopóry α-HL zanořené do elektro-rezistentní lipidové membrány používá ONT. Během let prošly značnou obměnou a vývojem. Posledním představeným prototypem je nanopór označovaný jako R10.3, avšak jeho přesné složené není známé, jelikož ho ONT doposud neodhalilo (Wang et al. 2021; Oxford Nanopore Technologies)

V sekvenační buňce jsou nanopóry umístěny v elektricky rezistentní membráně, která separuje dvě komory vyplněné elektrolytem. Mezi komory je přivedeno elektrické napětí, které je hnací silou pro translokaci elektricky nabitých biomolekul skrze pór. Během průchodu nanopórem způsobují jednotlivé báze (v případě DNA) různou změnu hustoty elektrického náboje uvnitř póru (v řádu stovek pikoampérů), která je prostřednictvím citlivého snímače detekována a zaznamenána. Modulace proudu jednotlivými nukleotidy odpovídá jejich vlastnostem jako jsou náboj, struktura a konformace (Lu et al. 2016; Deamer et al. 2016), obr. 4).



Obrázek 4: Princip sekvenování na platformě Oxford Nanopore; převzato z Wang et al. (2021)

1.4.2 Sekvenátor MinION

Prvním ze sekvenátorů třetí generace uvedených na trh byl přístroj MinION. Kromě tohoto prvenství se díky svým rozměrům 10,5 x 2,3 x 3,3 cm a hmotnosti 87 g (https://nanoporetech.com/) pyšní rovněž titulem nejmenšího sekvenačního aparátu na světě (Lu et al. 2016). Přístroj vykazuje značnou odolnost a může být používaný jak v laboratoři, tak v terénním výzkumu. Sekvenátor MinION byl v rámci reklamních kampaní dokonce úspěšně provozován ve vesmíru (Castro-Wallace et al. 2017) nebo v Suchých údolích v Antarktidě (Johnson et al. 2017). Tento systém umožňuje nejen čtení sekvencí DNA, ale potenciál má i v oblasti analýzy RNA, iontů, proteinů nebo i léčiv a polymerů (Feng et al. 2015). Další benefit představuje zanedbatelná pořizovací cena a zároveň nízká cena detekované báze. Díky sekvenování v reálném čase poskytuje okamžitý přístup k výsledkům, což je výhodné kupříkladu při identifikaci patogenů, stanovení rezistence k antibiotikům aj. (https://nanoporetech.com/).

Jádrem systému je sekvenační buňka („flowcell“) s 512 měřícími kanály v membráně a celkově 2048 nanopóry (Jain et al. 2016). Na každý kanál jsou tak napojeny čtyři nanopóry, avšak aktivní je v každém čase pouze jeden z nich. Každý nanopór má napojení na vlastní elektrodu. Pokud dojde k „zablokování“ póru průchodem více molekul DNA za sebou a ztrátě jeho funkčnosti, bude v dalším časovém cyklu (standardně 1.5 hod.) nahrazen dalším pórem ze čtveřice. Průměrná rychlost průchodu molekul DNA nanopóry je 400 bází za sekundu. Přístroj má uváděnou max. sekvenační kapacitu až 50 Gb (v praxi

spíše 10 až 20 Gb) během 72 hodin (Lin et al. 2021). Délka produkovaných sekvencí se, v závislosti na kvalitě sekvenovaného templátu, běžně pohybuje v desítkách až stovkách kilobází, výjimečně až jednotkách megabází (https://nanoporetech.com/).

K ovládání sekvenátoru je firmou ONT dodáván řídící software MinKNOW spolu s vestavěným programem („basecallerem“, viz dále) Guppy, který umožňuje online převádění snímaných elektrických charakteristik na sekvence DNA. Program nemá nijak zvlášť velké nároky na hardware používaného řídícího počítače (velmi doporučovaný je rychlý SSD disk) a může být provozován pod libovolným operačním systémem (Windows, Mac, Linux). Řídící PC je propojen se sekvenátorem prostřednictvím běžně používaného portu USB 3.0 (https://nanoporetech.com/).

Vlastnímu sekvenování předchází úprava izolované DNA do formy tzv. sekvenační knihovny, což je směs fragmentů DNA s připojenými adaptéry (technická DNA zajišťující průchod sekvenátorem) a dalších potřebných komponent. Připravená knihovna se bezprostředně před zahájením sekvenace nanáší do příslušného portu v sekvenační buňce. Důležité je dbát na rychlost nanášení knihovny, jelikož je velmi snadné s jejím roztokem zanést i vzduchové bubliny, který by měly za následek poškození nanopórů v daném místě a významné snížení sekvenační kapacity (Karamitros and Magiorkinis 2018; https://nanoporetech.com/).

1.5 Bioinformatické zpracování získaných sekvencí

Nedílnou součástí práce s genomickými daty je jejich zpracování. Vzhledem k obrovskému objemu generovaných dat probíhají analýzy téměř výhradně prostřednictvím výkonných výpočetních klastrů umožňujících paralelní zapojení velkého počtu procesorů a vyhrazení rozsáhlé operační paměti. Používané expertní bioinformatické postupy vyžadují znalost práce v příkazových řádcích unixového prostředí a nezbytnou součástí je práce také znalost programovací terminologie a schopnost nacházet pro jednotlivé projekty specifická řešení (Wang et al. 2015; 2021).

V případě mé práce zabývající se *de novo* sekvenováním a složením genomu cichlidy *C. semifasciata* zahrnovaly bioinformatické analýzy zejména převod elektrických charakteristik zaznamenaných sekvenátorem Nanopore na sekvence DNA, jejich úpravu a kvalitativní filtraci, dále sestavení sekvencí do delších celků na základě rozpoznání jejich vzájemných překryvů (vlastní asemblování genomu) a v poslední fázi taktéž vyhodnocení vzniklých sestav.

1.5.1 Basecalling

Bioinformatický proces, při kterém jsou primární data ze sekvenátoru ve formě elektrických charakteristik zaznamenaných při průchodu molekul DNA nanopórem převáděna na sekvence, se v angličtině nazývá těžko přeložitelným výrazem „basecalling“ („volání bází“). Primární data jsou sekvenátorem ukládána ve formátu FAST5 (modifikace formátu používaného v obecné informatice HDF5, Hierarchical Data Format) a vzhledem ke tvaru proudových křivek se jim rovněž říká „squiggles“ („klikyháky“). K převodu na sekvence se používá program Guppy pracující na bázi samoučících neuronových sítí vyvíjený přímo společností ONT, který se s rostoucím poznáním a množstvím osekvenovaných templátů neustále zdokonaluje. Velkou výhodou je, že se primární data dají s příchodem novější verze software opětovně „basecallovat“ a snížit tak významně jejich chybovost (Senol Cali et al. 2018; Wick et al. 2019).

1.5.2 Sestavení genomu a jeho vyhodnocení

Již při vlastním „basecallingu“ se provádí filtrace sekvencí na určitou minimální kvalitu (pravděpodobnost správného přečtení bází v sekvenci). Další úpravy pak zahrnují odstranění sekvenačních adaptérů (úseky technické DNA připojené při přípravě knihovny), odfiltrování příliš krátkých sekvencí, spojení všech získaných sekvencí do jediného souboru a případně i znáhodnění jejich pořadí v souboru (Senol Cali et al. 2018; Wick et al. 2019).

Při vytváření vlastního asembly (sestavy) genomu dochází ke spojování sekvencí na základě definice jejich překryvu či pozice. Nejbližším vyšším celkem je tzv. „contig“ (z anglického contiguous, souvislý), což je konsenzuální sestava překrývajících se úseků DNA. Ještě vyšším celkem je pak „scaffold“, což je sestava nepřekrývajících se contigů, u kterých ale známe jejich přibližnou vzájemnou polohu a orientaci. Scaffold tedy z definice obsahuje tzv. „gapy“ (mezery), tj. úseky neosekvenovaných bází (Wee et al. 2019).

Kvalitu získané sestavy genomu je možno hodnotit na základě několika parametrů. Čistě technické hodnocení zahrnuje posouzení délky získaných scaffoldů; čím menší počet co nejdelších scaffoldů asembly obsahuje, tedy čím méně je genom fragmentován, tím lépe. Pro toto zhodnocení je nejčastěji používána metrika N50, která vyjadřuje minimální délku scaffoldů ve kterých (a delších) je sestavena právě polovina osekvenovaných bází (Simão et al. 2015; Lu et al. 2016; Seppey et al. 2019)

Z biologického hlediska se kvalita asembly vyhodnocuje nejčastěji stanovením tzv. kompletnosti BUSCO (Benchmarking Universal Single-Copy Orthologs). Tento bioinformatický nástroj ve spojení s databází vyhledává v analyzované sestavě

předpokládanou množinu ortologních genů, univerzálních pro danou linii organismů, a je tak hodnověrným měřítkem kompletnosti přinejmenším kódující části genomu (ortologní geny jsou geny získané od společného předka daných organismů, které si zachovávají stejnou funkci v průběhu evoluce; (Simão et al. 2015; Seppey et al. 2019; Yang et al. 2021).

1.6 Cíle práce

Hlavními cíli mojí bakalářské práce tedy bylo:

1. Osekvenovat a sestavit genom cichlidy *Crenicichla semifasciata* pomocí platformy Oxford Nanopore za účelem získání referenčního genomu pro náš běžící projekt.
2. Vyhodnotit vliv objemu vstupních sekvencí na kvalitu vytvářených sestav genomů a navrhnout optimální hloubku sekvenace.

