

## Posudek oponenta na diplomovou práci

oponentský posudek

Jméno posuzovatele: **Tomáš Mašek**

Datum: 24.5.2011

Autor: **Anna Přistoupilová**

Název práce:

**Využití nových sekvenačních technik v biomedicinském výzkumu**

### Cíle práce

Cílem práce bylo vyhodnotit data získané sekvenací kódujících oblastí genomu pacienta s adultní autozomálně dominantní formou neuronální ceroid lipofuscinózy, na hrubých sekvenačních datech porovnat a zhodnotit aktuálně dostupné vyhodnocovací programy vhodné pro sekvenátor SOLiD4, anotovat nalezené mutace a porovnáním s výsledky vazebné analýzy postižené rodiny a pomocí stanovení exprese lidských genů vytipovat gen, či geny, jejichž proteinový produkt by mohl mít vztah k uvedenému onemocnění.

### Struktura (členění) práce, odpovídá požadovanému? **ANO**

Rozsah práce (počet stran): **90**

Je uveden anglický abstrakt a klíčová slova: **ANO**

Je uveden seznam zkratk? **ANO**

### Literární přehled:

Odpovídá tématu? **ANO**

Je napsán srozumitelně? **ANO**

Použil(a) autor(ka) v rešerši relevantní údaje z literárních zdrojů? **ANO**

Jsou použité literární zdroje dostatečné a jsou v práci správně citovány? **ANO**

Literární přehled je velice pěkně a srozumitelně napsán, nicméně si dovoluji malou poznámku. Přehled vysoko-průchodných sekvenačních technologií se v zásadě překrývá s informacemi, které jsou dostupné našim studentům např. v rámci přednášek "Functional Genomics" Dr. Vladimíra Beneše a kolektivu z EMBL Heidelberg nebo zazněly jako samostatná přednáška v rámci "Pokroků molekulární biologie". Myslím, že by práci prospělo, kdyby základní fakta o studované chorobě byly uvedeny už v literárním přehledu (začlenění do kapitoly Materiál a metody, nepovažuji osobně za nejvhodnější, ale jistě možné) a autorka uvedla další případy dědičných chorob, které už byly studovány pomocí "Next Generation Sequencing", jak se uvedené technologie osvědčily, případně jaké sekvenační platformy a vyhodnocovací programy byly použity, což by jistě autorce následně pomohlo diskutovat vlastní výsledky.

(malé nepřesnosti: 2.1.2. Sangerova metoda – je jistě i varianta se značeným sekvenačním primérem; 2.2.3.1. - dATP není substrátem luciferázy, nevolňuje se fosfát, ale pyrofosfát; 2.2.3.4. příprava knihovny – je navázán polyA adaptér, který poté hybridizuje k oligodT...)

### Materiál a metody:

Odpovídají použité metody experimentální kapitole? **ANO**

Kolik metod bylo použito? **Nelze zodpovědět**

Jsou metody srozumitelně popsány? **ANO**

Mám následující připomínky:

1: Jsem si vědom, že v závazných pravidlech pro sepisování diplomové práce na naší katedře je vyžadováno vysvětlení podstaty použitých metod, autorka však tuto povinnost pojala až příliš široce na úkor konkrétních informací. Mohu konstatovat, že popis principu metod je pojat v zásadě stejně jako literární úvod, tzn., autorka popisuje jejich všeobecné použití,

jejich výhody a nedostatky. Např. u kapitoly 3.3.3.1 (Expresní analýza) nepovažuji za účelné vysvětlovat, co je to transkriptom nebo, že množství mRNA nekoreluje s množstvím proteinu, atd.. U tzv. exomového sekvenování nechápu, proč autorka rozebírá jednotlivé metodické možnosti obohacování DNA o kódující oblasti, když byla použita jen jedna z nich?

