Oponentský posudek disertační práce Mgr. Petra Vojty

na téma

Zpracování dat z vysokokapacitního DNA sekvenování pro studium variability genomu a transkriptomu

Práce má 125 číslovaných stran, dále obsahuje 28 stran autorského manuskriptu, 16-ti stránkový prvoautorský článek v časopise New Biotechnology 33:676-691 (2016) a 14-ti stránkový spoluautorský článek Pospíšilová *et al.*, New Biotechnology 33:692-705 (2016). Časopis má IF=3.733. Dále práce obsahuje CD-ROM médium s 11 elektronickými přílohami tabulek v CSV nebo XLSX formátu. Médium by mělo být řádně popsáno jménem autora a názvem práce. Ve skutečnosti má Petr Vojta 11 publikací v impaktovaných časopisech a do své disertační práce vložil sotva ¼. Celkově mají jeho publikace podle Web Of Science 28 citací když nepočítám autocitace a H-index je 3.0.

Autor na straně 2 uvádí že si nepřeje aby byla jeho práce uložena v projektu theses.cz. Domnívám se že **je věcí studijního oddělení aby posoudilo jeho požadavek**.

Práce má na můj vkus neobvyklé členění. Bez jakéhokoli Úvodu a stanovení cílů začíná ihned na straně 8 jakýmsi Literárním přehledem ("Obecný úvod do problematiky"), který je ovšem směsí přehledu technologií, softwarových nástrojů a datových formátů spolu s názory autora práce. Vzhledem k tomu že publikované výstupy jsou na základě programů DeSeq2 a Cufflinks mohl se jim autor trochu věnovat a diskutovat proč zvolil zrovnat tuto kombinaci pro svoje analýzy. Obdobně se mohl více rozepsat o tom proč zvolil do MOLDIMED pipeline nástroje samtools, GATK UnifiedGenotyper, VarScan2 a proč je později vyměnil za Platypus a následně za GATK HaplotypeCaller (částečně to uvádí v kapitole 3 ale mohl je více charakterizovat v Literárním přehledu).

Kapitola 2 obsahuje seznam hypotéz a cílů. V krátkosti, cílem bylo zlepšit stávající možnosti analýzy a interpretace dat ze sekvenátorů nové generace. Specifické cíle 1-3 jsou prakticky realizovány v projektu MOLDIMED a 4. cíl je naplněn publikacemi o Diamond-Blackfanově anémii a transkriptomech ječmene. Bylo by vhodné zdůraznit že práce na vývoji softwaru pro projekt MOLDIMED je stěžejní ačkoli není dosud uspokojivě publikována.

Kapitoly 3, 4 a 5 jsou psány formou mini-článku. V případě kapitol 4 a 5 obsahují každá ne moc zestručněný výtah z manuskriptu a dvou publikací které autor zařadil do své disertační práce. Mají tedy vlastní úvod, metodiku, výsledky, diskusi a závěr. Bohužel zde chybí odkazy na konkrétní místa uvnitř samotného publikovaného článku a tak se špatně ověřují jednotlivá tvrzení v kontextu skutečných výsledků.

Autor píše text v množném čísle nebo trpném rodě, což prakticky znamená že není zřejmé co je jeho vlastní výsledek a co je výsledek jeho spolupracovníků. Např. na straně 19 na šestém řádku od konce je napsáno: "…, domníváme se, že vzhledem ke klesající ceně vyšetření exomu, ...". Množné číslo se opakuje například na straně 47 na konci druhého odstavce: "...Pro ilustraci … uvádíme příklad pro BYSL gen ...", na straně 48 a také 49 dole "...představujeme ...", "prezentujeme" na straně 67. Dalších příkladů by bylo ještě mnoho.

Navíc ani z jiných míst uvnitř samotného textu disertační práce nepoužívá autor odkaz na jiná místa ve své práci. Například v kapitolkách Diskuse chybí odkazy zpět na Výsledky, Obrázky, Tabulky téhož mini-článku. Konkrétně na straně 67 a 80-81.

Po přečtení anglického manuskriptu článku o Diamond-Blackfan anemii musím zkonstatovat že celé pasáže jsou stejné s českým textem mini-článku v hlavním textu disertační práce, občas v jiném pořadí nebo pod jiným názvem kapitolky/odstavce. Autora Petra Vojtu nicméně vidím v seznamu autorů kteří přispěli ke psaní pracovní verze textu manuskriptu a musím uznat že místy je text v mini-článkové verzi rozšířen (navíc jsou některé obrázky/tabulky). Nicméně mě netěší že jsem

místy shodný text četl dvakrát a především, autor měl místo využít k prezentaci výsledků a úvah které se do tištěné publikace nevešly. Obdobně je tomu v části věnované analýze ječmene a dvou publikací na toto téma. Vpodstatě to i vysvětluje proč autor používá množné číslo a nehovoří za sebe a o svých výsledcích. Text věnovaný vyvinutému MOLDIMED softwaru jako jediný nemá dvojí podobu protože ještě není publikován.

Další formální problém je, že v kapitole 6 na ¾ stránky následuje jen shrnutí předešlých kapitol 3, 4, a 5 bez nějakých odkazů/citací, v kapitole 7 je anglický překlad téhož textu a potom již následuje jen seznam zkratek, použité literatury , přílohy, životopis, přiložený manuskript a dvě publikace.
Postrádám text který by témata nějak více spojil dohromady, zdůvodnil proč zrovna tyto práce autor předkládá ve své disertační práci (na úkor jiných) a především, co do nich autor přinesl svým bioinformatickým přístupem. Chybějící text mohl být v dosud neexiistujícím Úvodu, nebo spolu s definicemi cílů a hypotéz, nebo v nějaké další všeobjímající kapitole Diskuse nebo Závěr, záleželo by na konkrétním obsahu a rozsahu textu.

Ačkoli je psaná práce zdánlivě bioinformatická, většina psaného textu se věnuje spíše molekulární biologii a interpretaci molekulárních výsledků než zdrojům dat, metodám zpracování, případným kontaminacím, separacím vzorků od sebe ze směsných dat a jen minimálně diskusi hrubých bioinformatických výstupů. Je to dáno tím že jsou to vpodstatě přeložené nebo přepsané původní odborné články, ve kterých bioinformatika byla jen pomocným přístupem. Výjimkou je více technický výstup ve formě softwarového nástroje MOLDIMED v kapitole 3.

Kapitoly 1.1.1 až 1.1.6 věnované různých komerčním sekvenačním platformám mají málo citací. Absolutně chybí například v místě kde autor píše o tom že se používají "pro resekvenování celého lidského genomu několika desítek jedinců v jednom běhu", na konci druhého odstavce na straně 9.
Obdobně na straně 9 na konci tvrdí že se 454 sekvenační technologie používala především proto že dosahovala délky čtení s průměrem 700nt. Nesouhlasím, těchto délek bylo možné dosáhnout až v roce 2012/2013 pokud si uživatel připlatil za aktualizaci. Prvotní sekvenační výstupy byly v délkách okolo 100nt, později od asi roku 2008 to bylo okolo 200-250nt, později 400-450nt, a teprve potom s nástupem FLX+ verze bylo možné dosáhnout délek kolem 700nt. Tedy asi po 6-7 letech od uvedení na trh. Tvrzení v následující větě že "V současnosti je metoda 454 na ústupu zejména kvůli vysoké sekvenační ceně a chybovosti ..." je dosti troufalé. Především již není od roku 2014 podporována výrobcem a již léta nelze koupit od firmy Roche potřebné kity s chemikáliemi.
Poslední věta prvního odstavce na straně 10 konstatující "I přes uvedené nedostatky se metoda nadále uplatňuje v sekvenování genomů *de novo* a metagenomického profilování." je podle mne nepravdivá a bez citace případných aktuálních projektů. Já o žádných v roce 2018 nevím.
Postrádám citace u Technologie SOLID (kap. 1.1.3) v poslední větě odstavce, kde autor tvrdí že je vhodná pro detekci SNP polymorfizmů a diagnostice jednonukleotidových variant (poslední věta odstavce 1.1.3). Obdobně v kapitole 1.1.4 věnované platformě IonTorrent na konci odstavce tvrdí něco o jejich vhodnosti pro stanovení rozsáhlých duplikací a delecí genomu a o preimplantační diagnostice. Citace k tomu neuvádí žádné.
Z tohoto pohledu se mi rešerše na téma sekvenačních technologií nejeví aktuální. Osobně si myslím že tyto 3-4 stránky textu mohl autor úplně vynechat.

Kapitola "1.2.1 FASTQ" nemá jedinou citaci.

Odhlédnu-li od toho že například vysvětlení co je "CIGAR string" na straně 13 neobsahuje žádný odkaz, pak alespoň na straně 14 kde autor píše o tom co je doporučováno by měl jasně napsat kdo to

doporučuje. Hádám že hovoří o "Guidelines" na webu amerického Broad Institute které tam jsou zveřejněny spolu s dokumentací jejich GATK nástroje.

Chybějící citace, anglické výrazy nejsou v uvozovkách natož s označedním že jde o anglický termín. Například 'linuxového shellu' na straně 18 na konci prvního odstavce.

Není uvedena žádná citace k databázím NCBI SRA a NCBI GEO (ani URL). Na straně 17-18 je tak vpodstatě jedna celá strana textu bez jediné citace. Vzhledem k tomu že se jedná o kapitolu "1 Obecný úvod do problematiky", je to neobvyklé.

Na konci strany 18 autor píše o "identifikaci homologických částí proteinového překladu", přičemž má zjevně na mysli identifikaci sekvenčně podobných částí scaffoldů přeložených z DNA do proteinového zápisu. Jedná se o sekvenční podobnost, ne o homologii. O homologii nic nevíme. Viz kniha od autorů Koonin a Galperin (2003): Sequence, Evolution and Function – Computational Approaches in Comparative Genomics. Kluwer Academic Publishers.

Na konci poslední věty prvního odstavce na straně 19 chybí citace.

Řazení citací v seznamu citací je podle anglické abecedy a tak citace "Chyra Kufova et al., 2018" je pod písmemen "C" a ne podle české pod písmenem "Ch".

Na straně 18 je podivný překlad z angličtiny: "Sekvenační assemblery mohou využívat různé algoritmy ...". Bylo by lepší použít "Assemblery sekvencí mohou využívat ...".

Na straně 19 opět chybí řada citací již v první větě odstavce "1.4 Klinické využití MPS". Dále autor ve spodní čtvrtině téže stránky píše bez citace "Ačkoliv je v současnosti doporučováno u hereditárních onemocnění ..., domníváme se, že vzhledem ke klesající ceně vyšetření exomu, ...". Zdá se, že věta je odněkud opsána, protože autor nehovoří za sebe v jednotném čísle. Dále pokračuje větou "U detekce somatických variant získaných onemocnění je kladen důraz na co nejvyšší hloubku hotspotových míst.". **Postrádám citaci a vysvětlení o jakých "hotspotových" místech je řeč.** Dosud o nich v Úvodu nebyla žádná zmínka.

Na straně 20 je náročný bleskový přehled laboratorních metod ale skoro všechno je psáno bez citací (na celé stránce jsou pouze dvě). **Již po několikáté v disertační práci autor zmínil "adaptorové sekvence" a "indexy" a ani na této stránce vůbec nevysvětluje o čem mluví, jaké jsou jejich sekvence, natož aby nějak shrnul jejich varianty a odlišnosti/(ne)výhody využití. To bych považoval za prospěšnější než nekompletní úvod do různě zastaralých sekvenačních přístrojových řešení. Navíc se to velmi dotýká bionformatického zpracování dat v jeho vlastní práci.**

Na straně 21 kde začíná úvod do samotné bioinformatiky se rázem dostáváme k označením variant sestavení lidského genomu hg38/GRCH38 a hg19, bez citací, bez vysvětlení rozdílů mezi nimi, bez vysvětlení rozdílů mezi americkým a evropským způsobem označování chromozómů, bez rozlišení ALT contigů a zejména bez zdůraznění odlišností ve zpracování dat, atd.

Na straně 21 je bez citace, natož alespoň URL, rozebrán mechanismus programu BWA. Ještě ke všemu pak plynule přechází v diskusi rekalibrace PHRED kvalit nukleotidů a ostraňování PCR duplikátů. **Zde již ale jistě není řeč o programu BWA. Na konci strany 21 autor píše o optimalizaci alignmentu, opět bez citace a zjevně má na mysli krok IndelRealigner z programu GATK v3**, který byl mimochodem zrušen ve verzi 4.

Typickým překlepem je "priortizace", např. strany 6 a 48. Dále "Ovlivněných" v legendách k tabulkám na str. 63, 64 a 65.

Autor používá velmi krkolomné věty, např. "Pro proteinovou funkční predikci nesynonymních variant byly použity ..." na straně 55 na konci druhého odstavce.

Věta na straně 56 na konci čtvrtého odstavce nedává smysl: "Ovlivněné genové interakce ... byly zjišťovány pomocí webové aplikace Reactome ... a vlastními skripty ... ve srovnání se zdrojovými soubory Gene ontology ..., kde byly nejvíce ovlivněné ...". Mimo jiné, jedná se o kapitolu Materiál a metody, ne Výsledky či Diskuse, řekl bych že to sem nepatří.

"Potencionálně" na straně 57. Na straně 67 je uvedeno "při následném skríninku detekována ...".

Na straně 68 chybí alespoň jedna citace v poslední větě třetího odstavce: "Velká pozornost byla v problematice DBA a ribozomálního stresu věnována především mechanizmům poruch hematopoézy ..., zatímco rozsáhlejší soubor prací věnujících se imunitnímu systému u ribozomálního stresu chybí.".

Na straně 56 v kapitole 4.3 Výsledky autor konstatuje počty jedinců v českém národním registru DBA a uvádí počty s molekulárně diagnostikovanou příčinou. Zdá se, že se jedná o konstatování stávajících znalostí a ne o jeho vlastní výsledek. V tom případě zde chybí citace, alespoň na ať již seznam vícero originálních prací nebo na nějaký shrnující článek, nebo by stačilo i prosté konstatování že se nejedná o výsledky autora s odkazem na web registru pokud nějaký má. Z první věty následujícího odstavce 4.3.1 mi vyplývá, že předmětem práce autora bylo charakterizovat zbylých 19 pacientů (52 – 33 = 19). Práce která je ve formě manuskriptu se věnovala jenom jedné pacientce.

Na straně 74 v části Materiál a metody uvádíte anotační nástroje a databáze ale současně i uvádíte výsledky, cituji: "Tato doplňující anotace GO termínů pomohla zvýšit počet anotovaných genů na 17 885 z celkového počtu 26 074 predikovaných genů v ječmeni.".

Na straně 74 v části 5.3 Výsledky uvádí že byly použity rostliny jak s rekombinantním proteinem s vakuolární signální sekvencí (vAtCKX1) tak s odstraněnou signální sekvencí které mají protein lokalizován v cytoplazmě (cAtCKX1). Na straně 72 ale uvedl že transgenní rostliny měly pod kontrolou kukuřičného promotoru beta-galaktosidázy gen pro cytokinin dehydrogenasu. Zdá se že snad píše o něčem jiném. Až z jiných prací čtenář zjistí že geny pro cytokinin dehydrogenázy mají isoformy s označením CKX1, CKX2, CKX3. Příště by měl autor uvádět vedle názvů genů i jejich zkratky, chybí nejenom zde ale i v Závěru mini-článku. Je to další příklad nekvalitní transformace publikovaných prací, tam tyto nedostatky nejsou (viz Abstrakt práce Pospíšilová *et al.*, 2016 kde je uvedena zkratka genu ihned za jeho plným názvem).

Otázky a připomínky

Můžete přehledně uvést jak jste přispěl do tištěných publikací uvedených na straně 122? Uveďte prosím i "Impact Factor" a počet citací (bez autocitací) podle ISI Web Of Science.

V příloze 3.14 zobrazujete výstup z Vašeho nástroje. Na řádku "Barcode" by měl být index/barcode použitý k označení sekvencí ve směsném vzorku. Jako hodnota v políčku Platform by mělo být slovo ILLUMINA velkými písmeny, pokud se chcete držet standardu pro SAM/BAM/VCF. Hodnota "Platform unit" by měla být kombinací proměnných `$flowcellname` a `$lane_number` pokud chcete aby správně fungovalo vyhledávání artefaktů a duplikátů pomocí programu Picard. Viz https://software.broadinstitute.org/gatk/documentation/article.php?id=6472 . Osobně nesouhlasím s názorem že by se do Platform Unit mělo psát `${flowcellname}.${lane_number}.${sample_name}` a v diskusích tento názor podporuje více lidí s vysvětlení proč (v dokumentu je logická chyba).

Grafy v příloze 3.14 nemají nadpis a první dvojice nemá popsánu osu X.

Můžete uvést kdy plánujete zpřístup rozhraní k vašemu nástroji MOLDIMED pro veřejnost a nebo třeba uvolnit i jen samostatný anotační nástroj? Případně plánujete zveřejnit dokumentaci a schéma například databázových tabulek, pokud nelánujete zveřejnit celý software? Můžete srovnat Váš software s jinými zpracovatelskými balíky, například Gemini nebo s komerčními nástroji typu CLC Bio, Illumina BaseSpace, SeqPilot, VariantStudio, atp.?

Plánujete podporovat i jiné verze lidského genomového assembly než hg38? Podporuje váš nástroj analýzu vůči referenčnímu genomu rozšířenému o ALT contigy? Vyžaduje zpracování dat nějaké změny během výpočtů či analýz? Jaké a co z toho musí vědět uživatel/interpret?

Pomocí jakých programů byly vyvolány SNP/MNP v datasetu z projektu GIAB který jste použil pro srovnání s vašimi výstupy? Domnívám se, že Váš závěr že nejlepších výsledků dosahuje Vaše MOLDIMED pipeline v5 pomocí GATK HaplotypeCaller není nijak překvapivý protože i v rámci projektu GIAB byl použit tentýž nástroj GATK HaplotypeCaller, a tedy že jste potvrdil jen to že dostáváte podobné výsledky stejným programem. Viz strana 50, závěr prvního odstavce. **DNA použitá v projektu GIAB byla na světě použita již vícekrát a sekvenčních dat je velké množství. Mohl byste zpřehlednit kterou z verzí předpovědí jste použil a proč?**

Zatímco na Obrázku 4.1 alespoň v legendě vysvětlujete že pod označením C1 je probandka, pod C2 je její matka a pod C3 je její asymptomatická sestra, v Tabulce 4.1 na stránce 59 jsou prakticky bez legendy pro čtenáře nová označení vzorků rps7_1, rps7_2 a rps7_3. Ani v hlavním textu jsem nenašel zavedení těchto zkratek, byť bych je tak jako tak očekával v legendě tabulky 4.1.

Dále, dedukcí z jediné věty na stránce 58 která odkazuje na tuto tabulku soudím že řádek "Přiřazená čtení" v tabulce obsahuje ve skutečnosti "Jednoznačně přiřazená čtení". Mimochodem, v hlavním textu hovoříte o "unikátně mapovaných čteních", buďte jednotný. Tabulka měla pro lepší přehlednost a možnost kontroly obsahovat i celkové počty čtení na vstupu, lépe by se pak kontroloval výpočet v procentech. Legenda by byla mnohem lepší kdyby uváděla že se jednalo o mapování párových čtení pomocí programu STAR vůči referenčnímu genomu hg38 (čtenář při úvaze nad těmi miliony nejednoznačně přiřazených čtení tápe kolik jich asi bylo v procentech a zda to je hodně a nebo málo). **Proč jste použil program STAR a ne nějaký jiný? Jinde v práci jste použil TopHat. Můžete zhodnotit jak by se případně mohl lišit výsledek za použití jiného programu pro alignment a nebo jiného programu pro následné vyhodnocení?**

Jak si vysvětlujete pozorování že programem Cufflinks jste v datasetu rps7 získal asi 3x více expresně dysregulovaných genů (3022) než pomocí DESeq2 (955) kdežto u datasetu rps19 naopak asi 18.5x více nalezl program DESeq2? Viz obrázek 4.2 na straně 60. **Dále, šlo by obsah obrázku 4.2A převést do kontextu obrázku 4.5 ze strany 66 (který zobrazuje pouze skupinu genů na průniku DESeq2 a Cuflinks analýz)? Můžete vysvětlit proč podle Vás jsou si podobnější výsledky z RPS7_cufdiff s výsledky RPS19 ať již pomocí DESeq2 nebo Cuffdiff a z nějakého důvodu je RPS7_DeSeq2 odlehlý?**

Na straně 67 by se čtenář rád dozvědel kde je v práci doloženo "výrazné klastrování všech členů rodiny spolu s RPS19 skupinou". Obrázek 4.5 se přeci zaměruje jen na geny v průniku predikčních programů DESeq2 a Cuffdiff. **Diskutujete někde v práci to že mimo průnik bylo 955+3022 genů v případě rps7 a 4370+236 genů v případě rps19?** Genů mimo průnik je výrazně více.

Jak si vysvětlujete že stejná mutace hg38[chr2:g.3,580,153G>T] se jeví jako kauzální u studované probandky ale je přitom asymptomatická u její matky i sestry? Viz strana 57 a Obrázek 4.1 na straně 58.

V práci jsem si nevšiml dokladu že pozorovaná mutace RPS7 p.V134F v heterozygotní konstituci u všech členů rodiny RPS7 **je skutečně zodpovědná za odlišnosti v expresi studovaných genů.** Rozumím špatně první větě v části 4.4 Diskuse na straně 67? Cituji: "V této kapitole prezentujeme první případ nesynonymní varianty v genu RPS7 **jako molekulární příčinu DBA**". V kapitole 4.3.1 jste potvrdili expresi obou alel ve všech studovaných ženách z rodiny RPS7 pomocí sekvenace cDNA produktů, viz též Obrázek 4.1C na straně 58. Nakonec vidím na straně 54 nahoře zmínku o *in vivo* modelu který ale není součástí práce. V manuskriptu se hovoří o *in vitro* modelu. **Můžete to celé lépe vysvětlit pokud to je vaším výsledkem?**

V části Výsledky na straně 74 zmiňujete nějakou provedenou analýzu a odkazujete na práci Kokas *et al.*, 2016. Zdá se že jde o cizí práci ale vy jste spoluautorem. **Proč není tato práce přiložena k vaší disertační práci?** Dohledávání co jste analyzoval a co jste našel zbytečně zdržuje čtenáře Vaší práce.

Tabulka 5.1 má nedostatečný popis v legendě. **Co znamená označení sloupce "V (%)"? Mám si jako čtenář nečeho v tabulce všímat? Třeba že jsou hodnoty od 72% do 94%? Jak si vysvětlujete rozdíly? Je to dáno nekvalitní knihovnou (fragmentovaná cDNA, nebo kontaminací nerostlinnou DNA, nebo z jiného důvodu)? Byla tím nějak ovlivněna interpretace výsledků?**

Na straně 75 v části 5.3.1 kde jsou výsledky diferenciální exprese genů na mě text dělá dojem že byl opsán z nějaké grantové zprávy a nebo jiného článku. Druhý odstavec začíná větou: "Transkriptomová studie provedená v rámci této práce ...". Čtenář jen těžko dedukuje že "této práce" neznamená Vaši disertační práci ale asi práci Vojta *et al.*, 2016 která je zmíněná později uvnitř tohoto odstavce. V tom případě je ale nepochopitelné že v předchozím odstavci kde hovoříte o výsledcích z práce Pospíšilová *et al.*, 2016 zamlčujete svůj vlastní výsledek a jen lakonicky uvádítě počty vychýlených genů a jediný výstup který z této práce uvádíte ve své práci je Obrázek 5.1 na kterém je fotografie rostlin ječmene. Spíše bych očekával odkaz kde mám ve článku Pospíšilová *et al.*, 2016 hledat oněch statisticky významně vychýlených asi 400 genů a kde mám hledat těch 2400 genů z kořenů rostlin které jste Vy našel. Našel jsem si na ně sám odkaz ze strany 699 publikované práce a míří na Přílohy S2, S3 a S4. Ty ale přiloženy do disertační práce nejsou. Myslím si že stačilo uvést odkaz na původní Figure 2 z práce Vojta *et al.*, 2016 (na straně 682) a místo ve Výsledcích své disertační práce jste mohl věnovat něčemu co v publikaci není rozvedeno. A nebo alespoň do legendy ve Vaší disertaci uvést že jde o přetisk Figure 2 z práce Vojta *et al.*, 2016.

Obdobně, Tabulka 5.1 je mizerný přetisk Table 2 z práce Vojta *et al.*, 2016 přičemž jste ji ale nevhodně zjednodušil a vypustil informace které byly v Table 2 tučně. Vypustil jste i vysvětlení proč jsou v tabulce hodnoty *Nezobrazeno*, a **mě by přesto zajímalo proč u kategorie "Cellular response to hormone stimulus" na předposledním řádku nemáte zobrazeny geny.**

Tabulka 5.2 je obdobně jako Table 3A a Tabulka 5.3 odpovídá Table 4 v původní práci Vojta *et al.*, 2016.

K Tabulce 5.3: Proč vůbec zobrazujete a třídíte GO kategorie podle toho kolik obsahují zatříděných bílkovin? Je podstatné že kategorie která má dva exempláře z nichž shodou okolností oba měly ve vašich analýzách zvýšenou expresi (kategorie je tedy na 100% ovlivněna) nějak směrodatné? Nebylo by lepší třídit kategorie podle počtu skutečně ovlivněných genů?
Kategorie Biologický proces budou asi vždy obsahovat více přiřazených genů než kategorie s Molekulární funkcí, pokud nebudu uvažovat sběrné kategorie typu "DNA binding, "Ca2+-binding", atp.

Proč v Tabulce 5.2 není i srovnání s cAtCKX1 skupinou? Došlo i u ní ke zvýšené expresi RUBISCO (ribulose-bisphospate carboxylase)? Viz první řádek v tabulce. To ale není v souladu s Vaším závěrem na straně 80 v části 5.4.1 kde tvrdíte že naopak fotosyntéza byla ve transgenních rostlinách negativně ovlivněna. Je to zjednodušený závěr?

Zajímalo by mě proč jste nezobrazil a nediskutoval názvy významně pod-exprimovaných kináz v Tabulce 5.3 (pátý řádek od konce).

Můžete předvést pomocí vícenásobného alignmentu jak si jsou podobné cDNA pro geny HvCKX1 až HvCKX11 (cytokinin dehydrogenases) a HvIPT2 až HvIPT7 (isopentenyl transferases) z Figure 2 v práci Pospíšilová *et al.*, 2016? Můžete rozvést jak je velké riziko že bylo sekvenační čtení umístěno na jiný/chybný gen? Jak funguje Vámi použitý program pro alignment v případě více kandidátních oblastí? Minimálně v práci Vojta *et al.*, 2016 jsem si všiml že jste sekvenovali pouze z jednoho konce. To nebyl dobrý nápad.