2 Materiály a metody

2.1 Odběr krve z kaudální cévy a izolace genomové DNA

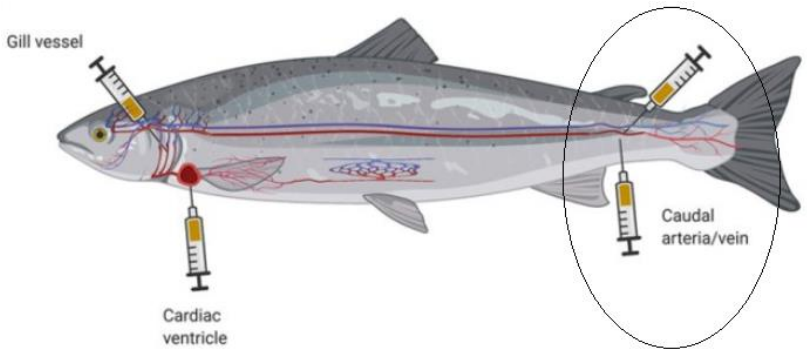
Kvalita genomového sestavení je primárně závislá na kvalitě vzorku DNA. Jelikož dochází k postupné fragmentaci DNA po odběru vzorku z organismu, je nutné použít pro izolaci tkáň čerstvou, která umožní sekvenovat delší úseky (Jain et al. 2016; Permenter et al. 2015). Faktorem pro snadnější sestavení sekvencí jsou právě dlouhé úseky v délce několika desítek nebo stovek tisíc bází. Tkáň jako je krev má vyšší obsah DNA než například vzorek ploutve. Co se týče rybích erytrocytů je krev výhodnou variantou, na rozdíl od lidských erytrocytů, jelikož jsou jaderné (Stoskopf 1993; de Lannoy et al. 2017). Na základě těchto skutečností byla jako materiál pro sekvenaci vybrána právě krev. Obvykle se provádí punkce v oblasti žaber, srdce nebo kaudální cévy (obr. 5).

Použité materiály

- Injekční stříkačka
- 1,5 ml zkumavky Eppendorf DNA LoBind
- Lyzační pufr z kitu QIAGEN (katalogové číslo: 67563)

Postup odběru krve

K odběru jsem zvolila ocasní cévu, jehlu jsem zavedla pod úhlem zhruba 45 °, odhadem ve vzdálenosti 1 cm kaudálně za řitní ploutví. Celkem bylo odebráno necelých 700 µl krve. Krev jsem ihned rozpipetovala do plastových zkumavek Eppendorf obsahujících lyzační pufr.



Obrázek 5: Metody odběru krve u ryby, převzato z (Narayan et al. 2021)

Vzhledem k charakteru experimentu a sekvenační techniky ONT umožňující číst vlákna délky až 2 Gb, bylo hlavním požadavkem izolace získat co nejdelší a nejkvalitnější

fragmenty DNA. Izolační kit MagAttract HMW DNA Mini Kit od firmy QIAGEN byl zvolen proto, že je schopný izolovat molekuly DNA o délce min. několik desítek tisíc bází (HMW = high molecular weigth). Použitý kit využívá princip separace na magnetických kuličkách. Tato metoda poskytuje vysoké výtěžky kvalitní DNA díky šetrnému postupu izolace a extrahovaná DNA je tak zvlášť vhodná pro TGS sekvenování (https://nanoporetech.com/).

Použité materiály

- MagAttract HMW DNA Mini Kit QIAGEN, katalogové číslo:67563
- Qubit Fluorometr (Invitrogen)
- E-gel (Invitrogen)
- Běžné laboratorní vybavení

Postup izolace genomové DNA

Ke směsím odebrané krve s lyzačním pufrem jsem do každé zkumavky přidala proteinasu K (20 μl), RNázu A (4 μl) a AL Buffer (150 μl). Po důkladném promíchání se připravené zkumavky lyzovaly při pokojové teplotě (dále pouze RT) 30 min. Po zcentrifugování jsem do vzorků přidala suspenzi magnetických kuliček MagAttract G (15 μl) a MB Buffer (280 μl). Vzorky jsem ponechala inkubovat (3 min, RT) a dala na magnetický stojánek za účelem vytvoření pelety doprovázené následným odsátím supernatantu. V dalším kroku jsem přidala PE Buffer (700 μl) a zkumavky umístila na třepačku. Promytí pomocí PE Buffer jsem ještě jednou zopakovala. Peletu jsem poté na magnetickém stojánku promyla PCR vodou (700 μl) a po opětovném vytvoření pelety supernatant opět odsála. Tento krok s promytím a odsátím jsem ještě jednou zopakovala. Závěrem jsem přidala AE Buffer (100 μl) a po inkubaci jsem extrakt HMW DNA (supernatant) přenesla do čisté plastové zkumavky. Na základě ověření kontroly kvality a kvantity získané DNA jsme vybrali vzorky vhodné k sekvenování (Qubit fluorometr a agarózový gel).

2.2 Příprava sekvenační knihovny

Celková příprava sekvenační knihovny pozůstává z kroků opravy DNA s přípravou jejích konců (časová náročnost zhruba 35 min) a ligace adaptérů (30 min). Pro tvorbu knihovny je doporučeno použít alespoň 1 μg (případně 100–200 fmol) HMW DNA.

Použité chemikálie

Tabulka 1: Přehled jednotlivých komponent obsažených v ligačním kitu SQK-LSK110

Zkratka	Reagencie	Zkratka	Reagencie
DCS	DNA Control Strand	SBII	Sequencing Buffer II
AMX-F	Adapter Mix F	EB	Elution Buffer
LNB	Ligation Buffer	LBII	Loading Beads II
SFB	S Fragment Buffer	LS	Loading Solution

- AMPure magnetické kuličky
- NEBNext FFPE Reapir Mix
- NEBNext End repair/ dA-tailing Module
- NEBNext Quick Ligation Module
- Voda bez nukleázové aktivity
- 1,5 ml zkumavky Eppendorf DNA LoBind
- Běžné laboratorní vybavení

2.2.1 Oprava DNA a příprava konců

Materiál byl použit na základě standardního protokolu v kitu Ligation Sequencing Kit (SQK-LSK110; tab. 1), Oxford Nanopore Technologies. DNA jsem naředila PCR vodou na doporučenou koncentraci. 1 µg genomové DNA jsem přenesla do LoBind zkumavky (dále jen zkumavka) a posléze přidala 49 µl vody. V tenkostěnné 0,2 ml zkumavce jsem smíchala následující reagencie (tab. 2), směs jemně promíchala pipetováním (se špičkou s větším vstupním otvorem P1000, kvůli fragmentaci DNA) a stočila. Pomocí termocykléru byla směs inkubována při 20 °C po dobu 5 min a následovně vystavena 65 °C, také 5 min, kvůli inaktivaci použitých enzymů. Vzorek opravené DNA jsem přenesla do↓čisté zkumavky a přidala 60 µl resuspendovaných kuliček AMPure. Zkumavku jsem opět promíchala proklepáním a vložila na 5 min na hula rotátor, poté přenesla na magnetický stojánek a vzniklou peletu promyla 200 µl čerstvě připraveného 70 % ethanolu. Promývání ethanolem jsem ještě jednou zopakovala. Vzorek jsem krátce stočila, přemístila

Tabulka 2: Reagencie potřebné na opravu DNA a přípravu konců obsažené v kitu SQK-LSK110

Reagencie	Objem (µl)
DCS	1
DNA	47
NEBNext FFPE DNA Repair Buffer	3,5
NEBNext FFPE DNA Repair Mix	2
Ultra II End-prep reaction buffer	3,5
Ultra II End-prep enzyme mix	3
Celkový objem	60

na magnetický stojánek a odpipetovala zbytek supernatantu ze dna zkumavky. Peletu jsem sušila ca. 30 s (při delším sušení může peleta vyschnout, což by mělo za následek fragmentaci DNA). Peletu jsem promyla 61 µl PCR vody, zkumavku umístila na magnetický stojánek a eluát opatrně odpipetovala do čisté zkumavky. 1 µl eluátu byl odebrán pro změření kvality a koncentrace DNA pomocí Qubit fluorometru. V tomto okamžiku je před následujícím krokem ligace adaptérů možno vzorek uchovat, nejlépe pouze přes noc, v chladničce při teplotě 4 °C a další den pokračovat v experimentu.

2.2.2 Ligace adaptérů a přečištění knihovny

Materiál byl použit na základě standardního protokolu v kitu Ligation Sequencing Kit (SQK-LSK110; tab. 1), Oxford Nanopore Technologies. Pakliže chceme obohatit DNA fragmenty delší nebo rovnající se 3 kb, použijeme Buffer LFB, pro zachování kratších délek se používá SFB. Já jsem pro charakter mého experimentu zvolila LFB. Nejprve jsem ve zkumavce smíchala reagencie (tab. 3), spolu s AMPure kuličkami vázajícími DNA (40 µl) a inkubovala po dobu 5 min na hula rotátoru, přenesla na magnetický stojánek. K promytí kuliček jsem použila 250 µl LFB, ponechala je vytvořit peletu a opět odstranila supernatant. Předešlý krok byl ještě jednou zopakován. Následovalo sušení vzorku (30 s) a promytí EB (15 µl) s inkubací v termocykléru (10 min, 37 °C; regenerace dlouhých fragmentů). Po ponechání zkumavky na magnetickém stojánku a vytvoření pelety jsem extrakt obsahující DNA knihovnu přenesla (15 µl) do čisté zkumavky a zbylý 1 µl jsem použila na změření pomocí fluorometru Qubit.

Tabulka 3: Reagencie potřebné na ligaci adaptérů obsažené v kitu SQK-LSK110 kitu

Reagencie	**Objem (ml)**
Vzorek DNA z předchozího kroku	60
LNB	25
Quick T4 DNA Ligase	10
AMX-F	5
Celkový objem	100

2.3 Sekvenování na platformě Oxford Nanopore přístrojem MinION

2.3.1 Nanesení vzorku na sekvenační buňku

Použité materiály

- Flow Cell Priming Kit (obsahuje Flush Buffer [FB] a Flush Tether [FLT])
- Běžné laboratorní vybavení

Nejprve jsem připravila tzv. priming mix (30 µl Buffer FLT napipetováno přímo do zkumavky FB). V této fázi jsem si připravila přístroj MinION, kdy je před každou

sekvenací nutné do přístroje umístit kompatibilní sekvenační buňku (flowcell) a poznamenat si její typ. K nanesení knihovny je nutné otevřít kryt tzv. priming portu, kam se bude připravená knihovna nanášet. Pipetováním jsem nanesla 800 μl priming mixu do sekvenační buňky přes priming port. V nové zkumavce jsem si připravila knihovnu smícháním výše zmíněných reagencií (tab 4). K dokončení aktivace sekvenační buňky jsem do priming portu nanesla dalších 200 μl priming mixu a do příslušného vzorkovacího portu knihovnu (75 μl).

Tabulka 4:Reagencie potřebné k dokončení nanesení směsi na sekvenační destičku obsažené v kitu SQK-LSK110

Reagencie	**Objem (ml)**
SBII	37,5
LBII	25,5
DNA knihovna	12
Celkový objem	75

2.3.2 Ovládání sekvenátoru MinION

Sekvenátor MinION se ovládá nejčastěji z osobního počítače prostřednictvím aplikace MinKNOW (Ip et al. 2015), pomocí které jsem nastavovala parametry sekvenace a spouštěla přístroj. Výhodou tohoto software je možnost monitorovat sekvenační běh v reálném čase a v kterémkoli okamžiku jej přerušit. Tato alternativa se využívá hlavně při neuspokojivých výsledcích (malý výtěžek nebo krátké sekvence) zaznamenaných během prvních několika hodin procesu. V tom případě je optimální strategií sekvenační běh zastavit, sekvenační buňku ošetřit promývacím kitem (EXP-WSH003 nebo novějším EXP-WSH004) a provést nanesení nové knihovny (nebo optimálního množství té stávající). Během sekvenování totiž dochází k postupné degradaci nanopórů a sekvenační kapacita buňky se tak postupně nevratně snižuje; po ca. 72 hodinách sekvenačního běhu dochází k jejímu úplnému vyčerpání (žádné funkční nanopóry). Včasné přerušení sekvenace tak umožní využít nevyčerpanou kapacitu sekvenační buňky pro nový experiment. Díky této možnosti jsme provedli celkem 16 experimentů (sekvenačních běhů) s využitím 8 ks sekvenačních buněk, z toho 6 ks typu FLO-MIN106 a 2 ks novějšího typu FLO-MIN111 (tab. 5).

Tabulka 5: Přehled jednotlivých sekvenačních experimentů

Sekv. běh	ID	Datum startu	Doba běhu [hh:mm]	Sekv. buňka - typ	Sekv. buňka - ID	Knihovna	Kit pro přípravu	Izolát	Poznámka
1	Pilot00	20.12.2019	72:00	FLO-MIN106	FAL27071	LIB00	SQK-LSK109	1-0	expired 12/2019
2	Adel01	25.2.2021	48:00	FLO-MIN111	FAP16694	LIB01	SQK-LSK110	1-1	new
3	Adel02	3.3.2021	52:16	FLO-MIN106	FAL27129	LIB02	SQK-LSK110	1-1 + 1-2	expired 12/2019
4	Adel03	5.3.2021	69:12	FLO-MIN106	FAL26934	LIB03	SQK-LSK110	1-4	expired 12/2019
5	Adel04	8.3.2021	67:20	FLO-MIN111	FAP15580	LIB04	SQK-LSK110	1-5 + 1-6	new
6	Adel05	11.3.2021	14:24	FLO-MIN106	FAL27129	LIB05	SQK-LSK110	2-5	washed
7	Adel06	12.3.2021	72:01	FLO-MIN106	FAL27129	LIB05 - restart	SQK-LSK110	2-5	restarted
8	Adel07	19.3.2021	24:06	FLO-MIN106	FAP33242	LIB06	SQK-LSK110	2-5	new
9	Adel08	20.3.2021	27:21	FLO-MIN106	FAP33242	LIB06 - restart	SQK-LSK110	2-5	restarted
10	Adel09	21.3.2021	16:41	FLO-MIN106	FAP51522	LIB07	SQK-LSK110	2-3	new
11	Adel10	22.3.2021	20:17	FLO-MIN106	FAP51522	LIB08	SQK-LSK110	2-11	washed
12	Adel11	23.3.2021	49:26	FLO-MIN106	FAP33242	LIB09	SQK-LSK110	2-4	washed
13	Adel12	25.3.2021	19:24	FLO-MIN106	FAP54211	LIB09	SQK-LSK110	2-4	new
14	Adel13	26.3.2021	2:48	FLO-MIN106	FAP54211	LIB10	SQK-LSK110	2-11	washed
15	Adel14	26.3.2021	66:21	FLO-MIN106	FAP54211	LIB10+LIB11	SQK-LSK110	2-11	no wash, library added
16	Adel15	29.3.2021	23:26	FLO-MIN111	FAP15580	LIB11	SQK-LSK110	2-11	washed

2.4 Bioinformatická analýza

Zpracování sekvenačních dat bylo vzhledem k jejich objemu a náročnosti analýz prováděno v prostředí České Národní Gridové Infrastruktury Metacentrum (https://www.metacentrum.cz/cs/), jehož kapacita je přístupná zdarma všem akademickým pracovníkům a studentům členů sdružení CESNET (https://www.cesnet.cz/). Veškeré operace a analýzy byly prováděny v operačním systému Linux.

2.4.1 Basecalling

Primární data ze sekvenátoru Nanopore MinION ve formátu FAST5 zahrnující záznamy elektrických charakteristik jednotlivých fragmentů DNA byly převáděny do podoby sekvencí ve formátu FAST5 v procesu zvaném basecalling s použitím programu Guppy v4.4.1. Formát FASTQ obsahuje mimo vlastní sekvence jednotlivých bází (A, C, G, T) pro každou převedenou bázi i vyjádření kvalitativního skóre (Q-score), které vyjadřuje pravděpodobnost, s jakou byla každá jednotlivá báze osekvenována. Během tohoto procesu byly odfiltrovány sekvence s Q-score menším než 7, což odpovídá pravděpodobnosti správného přečtení báze ca. 80 %. Program byl spouštěn s následujícími parametry:

```
guppy_basecaller \
 -i nanopore_runs/Adel01/fast5 \
 -s basecalling_results/Adel01 \
 -c dna_r9.4.1_450bps_hac.cfg \
 --qscore_filtering --min_qscore 7 \
 -x 'auto' -q 0 --compress_fastq \
 --calib_detect --chunk_size 1000 --num_callers 2 --chunks_per_runner 300\
 --chunks_per_caller 1000 --gpu_runners_per_device 4


# -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)
# --qscore_filtering = zapnutí procesu filtrování na základě Q-skóre
# --min_qscore 7 = minimální požadované skóre
# většina ostatních parametrů se týká nastavení grafické karty
#   optimalizovaného pro analýzu na výpočetních strojích v Metacentru
```

Vzhledem k výpočetní náročnosti procesu, který je založen na neuronových sítích a využívá grafické procesory, byl basecalling prováděn zvlášť pro data z každého sekvenačního

experimentu (případně byla v některých případech data z experimentu ještě dále rozdělena na několik částí).

2.4.2 Úprava a filtrace získaných dat

Výstupní soubory z jednotlivých sekvenačních běhů (obsahující sekvence ve formátu FASTQ) byly v případě rozdělení na několik částí sloučeny do jediného souboru linuxovým příkazem *cat*.

K vyhodnocení hlavních technických parametrů výstupních souborů byl použit softwarový balík Seqkit (Shen et al. 2016) pro analýzu a úpravu genomových dat. Základní statistiky (N50, počet sekvencí, počet bází atd.) byly zjištěny prostřednictvím příkazu:

```
seqkit stats -a *.fastq.gz
```

Součástí získaných sekvencí DNA jsou i sekvence tzv. adaptérů, což jsou úseky technické DNA nezbytné pro průběh sekvenačního procesu a připojené k templátu v průběhu přípravy DNA knihovny. Některé asemblery (programy pro sestavení genomu) včetně mnou používaného programu Flye zajistí v prvních krocích analýz jejich automatickou detekci a odstranění. Alternativní cestou je jejich odstranění prostřednictvím specializovaných programů ještě před vlastním procesem sestavení genomu. Tuto cestu jsem zvolila i já a to z následujících důvodů: 1) Parametry pro nalezení sekvence adaptéru (např. jak dlouhé úseky konců sekvencí procházet, jaké sekvence hledat a s jakou podobností) a jejich odstranění je možné přesně a podrobně definovat a zvolit přísnější podmínky než nabízí asembler (tj. vzhledem k vysoké chybovosti čtení na platformě Nanopore odstranit i méně podobné sekvence); 2) Jedním ze známých artefaktů Nanopore sekvenování je tvorba tzv. chimerických sekvencí DNA vzniklých spojením dvou různých fragmentů DNA a jejich osekvenování coby fragmentu jediného. Tyto chiméry se ale dají ve většině případů odhalit díky přítomnosti adaptéru uprostřed získané sekvence a specializované programy umožňují v tom případě adaptér odstranit a sekvenci v tomto místě rozdělit na dvě (nedochází tak ke ztrátě přečtených dat).

Pro odstranění adaptéru jsem používala program Porechop (https://github.com/rrwick/Porechop) spouštěný s následujícími parametry:

```
porechop-runner.py \
  -i Adel01_hac7.fastq.gz \
```

```
  -o Adel01_hac7_trim.fastq.gz \
  --format fastq.gz \
  --end_size 200 \
  --end_threshold 60

# -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
# --end_threshold = min. percentuální podobnost sekvence, aby byla
#   považována za adaptér
# odstranění adaptérů uprostřed sekvence a její rozdělení na dvě je
#   zajištěno v defaultním nastavení
```

Soubory se sekvencemi z jednotlivých experimentů byly po odstranění adaptérů spojeny do jediného datasetu pomocí linuxového příkazu *cat*. Vzhledem ke skutečnosti, že jednotlivé sekvenační běhy poskytly v závislosti na zdařilosti experimentu sekvence různé kvality (tj. především různé průměrné délky N50) a tato heterogenita by neumožnila porovnávat výsledná asembly získaná z jednotlivých částí datasetu (viz dále), bylo pořadí sekvencí ve sloučeném souboru znáhodněno pomocí programu Seqkit:

```
seqkit shuffle all_trim.fastq.gz > all_trim_random.fastq.gz
```

2.4.3 Sestavování genomu z různě velkých datasetů

Mimo vlastního sestavení kvalitního genomu cichlidy druhu *C. semifasciata* bylo dalším cílem mojí práce i zkoumání vlivu velikosti vstupních dat na kvalitu sestaveného genomu. Z datasetu s odstraněnými adaptéry jsem odfiltrovala sekvence kratší než 1000 bází (pro sestavování genomu neužitečné, pouze zvyšují výpočetní nároky) pomocí programu Seqkit:

```
seqkit seq -m 1000 -w 0 all_trim_random.fastq.gz > \
  all_trim_random_min1000.fastq.gz

# -m = minimální délka filtrovaných sekvencí
# -w 0 = každá sekvence bude v souboru umístěna na jedinou řádku bez
#   zalamování na určitou definovanou šířku
```

Výsledný objem sekvenačních dat činil po odstranění adaptérů a odfiltrování krátkých

sekvencí celkem ca. 86.5 Gb; vzhledem k předpokládané velikosti genomu *C. semifasciata* ca. 850-900 Mb jsme tak celkem studovaný genom osekvenovali v průměru stokrát (tzv. hloubka čtení = 100x).

Pro zkoumání závislosti kvality sestaveného genomu na objemu vstupních dat jsme celkový dataset rozdělili jednoduchým "odkrajováním" různě velkých částí vždy od začátku datasetu prostřednictvím linuxového příkazu *head*. Získali jsme tak deset různě velkých datasetů odpovídajících hloubce čtení 10x, 20x, 30x ... 100x.

Pro vlastní sestavování genomu jsem použila program Flye (Kolmogorov et al. 2020), což je *de novo* asembler určený pro vytváření genomů z dlouhých neamplifikovaných sekvencí s vysokou chybovostí (tj. sekvencí získaných z TGS platforem Oxford Nanopore a Pacific Biosciences). Výběr asembleru byl proveden na základě výsledků týmové kolegyně D. Kotalové, která se mj. zabývá porovnáním jednotlivých asemblerů pro dlouhé sekvence; program Flye v tomto srovnání nabízel nejkvalitnější sestavené genomy z hlediska délky sestavených scaffoldů (N50) i kompletnosti genomu (BUSCO), a navíc oproti jiným programům vyžaduje nejmenší výpočetní zdroje (hlavně procesorový čas, místo na disku a velikost operační paměti).

Asembler Flye byl spouštěn s následujícími parametry:

```
flye --nano-raw all_trim_min1000_random_coverage10.fastq.gz \
  --out-dir assembly_coverage10_rep1 \
  --threads 16 --iterations 1

# --nano-raw = vstupní data z platformy Nanopore, data použita bez
#   předběžného čištění (odstranění sekvenačních chyb)
# --out-dir = název adresáře pro uložení výsledků
# --threads = počet CPU použitých pro výpočet
# --iterations 1 = asembler provede jedenkrát rovněž vyčištění (tzv.
#   polishing) výsledné sestavy genomu; proces čištění je možné iterativně
#   opakovat
```

Vzhledem k tomu, že proces sestavování genomu je zčásti ovlivněn náhodnými faktory (např. volba pořadí sekvencí pro zapojení do sestavy), výsledná asembly vytvářená ze stejných datasetů se i při použití stejných programových parametrů liší. Proto bylo sestavení genomu provedeno celkem třikrát pro každý dataset.

2.4.4 Zhodnocení získaných genomových assembly

Kvalita získaných sestav genomů byla zhodnocena jednak z technického hlediska, tj. především jaká byla délka a počet výsledných scaffoldů, a dále z hlediska kompletnosti genomu posuzované na základě vyhledávání univerzálních ortologních genů programovým prostředkem BUSCO.

Technické statistiky každé sestavy genomu (N50, průměrná délka, počet sekvencí, množství gapů ad.) byly získány pomocí programu Seqkit:

```
seqkit stats -a *.fastq.gz
```

Analýza kompletnosti genomu prostřednictvím BUSCO byla spouštěna příkazem:

```
busco -i assembly_coverage10_rep1.fas \
  -l actinopterygii_odb10 \
  -o busco_cov10_rep1 \
  -m genome
```

Kompletnost byla odhadována na základě porovnání s databází ortologních genů univerzálních pro třídu paprskoploutvých ryb (Actinopterygii) obsahující celkem 3640 genů.

3 Výsledky

3.1 Sekvenování a úprava získaných dat

Ze dvou opakovaných odběrů krve ze zvoleného exempláře cichlidy *C. semifasciata* jsem extrahovala dostatečné množství DNA pro veškeré sekvenační experimenty. Celkem jsem provedla 16 izolací a výsledná koncentrace DNA se pohybovala mezi 35 a 240 ng/ul. Kvalita extrahované DNA byla odhadnuta prostřednictvím elektroforézy na agarózovém gelu. U šesti vzorků jsem zaznamenala významnou degradaci (fragmentaci) DNA, a proto jsem je vyřadila z dalšího zpracování. Zbývajících 10 vzorků vykazovalo vysokou kvalitu (kompaktní úzký proužek na gelu) a hrubý odhad průměrné délky získaných fragmentů DNA se u nich pohyboval mezi 20 a 40 kb. Tyto izoláty jsem použila pro přípravu celkem 12 sekvenačních knihoven s využitím ligačních kitů Nanopore SQK-LSK109 a SQK-LSK110.

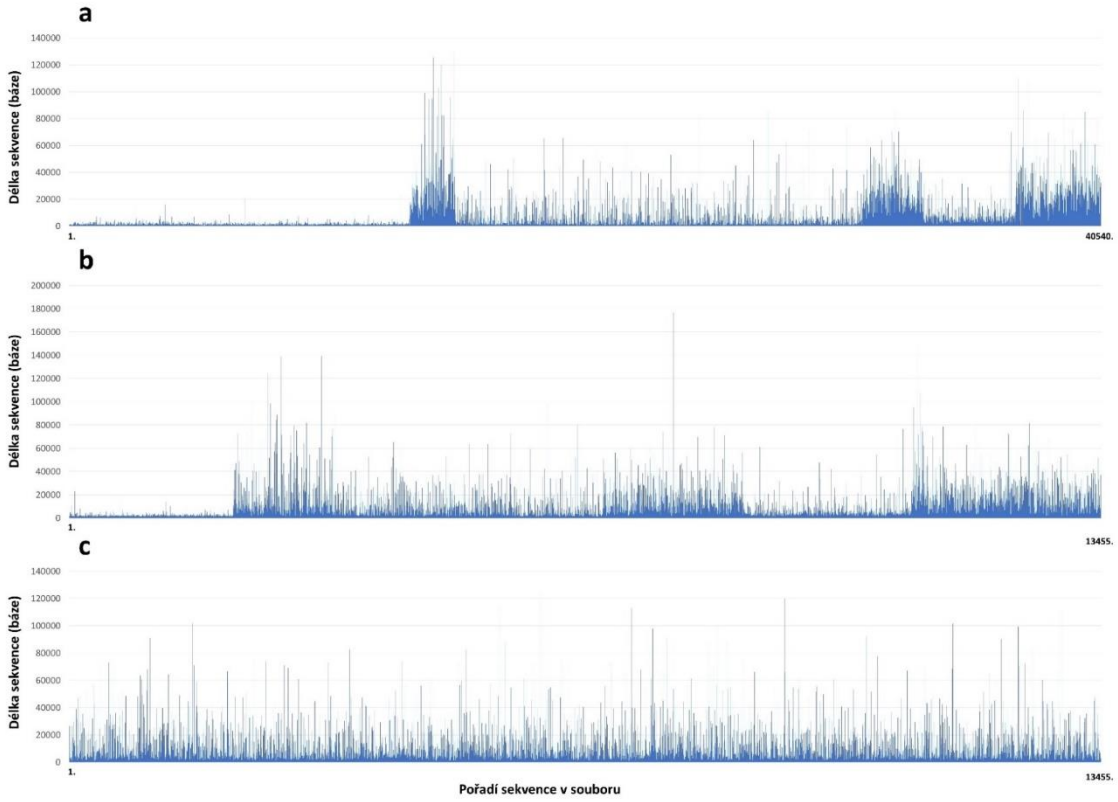
Celkem jsem provedla 16 sekvenačních experimentů (sekvenačních běhů) na přístroji Nanopore MinION; jednotlivé experimenty se lišily hlavně použitou knihovnou, typem sekvenační buňky a její kapacitou (některé byly použity opakovaně) a délkou sekvenačního běhu (tab. 6). Tabelované údaje (N50, počet získaných sekvencí a bází) jsou založené na odhadu získaném tzv. rychlým basecallingem; ovládací program sekvenátoru MinKNOW ve spojení s implementovaným basecallerem Guppy totiž nabízí možnost basecallingu v rychlém módu (model Fast) online již v průběhu sekvenace.

Tabulka 6: Přehled výsledků sekvenačních běhů

Sekv. běh	ID	Celkový výtěžek [Gb]	Počet sekvencí	N50 [b]
1	Pilot00	14,53	12 050 000	1090
2	Adel01	3,53	526 860	18400
3	Adel02	11,27	1 680 000	19600
4	Adel03	21,18	14 580 000	6600
5	Adel04	5,88	4 390 000	3660
6	Adel05	4,23	588 370	15840
7	Adel06	1,71	248 350	14500
8	Adel07	9,57	1 400 000	16930
9	Adel08	3,13	503 430	14370
10	Adel09	9,36	4 040 000	4380
11	Adel10	3,12	492 870	28190
12	Adel11	7,05	1 100 000	17010
13	Adel12	5,67	732 790	19250
14	Adel13	0,26	28 560	20800
15	Adel14	0,63	57 950	21970
16	Adel15	1,75	194 300	16190

Tyto předběžné výsledky slouží hlavně pro analytické účely – díky jejich vyhodnocení mohou být online upraveny parametry sekvenačního běhu nebo experiment přerušen. Celkem bylo na základě těchto dat osekvenováno 42.6 mil. fragmentů DNA obsahujících 102.9 mld. bází.

Finální basecalling byl prováděn v programu Guppy s modelem „High Accuracy“. V této fázi jsme datový soubor taktéž rozšířili o sekvenační data získaná ze stejného jedince a na stejném typu sekvenátoru u komerční firmy SeqMe v roce 2019 (jeden sekvenační běh, ca. 1.1 mil. sekvencí obsahujících 10.2 mld. bází). Po basecallingu a odstranění adaptérů obsahoval datový soubor celkem 40.5 mil. sekvencí a 97.7 mld. bází (N50 = 12.1 kb), po odfiltrování sekvencí kratších než 1000 bází pak 13.5 mil. sekvencí a 86.5 mld. bází (N50 = 14.9 kb). Před vytvořením deseti finálních datasetů (pro hloubku čtení 10x, 20x, ... 100x) bylo provedeno znáhodnění pořadí sekvencí a jeho výsledek zkontrolován graficky v programu MS Excel (vzorkována každá 1000. sekvence; obr. 6).



Obrázek 6: Délka sekvencí v datových souborech, zobrazena každá tisící sekvence. Případ a) představuje dataset po provedení basecallingu a po odstranění adaptérů, b) tentýž dataset po odfiltrování sekvencí kratších než tisíc bází; c) kompletní dataset se znáhodněným pořadím sekvencí.

3.2 Sestavování genomů asemblerem Flye a jejich zhodnocení

Pro každý z deseti datasetů jsem provedla celkem třikrát sestavení genomu v asembleru Flye včetně jednoho následného přečištění (polishing) získané sestavy vestavěným programem Minimap (Li 2018). V závislosti na objemu vstupních dat a přiděleném výpočetním stroji v akademické síti Metacentrum trvalo jedno sestavení genomu jeden až čtyři dny.

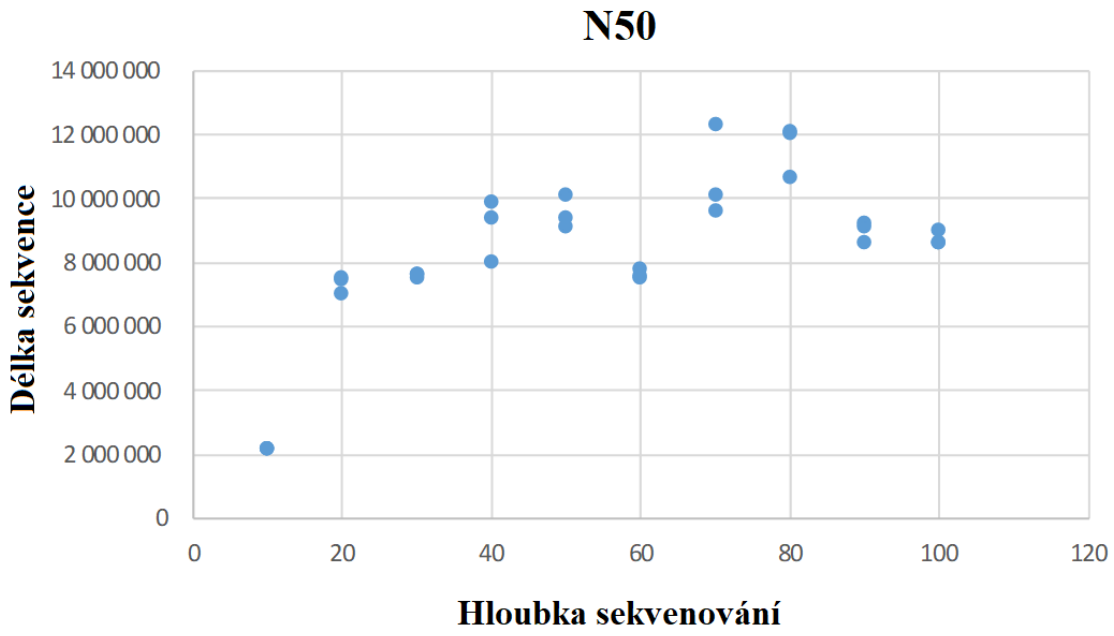
Získaná asembly obsahovala 1302 až 1841 sestavených sekvencí (scaffoldů a contigů) s maximální délkou 10.2 až 32.1 Mb, celková velikost genomů se pohybovala v rozmezí 797.0 až 806.7 Mb a délka sekvencí vyjádřená metrikou N50 byla 2.1 až 12.3 Mb (tab. 7).

Kompletnost genomů posouzená prostřednictvím vyhledávání 3640 ortologních genů z databáze BUSCO pro třídu Actinopterygii se pohybovala v rozmezí 91,8 až 96.7 % (tab. 7).

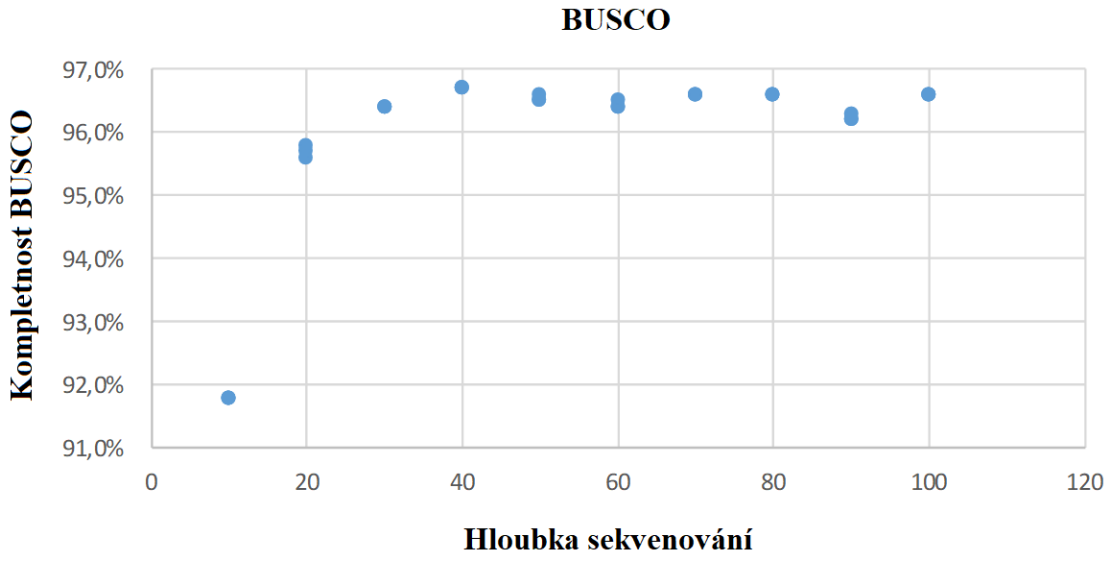
Porovnáme-li výsledná asembly v závislosti na objemu vstupních dat, zjistíme že v rozmezí hloubky čtení 10x až 40x technická kvalita sestav reprezentovaná metrikou N50 významně roste (prodlužuje se průměrná i maximální délka scaffoldů), v rozmezí 50x až 100x už ale zjevně došlo k saturaci dostatečně velkým pokrytím a kvalita genomu kolísá (někdy i velmi významně), pravděpodobně vlivem náhodných faktorů (obr. 7). Rovněž jsem zaznamenala kolísání kvality mezi jednotlivými asembly pro stejný dataset (nejvýznamněji v případě pokrytí 70x, 40x, 80x a 50x; obr. 7). Podobnou závislost vykazuje i kompletnost genomu BUSCO, která nejprve strmě roste až po pokrytí 40x a následně slabě kolísá. V tomto případě jsou ale rozdíly ve stanovené kompletnosti pro opakovaná asembly téhož datasetu minimální (obr. 8).

Tabulka 7: Výsledná asembly pro jednotlivé datasety

Sestava genomu	Coverage	N50 [b]	Kompletnost	Počet sekvencí	Celková délka [b]	Průměrná délka [b]	Max. délka [b]	Počet gapů
flye_cov10_rep1.fas	10	2 154 245	91,8%	1 814	797 034 637	439 380	10 183 263	4 100
flye_cov10_rep2.fas	10	2 146 191	91,8%	1 822	797 243 396	437 565	10 183 230	4 100
flye_cov10_rep3.fas	10	2 154 760	91,8%	1 841	797 239 137	433 047	10 186 047	4 100
flye_cov20_rep1.fas	20	7 471 963	95,8%	1 543	801 224 742	519 264	27 185 656	4 200
flye_cov20_rep2.fas	20	7 014 117	95,6%	1 552	801 072 675	516 155	27 693 467	4 100
flye_cov20_rep3.fas	20	7 446 699	95,7%	1 540	801 518 424	520 467	25 968 678	4 100
flye_cov30_rep1.fas	30	7 601 276	96,4%	1 560	804 001 773	515 386	27 709 516	4 300
flye_cov30_rep2.fas	30	7 599 554	96,4%	1 552	803 590 768	517 778	27 708 464	4 900
flye_cov30_rep3.fas	30	7 488 884	96,4%	1 575	803 673 925	510 269	27 683 906	4 500
flye_cov40_rep1.fas	40	7 966 696	96,7%	1 607	803 668 395	500 105	31 371 307	3 300
flye_cov40_rep2.fas	40	9 839 883	96,7%	1 657	803 626 013	484 989	31 394 988	3 400
flye_cov40_rep3.fas	40	9 354 322	96,7%	1 597	803 086 111	502 872	31 395 638	3 700
flye_cov50_rep1.fas	50	9 068 141	96,5%	1 529	805 484 092	526 805	24 053 513	3 300
flye_cov50_rep2.fas	50	10 084 633	96,5%	1 547	805 268 129	520 535	24 053 527	3 400
flye_cov50_rep3.fas	50	9 363 819	96,6%	1 542	804 436 600	521 684	26 843 367	3 100
flye_cov60_rep1.fas	60	7 752 198	96,4%	1 343	804 249 057	598 845	25 616 789	3 300
flye_cov60_rep2.fas	60	7 541 353	96,4%	1 365	804 838 023	589 625	26 785 982	3 400
flye_cov60_rep3.fas	60	7 496 068	96,5%	1 344	804 823 645	598 827	25 628 439	3 300
flye_cov70_rep1.fas	70	12 306 868	96,6%	1 375	805 877 042	586 092	29 121 771	3 200
flye_cov70_rep2.fas	70	9 562 327	96,6%	1 391	805 961 543	579 412	29 111 381	3 200
flye_cov70_rep3.fas	70	10 084 514	96,6%	1 355	805 867 489	594 736	29 092 656	3 400
flye_cov80_rep1.fas	80	12 069 785	96,6%	1 303	806 691 148	619 103	32 105 741	2 900
flye_cov80_rep2.fas	80	12 006 414	96,6%	1 302	806 480 162	619 416	28 349 418	3 000
flye_cov80_rep3.fas	80	10 655 181	96,6%	1 325	805 988 795	608 293	26 682 950	2 800
flye_cov90_rep1.fas	90	9 194 033	96,2%	1 381	806 078 189	583 692	31 452 975	3 000
flye_cov90_rep2.fas	90	8 616 285	96,2%	1 396	805 490 074	576 999	26 685 417	3 200
flye_cov90_rep3.fas	90	9 087 745	96,3%	1 388	806 288 880	580 900	26 980 723	3 200
flye_cov100_rep1.fas	100	8 980 240	96,6%	1 360	805 241 104	592 089	28 035 310	3 000
flye_cov100_rep2.fas	100	8 576 802	96,6%	1 352	806 272 141	596 355	28 046 705	2 700
flye_cov100_rep3.fas	100	8 576 890	96,6%	1 367	806 421 366	589 921	27 569 165	2 700



Obrázek 7: Závislost délky sekvencí na hloubce sekvenování



Obrázek 8: Závislost kompletnosti BUSCO na hloubce sekvenování

4 Diskuse

Bouřlivý vývoj v oblasti sekvenačních technologií, jehož jsme v poslední době svědky umožňuje biologům řešit nové a stále složitější úlohy, navíc s vynaložením stále nižších nákladů. Zatímco ještě na začátku tohoto století bylo osekvenování neznámého živočišného nebo rostlinného genomu většinou předmětem mnohaletého úsilí několika vědeckých týmů a podmíněné shromážděním astronomických finančních prostředků (Anon. 2004), s nástupem nových platforem masivně paralelního sekvenování (MPS) a bioinformatických technologií se vytvoření základního asembly středně velkého genomu stalo v podstatě rutinní záležitostí. V dnešní době tak mohou i menší vědecké týmy vytvořit, zpracovat a využít celogenomová data pro řešení svých výzkumných otázek. Zatímco druhá generace MPS, charakteristická produkováním krátkých sekvencí DNA (max. několik set bází), umožňovala v nemodifikované formě „pouze“ rychle a levně resekvenovat jedince druhů s již přečteným genomem, třetí generace sekvenátorů významně posouvá obor celogenomového sekvenování do dříve neprobádaných oblastí. Sekvenátory platforem Oxford Nanopore a Pacific Biosciences umožňující souvislé čtení sekvencí o délce běžně statisíců a v některých případech dokonce až několika miliónů bází, a to konkrétně 2 272 580 bází (https://nanoporetech.com/). nabízejí nástroj pro „rozluštění“ těch nejkomplexnějších oblastí genomů jako jsou např. úseky repetetivní DNA (Collins et al. 2003; Anon. 2004).

V případě sekvenování referenčního genomu cichlidy jihoamerického rodu *Crenicichla* nezbytného pro náš výzkumný projekt zabývající se studiem mechanismů paralelní diverzifikace dvou druhových hejn tohoto rodu z povodí Río de la Plata (Piálek et al. 2012; Burress et al. 2022) se tak sekvenování na platformě Nanopore přirozeně stalo naší první volbou. Cena nejmenšího ze systémů této platformy, přístroje MinION, je ve srovnání s ostatními existujícími platformami až neuvěřitelně nízká (ca. 1000 USD, v případě nákupu většího množství sekvenačních kitů i méně, https://nanoporetech.com/) a jeho pořízení je tak možné i z rozpočtu těch nejmenších projektů bez nutnosti shánění a schvalování investičních prostředků. Jako nejmenší sekvenátor platformy nabízí ale přístroj i nejmenší sekvenační kapacitu, v současnosti průměrně 10 Gb na jednu sekvenační buňku typu FLO-MIN106/111, což při sekvenování středně velkého genomu (v případě cichlidy ca. 1 Gb) znamená nutnost provést více než jeden sekvenační běh (https://nanoporetech.com/). Do jaké hloubky je ale potřeba sekvenovat, abychom mohli ze získaných sekvencí sestavit kvalitní genom? Obecně platí, že kvalita výsledného asembly roste s množstvím osekvenovaných bází a větší množství sekvencí znamená i šanci detekovat větší množství hodnověrných oblastí jejich

překryvu, což vede k prodloužení výsledných scaffoldů a contigů. Je potřeba ale vzít v potaz i nezanedbatelné finanční náklady, především cenu spotřebního materiálu pro sekvenování (sekvenační buňky, kity pro izolaci HMW DNA, kity pro úpravu DNA, kity pro přípravu knihovny a doprovodné chemikálie a plasty), která činí ca. 1200 až 1500 USD na jeden experiment. Rovněž stále ještě podceňované náklady na bioinformatické zpracování a uložení dat s větším počtem experimentů výrazně stoupají; navíc výpočetní čas většiny analýz roste s objemem vstupních dat strměji než lineárně. Počet sekvenačních běhů je tak při sekvenování genomu *de novo* zcela jistě potřeba optimalizovat. Proto jsem si stanovila jako další cíl mé bakářské práce, mimo vytvoření kvalitního asembly referenčního genomu, zjištění závislosti kvality získaného asembly na objemu vstupních sekvenačních dat (Ouzounis and Valencia 2003; Wang et al. 2021; Wee et al. 2019; Pereira et al. 2020; Wang et al. 2020).

Ze všech sekvenačních experimentů provedených v naší laboratoři (plus jednoho sekvenování téhož vzorku u komerční firmy) jsem po základní úpravě a filtraci dat získala soubor sekvencí s maximální délkou 387 kb a N50 ca. 14.9 kb. Celkový počet osekvenovaných bází činil téměř 87 Gb a představoval tak přibližně 100x velikost analyzovaného genomu. Výsledné asembly sestavené v programu Flye vykazovalo po technické stránce velmi uspokojivé parametry, nejzdařilejší sestava se skládala z 1375 scaffoldů (>1000 bází), nejdelší scaffold obsahoval 29.1 Mb a N50=12.3 Mb (Tab.). Vezmeme-li do úvahy, že počet chromozómů je u studovaného druhu *C. semifasciata* 46 (Feldberg a Bertollo 1985), tedy 24 párů, a při sestavování genomu dochází ke tvorbě haploidního konsensu, tak při odhadované velikosti genomu 850 Mb je hypotetická průměrná délka jednoho chromozómu ca. 35.4 Mb. Můžeme tedy říci, že nejdelší scaffoldy našeho asembly velikostně odpovídají významným částem celých chromozómů.

Při sestavování genomu jsme se snažili v co největší možné míře zabránit arteficiálnímu spojování nesouvisejících úseků genomu (konzervativní volbou parametrů příslušných programů, striktním odstraňováním sekvencí podobných adaptérům atd.). Přesto je potřeba technické vyhodnocení jednotlivých asembly z hlediska délkových metrik brát s určitou rezervou, neboť se nejedná o kontrolu správnosti jejich sestavení. Výjimku tvoří mitochondriální genom, jehož sekvenci lze i u neznámého organismu poměrně snadno zkontrolovat (u našich sestav s pozitivním výsledkem; L. Piálek pers. comm.), ale vzhledem k jeho nepatrné délce ca. 17 kb nelze z úspěšného sestavení vyvozovat žádné závěry. Vhodnější představu o kvalitě výsledné sestavy z biologického hlediska proto nabízí nástroj BUSCO analyzující přítomnost ortologních genů univerzálních pro určitou linii organismů

(Seppey et al. 2019; Simão et al. 2015). V našem případě databáze pro třídu paprskoploutvých ryb obsahovala 3640 ortologních genů a zjištěná kompletnost sestavených genomů dosahovala 96.7 % (tab. 7). Tyto hodnoty jsou překvapivě vysoké, neboť naše sestavy vznikly pouze na základě dlouhých sekvencí z platformy Nanopore vyznačující se vysokou chybovostí (navíc hodnota 100 % je pravděpodobně nedosažitelná vzhledem k možné ztrátě/převzetí funkce jiným genem v evoluci příslušného organismu). Pro srovnání jsem analyzovala kompletnost rovněž u publikovaného a anotovaného genomu africké cichlidy *Oreochromis niloticus* (https://www.ncbi.nlm.nih.gov/genbank/; txid8128 [Organism:exp]), který je bezesporu jedním z velmi kvalitních zveřejněných genomů nemodelových organismů a výsledná kompletnost byla jen nepatrně vyšší, 98.6 %.

Zajímavé zjištění přineslo zkoumání vlivu objemu vstupních sekvenačních dat na kvalitu sestaveného genomu. Jak technická kvalita, tak kompletnost genomu nejprve strmě rostla až do velikosti datasetu odpovídajícímu hloubce čtení ca 40x (tab. 7), při dalším zvětšování datasetu nad 50x ale docházelo ke stagnaci a kolísání sledovaných parametrů (v případě N50 i velmi významnému) mezi genomy sestavenými z jednotlivých datasetů (obr. 7 a obr. 8). Z praktického hlediska se tak jeví užitečné sekvenovat kvalitativně podobné izoláty DNA na platformě Nanopore pouze do hloubky ca. 60x. K dalšímu zlepšení kvality výsledného asembly vedou patrně jiné cesty, např. izolace významně delších fragmentů DNA s použitím speciálních technik nebo tzv. hybridní asemblování, při kterém jsou kombinována data ze dvou a více různých sekvenačních platforem (Tanaka et al. 2020; Wallberg et al. 2019; Ma et al. 2019). V případě sekvencí Nanopore se jako vhodné ke kombinování jeví např. krátké a přesné (a relativně velmi levné) sekvence z platformy druhé generace Illumina, což dokazují i předběžné výsledky mojí týmové kolegyně D. Kotalové.

5 Závěr

V předkládané práci bylo nutné k vyřešení evoluční otázky týkající se dvou nepříbuzných druhových hejn rodu *Crenicichla* vytvoření referenčního genomu, který se podařilo sestavit. Vyhodnotila jsem také vliv objemu vstupních sekvencí na kvalitu vytvářených sestav genomů a navrhla hloubku sekvenace. Ve své magisterské práci bych se chtěla zaměřit na sekvenaci transkriptomu stejného jedince s cílem zdokonalit a anotovat sestavený genom.

Seznam citované literatury

Anon., 2004. Finishing the euchromatic sequence of the human genome. *Nature* [online]. **431**(7011), 931–945. ISSN 0028-0836. Available at: doi:10.1038/nature03001

BAYLEY, Hagan, 2014. Nanopore Sequencing: From Imagination to Reality [online]. Available at: doi:10.1373/clinchem.2014.223016

BERRA, Tim M., 2001. *Freshwater fish distribution*. B.m.: Academic Press. ISBN 9780120931569.