2: Z uvedené práce není jasné, které metody autorka skutečně použila, alespoň v některém z kroků (např. příprava vzorku, apod.). Myslím si, že pokud autorka práce k uvedeným metodikám přistoupila až ve fázi vyhodnocení hrubých nebo již zpracovaných hrubých dat, tak měla popisovat jen kroky, které skutečně dělala a zbytek informací vhodně začlenit do výsledků, či diskuze. V opačném případě u všech experimentálních přístupů kromě přímé re-sequenace chybí konkrétní informace o experimentech (způsob izolace DNA, RNA, ověření kvality vzorku, či jeho další zpracování).

### **Experimentální část:**

Je vysvětlen cíl experimentů? **ANO**

Je dokumentace výsledků dostačující? **NE**

Hlavní nedostatek shledávám v tom, že autorka prezentuje pouze konečné výsledky často velice dlouhých a komplexních vyhodnocování dat, což čtenáři znemožňuje jakékoli kritické posouzení a uvedené závěry musí brát jako pouhá konstatování. To se týká především prvních třech kapitol. Např. u vazebné analýzy není specifikováno, u kterých členů rodiny byla provedena. U výsledků počtu kopií genomové DNA a expresní analýzy nejsou uvedena, žádná primární data, ani hrubě zpracované výsledky, např. nejsou uvedeny konkrétní identifikované transkripty, výsledek je prezentován pouze formou GO anotace do jednotlivých metabolických drah. V kapitole zabývající se přímým sekvenováním genu Hsp40 není specifikováno, u kterých členů rodiny analýza probíhala, specifita PCR produktů není dokumentována např. fotodokumentací příslušných agarózových gelů (což byl považoval za vhodné vzhledem k tomu, že bylo analyzováno pouze 5  $\mu$ l z reakce a purifikace neproběhla vyřezáním z gelu). Sekvenační chromatogram je vložen jen pro ilustraci výsledků. Autorka popisuje delecí Leu v pozici 116, pro niž jsou nemocní jedinci heterozygotní. Uvedené konstatování se mi na uvedeném chromatogramu sekvence nepodařilo odhalit (sekvence má 6N, vidím původní sekvenci CTCACG kódující Leu a Thr, druhou alelu jsem přečetl jako ATGACG kódující Met a Cys, rozhodně ne jako delecí, tak tedy nevím?) Dále v diskuzi autorka rozvíjí hypotézu, že delece Leu v pozici 116 by mohla mít vliv na správné fungování proteinu. Tento Leu se skutečně nachází v transmembránové doméně, respektive je též součástí domény o nízké homologii bohaté na cystein. Avšak ani delece Leu, ani záměna za Met-Thr nijak výrazně nemění transmembránový charakter této domény, není mi ani jasné, jaký vliv by měla mít uvedená mutace na palmitoylaci (tedy pravděpodobně Cys v pozici 118, který je zachován)? Autorka uvádí, že provedla analýzu vlivu nalezených mutací na strukturu a funkci proteinu, avšak žádná data v práci uvedena nejsou, takže výše uvedeným konstatováním nerozumím. Ačkoliv chápu, že práce s pacienty má svá specifika, nejsou mi jasná ani následující fakta a premisy. Expresní analýza byla provedena u 4 nemocných ze IV. generace pocházejících ze dvou manželství. Bylo nalezeno 330 genů se zvýšenou expresí, ale není specifikováno, jaké vzorky byly použity jako kontrola. RNA byla izolována z leukocytů, což nehovoří o expresi v mozku (což šlo alespoň hrubě zjistit pomocí "mikroarayových" databází). Nelze ani souhlasit s tím, že příčinná mutace se musí nacházet v genu se změněnou expresí (analyzovány byly kódující oblasti, ne regulační?).