Můžete uvést design sekvenování v práci Pospíšilová *et al.* (2016) a Vojta *et al.* (2016) (počet čipů, drah, knihoven, single- vs. paired-end sekvenace, délky čtení)?

Proč jste prováděli druhý tzv. technický replikát (sekvenování téže knihovny) když jste ho nijak nevyužili vyjma toho, že jste 2x provedl DESeq2 analýzu? Odstraňoval jste duplikáty čtení nebo artefakty vzniklé při přípravě knihovny? Technický replikát by měl zahrnovat i přípravu nové knihovny.

Při sekvenování jste nepoužívali indexy? V metodice se o nich nehovoří natož o jejich odstranění, vlastně se nehovoří ani o odstraňování adaptérů. Bylo tomu skutečně tak?

Osvětlete prosím obsah Figure 2 v publikaci Pospíšilová *et al.*, 2016. Označuje "Line" sekvenační dráhu (správně "Lane")? Můžete uvést kolik čtení jste měl pro geny ve Figure 2 a jak se to lišilo mezi "technickými replikáty"? Například v Table 2 v publikované práci jsou čísla v posledních 2 sloupcích dost odlišná. Jak si to vysvětlujete jestliže se jednalo o stejnou sekvenační knihovnu?

Shrnutí

Petr Vojta zjevně vyvinul velké úsilí na vývoj vlastního softwarového řešení ke skloubení cizích programů pro předpovídání SNP/MNP/InDel mutací které využívá jako zásuvné moduly a které občas vymění za jiné. Vyvinul ale vlastní samonosný anotační nástroj pro interpretaci významu mutací pro diagnostickou a klinickou praxi který funguje nejen nad výstupy z diploidních organismů ale i v případech s vyšší ploidií. Jistě nemalé úsilí bylo věnováno akreditaci softwarového řešení na analýzu celo-exomových sekvenací pro potřeby lékařské diagnostiky v ÚMTM v Olomouci.

Zvolené softwarové řešení odpovídá stávajícím požadavkům na uživatelsky přívětivý software a usnadňuje interpretaci výstupů. Nástroje a postupy použité pro analýzu transkriptomů rovněž odpovídají běžným standardům a jejich aplikací a interpretaci Petr Vojta zvládl náročný úkol. Jeho odborný záběr od medicíny přes rostlinnou fyziologii k programování v jazyce Python je obdivuhodný. Pokud bude nástroj MOLDIMED využíván širší skupinou uživatelů bude jistě přínosem pro své uživatelea se zpětnou vazbou pro Vás.

Předložené již otištěné publikace mají dobrou úroveň ačkoli nejsou špičkou v oboru. Jako plus pro předkladatale považuji to že je spoluautorem řady dalších prací a je zjevně zván ke spolupráci z jiných pracovišť. **Jeho disertační práci navrhuji k přijetí byť s výhradami vzhledem k mnoha připomínkám k textu předložené práce.** Domnívám se že samotné zpracování textu mohlo být výrazně lepší a autor se mohl více věnovat diskusi zvolených řešení a lépe pojmout úvod do problematiky. Pro úspěšnou vědeckou kariéru bude muset autor značně zlepšit svojí pečlivost při psaní odborného textu. Každopádně před ním nyní stojí nelehký úkol publikovat své softwarové řešení v recenzovaném časopise.

V Praze,
8. září 2018
RNDr. Martin Mokrejš, PhD.

Martin Mokrejs mok0039@vsb.cz

Digitally signed by :Martin Mokrejs mok0039@vsb.cz
Reason: I am the author of this document
Email: martin.mokrejs@vsb.cz
Date: 2018/09/08 14:45:09 +02'00'