BURRESS, Edward D, Lubomír PIÁLEK, Jorge CASCIOTTA, Adriana ALMIRÓN and Oldřich ŘÍČAN, 2022. Rapid Parallel Morphological and Mechanical Diversification of South American Pike Cichlids (Crenicichla). *Systematic Biology* [online]. [accessed. 2022-04-01]. ISSN 1063-5157. Available at: doi:10.1093/SYSBIO/SYAC018

BURRESS, Edward D., Lubomír PIÁLEK, Jorge R. CASCIOTTA, Adriana ALMIRÓN, Milton TAN, Jonathan W. ARMBRUSTER and Oldřich ŘÍČAN, 2018. Island- and lake-like parallel adaptive radiations replicated in rivers. *Proceedings of the Royal Society B: Biological Sciences* [online]. **285**(1870) [accessed. 2022-04-12]. ISSN 14712954. Available at: doi:10.1098/RSPB.2017.1762

CASTRO-WALLACE, Sarah L., Charles Y. CHIU, Kristen K. JOHN, Sarah E. STAHL, Kathleen H. RUBINS, Alexa B.R. MCINTYRE, Jason P. DWORKIN, Mark L. LUPISELLA, David J. SMITH, Douglas J. BOTKIN, Timothy A. STEPHENSON, Sissel JUUL, Daniel J. TURNER, Fernando IZQUIERDO, Scot FEDERMAN, Doug STRYKE, Sneha SOMASEKAR, Noah ALEXANDER, Guixia YU, Christopher E. MASON and Aaron S. BURTON, 2017. Nanopore DNA Sequencing and Genome Assembly on the International Space Station. *Scientific Reports* [online]. **7**(1) [accessed. 2022-03-30]. ISSN 20452322. Available at: doi:10.1038/S41598-017-18364-0

COLLINS, Francis S., Michael MORGAN and Aristides PATRINOS, 2003. The Human Genome Project: lessons from large-scale biology. *Science (New York, N.Y.)* [online]. **300**(5617), 286–290 [accessed. 2022-04-13]. ISSN 1095-9203. Available at: doi:10.1126/SCIENCE.1084564

DE LANNOY, Carlos, Dick DE RIDDER and Judith RISSE, 2017. The long reads ahead: De novo genome assembly using the MinION. *F1000Research* [online]. **6**

[accessed. 2022-04-01]. ISSN 1759796X. Available at: doi:10.12688/F1000RESEARCH.12012.2

DE LUCENA, Carlos A.S., 2007. Two new species of the genus Crenicichla Heckel, 1840 from the upper rio Uruguay drainage (Perciformes: Cichlidae). *Neotropical Ichthyology* [online]. **5**(4). ISSN 16796225. Available at: doi:10.1590/s1679-62252007000400004

DEAMER, David, Mark AKESON, Daniel BRANTON and Nat Biotechnol AUTHOR, 2016. Three decades of nanopore sequencing HHS Public Access Author manuscript. *Nat Biotechnol* [online]. **34**(5), 518–524. Available at: doi:10.1038/nbt.3423

ESCHMEYER, W.N., R., FRICKE and R. VAN DER LAAN, 2020. *Eschmeyer's Catalog of Fishes.* [online] [accessed. 2022-02-01]. Available at: https://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp

FENG, Yanxiao, Yuechuan ZHANG, Cuifeng YING, Deqiang WANG and Chunlei DU, 2015. Nanopore-based Fourth-generation DNA Sequencing Technology. *Genomics, Proteomics & Bioinformatics* [online]. **13**(1), 4–16 [accessed. 2022-02-17]. ISSN 1672-0229. Available at: doi:10.1016/J.GPB.2015.01.009

HEATHER, James M. and Benjamin CHAIN, 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics* [online]. **107**(1), 1–8 [accessed. 2022-02-11]. ISSN 0888-7543. Available at: doi:10.1016/J.YGENO.2015.11.003

HOWORKA, Stefan, Stephen CHELEY and Hagan BAYLEY, 2001. Sequence-specific detection of individual DNA strands using engineered nanopores. *Nature biotechnology* [online]. **19**(7), 636–639 [accessed. 2022-02-17]. ISSN 1087-0156. Available at: doi:10.1038/90236

HU, Taishan, Nilesh CHITNIS, Dimitri MONOS and Anh DINH, 2021. Next-generation sequencing technologies: An overview. *Human Immunology* [online]. **82**(11), 801–811 [accessed. 2022-02-13]. ISSN 0198-8859. Available at: doi:10.1016/J.HUMIMM.2021.02.012

CHARGAFF, E., 1950. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia* [online] [accessed. 2022-02-17]. Available at: doi:10.1007/bf02173653

IP, Camilla L C, Matthew LOOSE, John R TYSON, Mariateresa DE CESARE, Bonnie L BROWN, Miten JAIN, Richard M LEGGETT, David A ECCLES, Vadim ZALUNIN, John M URBAN, Paolo PIAZZA, Rory J BOWDEN, Benedict PATEN, Solomon MWAIGWISYA, Elizabeth M BATTY, Jared T SIMPSON, Terrance P SNUTCH, Ewan BIRNEY, David BUCK, Sara GOODWIN, Hans J JANSEN, Justin O'GRADY and Hugh E OLSEN, 2015. Open Peer Review MinION Analysis and Reference Consortium: Phase 1 data release and analysis [version 1; referees: 2 approved] [online]. Available at: doi:10.12688/f1000research.7201.1

JAIN, Miten, Hugh E OLSEN, Benedict PATEN and Mark AKESON, 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community [online]. Available at: doi:10.1186/s13059-016-1103-0

JOHNSON, Sarah S, Elena ZAIKOVA, David S GOERLITZ, Yu BAI and Scott W TIGHE, 2017. Real-Time DNA Sequencing in the Antarctic Dry Valleys Using the Oxford Nanopore Sequencer [online]. Available at: doi:10.7171/jbt.17-2801-009

KARAMITROS, Timokratis and Gkikas MAGIORKINIS, 2018. Multiplexed Targeted Sequencing for Oxford Nanopore MinION: A Detailed Library Preparation Procedure. *Methods in molecular biology (Clifton, N.J.)* [online]. **1712**, 43–51 [accessed. 2022-04-13]. ISSN 1940-6029. Available at: doi:10.1007/978-1-4939-7514-3_4

KASIANOWICZ, John J, Eric BRANDIN, Daniel BRANTON and David W DEAMER, 1996. *Characterization of individual polynucleotide molecules using a membrane channel*.

KOBLMÜLLER, Stephan, R Craig ALBERTSON, Martin J GENNER, Kristina M SEFC and Tetsumi TAKAHASHI, 2011. Access to. *Research International Journal of Evolutionary Biology* [online]. **2011** [accessed. 2022-02-06]. Available at: doi:10.4061/2011/847969

KOCHER, Thomas D, 2004. ADAPTIVE EVOLUTION AND EXPLOSIVE SPECIATION: THE CICHLID FISH MODEL [online]. [accessed. 2022-02-01]. Available at: doi:10.1038/nrg1316

KOLMOGOROV, Mikhail, Derek M. BICKHART, Bahar BEHSAZ, Alexey GUREVICH, Mikhail RAYKO, Sung Bong SHIN, Kristen KUHN, Jeffrey YUAN, Evgeny POLEVIKOV, Timothy P.L. SMITH and Pavel A. PEVZNER, 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature methods* [online]. **17**(11),

1103–1110 [accessed. 2022-04-10]. ISSN 1548-7105. Available at: doi:10.1038/S41592-020-00971-X

KULLANDER, Sven O, 1999. *Fish species-how and why*.

KUMAR, Kishore R, Mark J COWLEY, Ryan L DAVIS, David John RABBOLINI and Maha OTHMAN, 2019. Next-Generation Sequencing and Emerging Technologies. *Semin Thromb Hemost* [online]. **45**, 661–673. ISSN 0094-6176. Available at: doi:10.1055/s-0039-1688446

LI, Heng, 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* [online]. **34**(18), 3094–3100. ISSN 1367-4803. Available at: doi:10.1093/bioinformatics/bty191

LIN, Bo, Jianan HUI and Hongju MAO, 2021. Nanopore technology and its applications in gene sequencing. *Biosensors* [online]. **11**(7) [accessed. 2022-03-21]. ISSN 20796374. Available at: doi:10.3390/BIOS11070214

LU, Hengyun, Francesca GIORDANO and Zemin NING, 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics* [online]. **14**(5), 265–279 [accessed. 2022-02-11]. ISSN 1672-0229. Available at: doi:10.1016/J.GPB.2016.05.004

MA, Zhanshan (Sam), Lianwei LI, Chengxi YE, Minsheng PENG and Ya Ping ZHANG, 2019. Hybrid assembly of ultra-long Nanopore reads augmented with 10x-Genomics contigs: Demonstrated with a human genome. *Genomics* [online]. **111**(6), 1896–1901 [accessed. 2022-04-13]. ISSN 1089-8646. Available at: doi:10.1016/J.YGENO.2018.12.013

MABUCHI, Kohji, Masaki MIYA, Yoichiro AZUMA and Mutsumi NISHIDA, 2007. Independent evolution of the specialized pharyngeal jaw apparatus in cichlid and labrid fishes [online]. Available at: doi:10.1186/1471-2148-7-10

MARDIS, E.R., 2013. Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry* [online] [accessed. 2022-02-13]. Available at: doi:10.1146/annurev-anchem-062012-092628

MAXAM, Allan M and Walter GILBERT, 1977. *A new method for sequencing DNA (DNA chenistry/dimethyl sulfate cleavage/hydrazine/piperidine)*.

MCCOMBIE, W Richard, John D MCPHERSON and Elaine R MARDIS, 2019. Next-Generation Sequencing Technologies [online]. Available at: doi:10.1101/cshperspect.a036798

METZKER, Michael L., 2009. Sequencing technologies — the next generation. *Nature Reviews Genetics* [online]. 31–46 [accessed. 2022-02-11]. Available at: https://sci-hub.se/10.1038/nrg2626

MUNSHI ANJANA, 2012. DNA SEQUENCING-METHODS AND APPLICATIONS Edited by Anjana Munshi [online]. [accessed. 2022-01-31]. Available at: www.intechopen.com

NARAYAN, Edward, Janicke NORDGREEN, Yifei YANG, Alexander REBL, Henrike SEIBEL and Björn BASSMANN, 2021. Blood Will Tell: What Hematological Analyses Can Reveal About Fish Welfare. *Frontiers in Veterinary Science | www.frontiersin.org* [online]. **8**, 616955. Available at: doi:10.3389/fvets.2021.616955

NELSON, J. S., T. C. GRANDE and M. V. H. WILSON, 2016. *Fishes of the World* [online]. ISBN 9781118342336. Available at: doi:10.1002/9781119174844

OUZOUNIS, Christos A. and Alfonso VALENCIA, 2003. Early bioinformatics: the birth of a discipline--a personal view. *Bioinformatics (Oxford, England)* [online]. **19**(17), 2176–2190 [accessed. 2022-04-12]. ISSN 1367-4803. Available at: doi:10.1093/BIOINFORMATICS/BTG309

PEREIRA, Rute, Jorge OLIVEIRA and Mário SOUSA, 2020. Clinical Medicine Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics [online]. Available at: doi:10.3390/jcm9010132

PÉREZ-MIRANDA, Fabian, Omar MEJÍA, Eduardo SOTO-GALERA, Héctor ESPINOSA-PÉREZ, Lubomír PIÁLEK and Oldřich ŘÍČAN, 2018. Phylogeny and species diversity of the genus Herichthys (Teleostei: Cichlidae). *Journal of Zoological Systematics and Evolutionary Research* [online]. **56**(2), 223–247 [accessed. 2022-04-12]. ISSN 1439-0469. Available at: doi:10.1111/JZS.12197

PERMENTER, Jessalyn, Arjun ISHWAR, Angie ROUNSAVALL, Maddie SMITH, Jennifer FASKE, Charles J. SAILEY and Maria P. ALFARO, 2015. Quantitative analysis of genomic DNA degradation in whole blood under various storage conditions for molecular

diagnostic testing. *Molecular and cellular probes* [online]. **29**(6), 449–453 [accessed. 2022-04-13]. ISSN 1096-1194. Available at: doi:10.1016/J.MCP.2015.07.002

PIÁLEK, Lubomír, Edward BURRESS, Klára DRAGOVÁ, Adriana ALMIRÓN, Jorge CASCIOTTA and Oldřich ŘÍČAN, 2019a. 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* [online]. **832**(1), 355–375 [accessed. 2022-04-02]. ISSN 15735117. Available at: doi:10.1007/S10750-018-3733-6/FIGURES/5

PIÁLEK, Lubomír, Jorge CASCIOTTA, Adriana ALMIRÓN and Oldřich ŘÍČAN, 2019b. A new pelagic predatory pike cichlid (Teleostei: Cichlidae: Crenicichla) from the C. mandelburgeri species complex with parallel and reticulate evolution. *Hydrobiologia* [online]. **832**(1), 377–395 [accessed. 2022-04-02]. ISSN 15735117. Available at: doi:10.1007/S10750-018-3754-1/FIGURES/12

PIÁLEK, Lubomír, Klára DRAGOVÁ, Jorge CASCIOTTA, Adriana ALMIRÓN and Oldřich ŘÍČAN, 2015. Description of two new species of *Crenicichla* (Teleostei: Cichlidae) from the Lower Iguazú River with a taxonomic reappraisal of *C. iguassuensis*, *C. tesay* and *C. yaha*. *Historia Natural, Tercera Serie*. **5**(2).

PIÁLEK, Lubomír, Oldřich ŘÍČAN, Jorge CASCIOTTA and Adriana ALMIRÓN, 2010. Crenicichla hu, a new species of cichlid fish (Teleostei: Cichlidae) from the Paraná basin in Misiones, Argentina. *Zootaxa* [online]. (2537), 33–46 [accessed. 2022-02-04]. ISSN 11755334. Available at: doi:10.11646/ZOOTAXA.2537.1.3

PIÁLEK, Lubomír, Oldřich ŘÍČAN, Jorge CASCIOTTA, Adriana ALMIRÓN and Jan ZRZAVÝ, 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* [online]. **62**(1), 46–61 [accessed. 2022-04-02]. ISSN 10557903. Available at: doi:10.1016/J.YMPEV.2011.09.006

RHOADS, Anthony and Kin Fai AU, 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics* [online]. **13**(5), 278–289 [accessed. 2022-04-13]. ISSN 22103244. Available at: doi:10.1016/J.GPB.2015.08.002

RICE, Aaron N. and Phillip S. LOBEL, 2003. The pharyngeal jaw apparatus of the Cichlidae and Pomacentridae: function in feeding and sound production. *Reviews in Fish*

Biology and Fisheries [online]. **13**(4), 433–444. ISSN 0960-3166. Available at: doi:10.1007/s11160-004-8794-0

ŘÍČAN, Oldřich, Štěpánka ŘÍČANOVÁ, Adriana ALMIRÓN, Jorge CASCIOTTA and Lubomír PIÁLEK, 2021. Parallel evolution of a new sympatric species pair of Crenicichla (Teleostei: Cichlidae) from Misiones, Argentina with a review of biogeography and mitonuclear discordance in the C. mandelburgeri species complex. *Journal of Zoological Systematics and Evolutionary Research* [online]. **59**(8), 2134–2149 [accessed. 2022-01-30]. ISSN 14390469. Available at: doi:10.1111/JZS.12566

SALZBURGER, Walter, Bert van BOCXLAER and Andrew S. COHEN, 2014. Ecology and evolution of the African great lakes and their faunas. *Annual Review of Ecology, Evolution, and Systematics* [online]. **45**, 519–545 [accessed. 2022-02-06]. ISSN 15452069. Available at: doi:10.1146/ANNUREV-ECOLSYS-120213-091804

SALZBURGER, Walter and Axel MEYER, 2004. The species flocks of East African cichlid fishes: Recent advances in molecular phylogenetics and population genetics. *Naturwissenschaften* [online]. **91**(6), 277–290 [accessed. 2022-02-06]. ISSN 00281042. Available at: doi:10.1007/S00114-004-0528-6

SANGER, F, S NICKLEN and A R COULSON, 1977. *DNA sequencing with chain-terminating inhibitors (DNA polymerase/nucleotide sequences/bacteriophage 4X174)*.

SENOL CALI, Damla, Jeremie S. KIM, Saugata GHOSE, Can ALKAN and Onur MUTLU, 2018. Nanopore sequencing technology and tools for genome assembly: Computational analysis of the current state, bottlenecks and future directions. *Briefings in Bioinformatics* [online]. **20**(4), 1542–1559 [accessed. 2022-04-06]. ISSN 14774054. Available at: doi:10.1093/bib/bby017

SEPPEY, Mathieu, Mosè MANNI and Evgeny M. ZDOBNOV, 2019. BUSCO: Assessing genome assembly and annotation completeness. In: *Methods in Molecular Biology* [online]. B.m.: Humana Press Inc., p. 227–245. ISSN 19406029. Available at: doi:10.1007/978-1-4939-9173-0_14

SHEN, Wei, Shuai LE, Yan LI and Fuquan HU, 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE* [online]. **11**(10), e0163962. ISSN 1932-6203. Available at: doi:10.1371/journal.pone.0163962

SIMÃO, Felipe A., Robert M. WATERHOUSE, Panagiotis IOANNIDIS, Evgenia v. KRIVENTSEVA and Evgeny M. ZDOBNOV, 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)* [online]. **31**(19), 3210–3212 [accessed. 2022-04-04]. ISSN 1367-4811. Available at: doi:10.1093/BIOINFORMATICS/BTV351

STOSKOPF, M.K., 1993. *Fish medicine*. Philadelphia: W.B. Sannders.

TANAKA, Tsuyoshi, Ryo NISHIJIMA, Shota TERAMOTO, Yuka KITOMI, Takeshi HAYASHI, Yusaku UGA and Taiji KAWAKATSU, 2020. De novo Genome Assembly of the indica Rice Variety IR64 Using Linked-Read Sequencing and Nanopore Sequencing [online]. Available at: doi:10.1534/g3.119.400871

WALLBERG, Andreas, Ignas BUNIKIS, Olga VINNERE PETTERSSON, Mai-Britt MOSBECH, Anna K CHILDERS, Jay D EVANS, Alexander S MIKHEYEV, Hugh M ROBERTSON, Gene E ROBINSON and Matthew T WEBSTER, 2019. BMC Genomics [online]. Available at: doi:10.1186/s12864-019-5642-0

WANG, Yue, Qiuping YANG, Zhimin WANG and Genomics LAB, 2015. The evolution of nanopore sequencing [online]. Available at: doi:10.3389/fgene.2014.00449

WANG, Yunhao, Yue ZHAO and Audrey ✉ BOLLAS, 2021. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology* [online]. **39**, 1348–1365. Available at: doi:10.1038/s41587-021-01108-x

WANG, Ziye, Ying WANG, Jed A FUHRMAN, Fengzhu SUN, Shanfeng Zhu CORRESPONDING and Shanfeng ZHU, 2020. Assessment of metagenomic assemblers based on hybrid reads of real and simulated metagenomic sequences. *Briefings in Bioinformatics* [online]. **21**(3), 777–790. Available at: doi:10.1093/bib/bbz025

WATSON, J. D. and F. H.C. CRICK, 1953. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* [online]. **171**(4356), 737–738 [accessed. 2022-02-11]. ISSN 00280836. Available at: doi:10.1038/171737A0

WEE, Yongkiat, Salma Begum BHYAN, Yining LIU, Jiachun LU, Xiaoyan LI and Min ZHAO, 2019. The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. *Briefings in Functional Genomics* [online]. **18**(1), 1–12 [accessed. 2022-04-06]. ISSN 20412657. Available at: doi:10.1093/bfgp/ely037

WICK, Ryan R., Louise M. JUDD and Kathryn E. HOLT, 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology* [online]. **20**(1), 129. ISSN 1474-760X. Available at: doi:10.1186/s13059-019-1727-y

YANG, Chao, Debajyoti CHOWDHURY, Zhenmiao ZHANG, William K CHEUNG, Aiping LU, Zhaoxiang BIAN and Lu ZHANG, 2021. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and structural biotechnology journal* [online]. **19**, 6301–6314 [accessed. 2022-04-01]. ISSN 2001-0370. Available at: doi:10.1016/j.csbj.2021.11.028