Oponentský posudek na bakalářskou práci Kseniye Bobryshavy s názvem „Příběh hladové housenky: anotace chuťových receptorů předivky *Yponomeuta evonymella*

Bakalářská práce Kseniye Bobryshavy s poutavým názvem se zabývá detekcí, sekvenováním a anotací genů pro cukerné chuťové receptory předivky *Yponomeuta evonymella*. Studie je zasazena do širšího evolučního kontextu funkce a významu těchto receptorů pro adaptaci a speciaci.

Práce je psána česky a na prvním místě bych proto chtěl ocenit úsilí věnované překladu nepřeložitelných odborných výrazů do češtiny, které jsou vcelku vhodně volené a nejsou na úkor srozumitelnosti textu. Možná bych jen sekvenační element u platformy Illumina nenazýval destičkou (termín vhodný spíše pro 454 technologii), ale sekvenační/průtočnou buňkou, komůrkou či celou, ale to je samozřejmě nevýznamný detail. Práce obsahuje vzhledem ke svému rozsahu jen malé množství překlepů a stylistických chyb.

Úvod práce je poměrně rozsáhlý (12 stran včetně cílů). Za velmi zdařilou považuji jeho biologickou část, která poskytne neznalému čtenáři dostatek informací o stavbě, funkci a vývojové dynamice chuťových receptorů a nadchne ho pro další čtení. Druhá a techničtější část tohoto oddílu shrnuje historii sekvenování od objevu DNA až po třetí generaci sekvenátorů a informace v ní podané nejsou z jisté části pro pochopení práce relevantní. Např. odstavec věnovaný předvěkému Maxam-Gilbertovu sekvenování je zcela určitě možné oželet (zvlášť když k němu chybí obvyklý obrázek), na druhou stranu u použité technologie Oxford Nanopore by neškodilo blíže popsat proces průchodu makromolekuly pórem nebo se zmínit o způsobu převodu elektrického signálu na sekvenci bazí, případně vyzdvihnout obvyklé artefakty použitých sekvenačních metod atd. Cíle práce jsou v závěru definovány jasně a srozumitelně.

V metodické části bych uvítal základní schéma nebo diagram celého postupu, který by pomohl čtenáři zorientovat se v celém procesu hned na začátku této kapitoly, která je občas psána poněkud nepřehledně. Některé postupy jsou popsány dostatečně, jiné (např. Chromosome walking a Příprava BAC DNA pro ONT) jsou osvětleny pouze zčásti a další metodické informace k nim se skrývají až v oddíle Výsledky. Poněkud stručně je popsáno zpracování ONT dat, tj. basecalling, asemblování BAC klonů a mapování genů. Naopak velmi oceňuji publikování všech použitých skriptů na GitHubu.

V kapitole Výsledky mě mimo výše zmíněných čistě metodických pasáží zaujalo několik nesrovnalostí, které zmiňuji dále v dotazech. Dobrou zprávou je, že navržený postup vedl (alespoň částečně) k cíli a některé z genů pro cukerné receptory se podařilo zdárně osekvenovat a anotovat. Diskuse se pak zabývá z velké části rozborem problémů vyskytnuvších se během práce a hledáním jejich řešení včetně alternativních postupů, což shledávám velmi užitečným pro další práci na tématu. Zvolený úkol je totiž značně netriviální, např. proto že se při detekci cílových genů spoléhá na transkriptom poměrně nepříbuzného druhu, navíc cesta vedoucí k cíli musí v nějaké formě využívat sekvenování třetí generace jehož chybovost je v současné době větší než odhadovaná variabilita zkoumaných genů.

K práci mám následující připomínky a dotazy:

1) Autorka uvádí, že pro sestavení genomu *Y. evonymella* bylo v databázi Darwin tree of life k dispozici 211.5 k readů (délka 151 bp), pro asemblování ale použila 140 M, 250 M a 350 M readů...

2) Sestavený referenční genom z 10X Genomics dat byl z programu Supernova exportován ve formě částečně diploidního assembly (--style=megabubbles). Tento formát podle mě není vhodný pro následné vyhodnocení kompletnosti v BUSCO, neboť velká část vyhledávaných genů musí zákonitě vyjít jako duplikovaná. Pro tento účel doporučuji exportovat v pseudohaploidním formátu (--style=pseudohap).
3) Autorka reagovala na enormní množství duplikovaných genů poněkud překvapivě následnou deduplikací (v programu Purge-dups), která musí nepochybně zhoršit kvalitu assembly - jak vidno např. z celkového počtu kompletních BUSCO genů v Tab. 6, který po deduplikaci poklesl. Navíc může deduplikace nekontrolovaně pracovat s ortology/paralogy. Je proto otázkou, zda mohl tento krok případně ovlivnit konečný výsledek a zda pro vlastní mapování genů *P. xylostella* není vhodnější ponechat genom v onom částečně diploidním stavu (větší diverzita cílových sekvencí).
4) Pro deduplikaci byly využity rovněž primární 10XG ready zbavené barkódů v programu Trimmomatic. Jak dlouhý úsek byl odstřižen?
5) Soubor ortologních genů BUSCO pro Insecta obsahuje dle autorky 1658 genů (text a Tab. 4), v Tab. 6 je ale redukován na 1367 – jak a proč? Zkratka BUCBO v posledním řádku všech souvisejících tabulek je jen překlep nebo mi její význam uniká?
6) Osekvenovaná BAC DNA byla filtrovaná na kvalitu Q15 a délku 30 kbp – proč tak dramaticky?
7) Proč byl basecalling v Guppy proveden algoritmem High a nikoli Super Accuracy, který by mohl zvýšit přesnost sekvencí o několik procent?

Předkládaná bakalářská práce Kseniye Bobryshavy dle mého názoru jednoznačně splňuje požadavky kladené na daný typ práce na KMB PřF JU. Svým rozsahem je bezesporu nadprůměrná a autorka si v jejím rámci musela osvojit značné množství laboratorních technik i bioinformatických postupů, což je v tomto stupni studia podstatnější než uvedené obsahové a formální nedostatky nebo nekompletní výsledek. Práci proto rád doporučuji k obhajobě a budu hodnotit stupněm výborně.

V Českých Budějovicích 15.5. 2023

Lubomír Piálek