Ohledně porovnávání jednotlivých mapovacích programů, jsem toho názoru, že měly být uvedeny všechny vstupní parametry i "default" charakteru, protože s novými verzemi softwaru se mohou měnit i přednastavené hodnoty. Předpokládám, že pro opravdu plnohodnotné srovnání uvedených programů, by bylo lepší porovnat jejich vhodnost na

několika různých sekvenováních lišících se např. délkou čtení, pokrytím, atd., nebo naopak otestovat vliv vstupních parametru analýzy na jednom experimentu (např. typ seedu u PerM, respektive jeho váha, počet povolených záměn apod.). V práci není uvedeno, zda delec Leu v pozici 116 odhalily i ostatní programy (asi ne, když některé neumí vyhledávat malé delece a inserce?), potažmo jak by se překrýval seznam unikátních 617 SNP nalezených programem Novoalign s ostatními programy. V kap 4.1.5. je uvedeno, že bylo pracováno se 25352 sekvenčními variantami identifikovanými programem Novoalign (počet mírně vyšší než předpokládaný), tab. 12 uvádí 14204? Myslím, že pro experimentální důkaz přesnosti mapování jednotlivých programů mělo být přímo sekvenováno více lokusů. V tab. 15 je výsledek bioinformatických analýz založený na postupném filtrování dat, jedním z kritérií byla analýza mutací pouze v kódujících oblastech, jak s tímto souhlasí uvedení mutace v genu RANBP1, která je v intronu?

#### **Diskuze:**

Je opravdu diskuzí, nejde jen o konstatování vlastních výsledků? **ANO**

Jsou výsledky porovnávány s literaturou? **Ano**

Jsou uvedeny nějaké hypotézy či návrhy na další řešení problematiky? **ANO**

#### **Závěry (Souhrn) :**

Jsou výstižné? **ANO**

#### **Formální úroveň práce** (obrazová dokumentace, grafika, text, jazyková úroveň):

Práce je po formální stránce výborná, minimum překlepů, velké plus celé práce....

(dNTP jsou nukleozidtrifosfáty, ne nukleotidtrifosfáty..)

#### **Splnění cílů práce a celkové hodnocení:**

Cíle práce byly splněny. Výsledky A. Přistoupilové budou mít určitě své uplatnění při vyhodnocování dalších sekvenčních dat, zejména pro domovskou laboratoř.

Za závažné nedostatky práce považují úroveň dokumentace výsledků, popis použitých metod a chybějící údaje o původu a přípravě vzorků pro jednotlivé analýzy. Práce je napsána v množném čísle (pravděpodobně z vrozené skromnosti), proto je těžké poznat, které pokusy autorka provedla samostatně, nebo na nich spolupracovala. Práce zpracovává poměrně specifické bioinformatické téma, s kterým nemám zkušenosti, a proto se autorce předem omlouvám, pokud jsem se mýlil v některých svých námitkách. Návrh klasifikace ponechám až na nadcházející obhajobu.

#### **Otázky a připomínky oponenta:**

1. Prosim, rekapitulujte, jaké části práce jste vypracovala samostatně, jaké ve spolupráci, a které byly vypracovány na zakázku, či v jaké fázi jste se zapojila do uvedených analýz. Jak dlouho jste uvedené výsledky zpracovávala?
2. Máte již nějaké výsledky vlivu delece Leu 116 na funkci proteinu Hsp40? Jaká je pravděpodobnost, že uvedeným filtrováním dat jste neodhalili další kauzální mutaci? Může být uvedené onemocnění multigenní povahy?
3. Jaký je vztah mezi rychlostí algoritmu, schopnosti namapování na referenční genom a nalezení známých variant? (Pomalejší algoritmus obecně nalezne více známých SNP? Je reciproční vztah mezi mírou detekce neznámých variant a mírou namapování na referenční genom?) Jaký je vztah mezi počtem povolených "mismatch bází", délkou jednotlivých "reads", rychlostí algoritmu a správnou detekcí SNP?
4. Lze si stáhnout hrubá sekvenční data, např. z databáze "1000 genomes" a testovat parametry jednotlivých programů na větším souboru dat?

Návrh hodnocení oponenta (známka nebude součástí zveřejněných informací)

výborně  velmi dobře  dobře  nevyhověl(a)

Podpis oponenta: