

Univerzita Karlova

Přírodovědecká fakulta

Studijní program: Bioinformatika

Studijní obor: Bioinformatika



Bc. Lucie Korená

*Analýza transkriptomu tasemnice *Mesocetoides Corti**

*Transcriptomic analysis of *Mesocetoides Corti**

Diplomová práce

Vedoucí práce: Mgr. Roman Leontovyč, Ph.D.

Praha, 2022

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 10.8.2022

.....

Bc. Lucie Korená

Poděkování

Zde bych chtěla poděkovat všem lidem, kteří mě podporovali během mého magisterského studia a kteří mi byli velkou psychickou oporou. Hlavní díky patří vedoucímu mé práce Mgr. Romanu Leontovyčovi, Ph.D. za velkou dávku trpělivosti a cenné rady při psaní celé práce. Dále bych chtěla poděkovat infrastruktuře Metacentrum, na které probíhaly výpočty k praktické části práce a členům parazitologické katedry UK za pomoc a poskytnutí materiálu k praktické části práce. Dále bych chtěla poděkovat všem svým přátelům a rodině, kteří při mně vždy stáli a pomohli mi zvládnout těžké i dobré chvíle během studia. Děkuji Danovi za tvrdou zkoušku jeho trpělivosti a nervů.

Abstrakt

Některé druhy parazitů včetně helmintů mohou v hostiteli inhibovat karcinogenezi. Protinádorový efekt byl v minulosti prokázán u tasemnic *Taenia crassiceps* a *Echinococcus granulosus*, u kterých byly identifikovány geny související s regresí rakoviny. V rámci výzkumu Helmintologické laboratoře katedry Parazitologie Přf UK byl zjištěn efekt suprese melanomu u tasemnicí *Mesocestoides corti*, avšak mechanismus účinku zatím zůstává neznámý. Pro navazující výzkum bylo důležité získat komplexní molekulární data v podobě transkriptomu vývojových stadií *M. corti*. Tato práce se zabývá studiem transkriptomického profilu tasemnice *M. corti* a rozdílem v genové expresi u tasemnic kultivovaných *in vitro* a v myších hostitelích (inbrední a outbrední kmen) pomocí metody RNA-seq. Cílem bylo nalézt upregulované transkripty tasemnic v myších hostitelích, které by mohly mít potenciální vliv na regresi rakoviny. Byla provedena analýza diferenciální genové exprese, ze které vyplývá, že tasemnice v myších hostitelích (bez ohledu na kmen) mají více upregulovaných transkriptů než tasemnice kultivované *in vitro*. Analýzou vysoce upregulovaných transkriptů tasemnic v myších hostitelích bylo identifikováno několik transkriptů, které byly zmíněny ve vědeckých článcích jako možné supresory rakoviny. Mezi ně patří například *KATNAL1*, *PTPRJ*, *ZC3H14* a *CD148*. Analýza genového obohacení (gene enrichment analysis) ukázala, že většina upregulovaných drah u tasemnic v myších hostitelích souvisí s metabolismem. Naproti tomu u tasemnic kultivovaných *in vitro* souvisí nejvíce upregulované dráhy s buněčnými, environmentálními a genetickými procesy. Studium protirakovinného efektu tasemnic a jeho pochopení na molekulární úrovni může výrazně pomoci v budoucím výzkumu léčby nádorových onemocnění.

Klíčová slova: transkriptom, diferenciální genová exprese, rakovina, bioinformatická analýza, tasemnice, *Mesocestoides corti*

Abstract

Some species of parasites, including helminths, can inhibit carcinogenesis in their hosts. The antitumoral effect has been discovered in the tapeworms *Taenia crassiceps* and *Echinococcus granulosus*, which genes associated with cancer regression have been identified. The effect of melanoma suppression has also been observed in tapeworm *Mesocostoides corti* by the Laboratory of Helminthology, Department of Parasitology, Faculty of Science, Charles University, however the mechanism-of-action, remains unknown. For the upcoming research it was essential to have the complex molecular data such as transcriptome of the developmental stages of *M. corti*. This work is focused on the transcriptomic profiling of the tapeworm *M. corti* and the differential gene expression in two different strains of murine hosts (inbred and outbred) using the RNA-Seq. The main goal was to identify upregulated transcripts in the tapeworms from the murine hosts that could have a potential effect on cancer regression. Differential gene expression analysis was performed, and the results showed that tapeworms in murine hosts (regardless of strain) had more upregulated transcripts than tapeworms cultured *in vitro*. Analysis of highly upregulated transcripts in the tapeworms that were grown in the murine hosts identified several transcripts that have been previously described as potential tumor suppressors. Some of these proteins/genes include: KATNA1, PTPRJ, ZC3H14 and CD148. On the other hand, in *in vitro* cultured tapeworms, the most upregulated pathways are related to cellular, environmental, and genetic processes. Studying of the anticancer effect of tapeworms and understanding the molecular mechanisms could significantly help in further research of the cancer treatment.

Keywords: transcriptome, differential gene expression, cancer, bioinformatics analysis, tapeworm, *Mesocostoides corti*

Obsah

Slovník použitých pojmů	1
1 Úvod a cíle práce	2
2 Literární přehled	3
2.1 Parazit jako supresor rakoviny	3
2.1.1 <i>Mesocestoides corti</i> (syn. <i>M. vogae</i>)	6
2.2 Bioinformatické zpracování RNA-Seq dat	8
2.2.1 Kontrola kvality vstupních dat	9
2.2.2 Předzpracování sekvenačních readů (Preprocessing)	10
2.2.3 Kvantifikace genové exprese	11
2.2.4 Normalizace počtu readů	13
2.2.5 Diferenciální genová exprese	15
2.2.6 Anotace transkriptomu	16
2.2.7 Analýza obohacení genové sady (Gene enrichment analysis)	20
3 Materiál a metody	23
3.1 Sběr materiálu, izolace RNA a sekvenace	23
3.2 Bioinformatická analýza	24
3.2.1 Sekvenační data, předzpracování dat a kontrola kvality	24
3.2.2 Anotace referenčního transkriptomu	25
3.2.3 Kvantifikace a analýza diferenciální genové exprese	25
3.2.4 Sloučení výstupů a analýza obohacení genové sady	25
4 Výsledky	27
4.1 Izolace RNA a sekvenační data <i>M. corti</i>	27
4.2 Anotace referenčního transkriptomu	28
4.3 Kvantifikace a analýza diferenciální exprese	30
4.4 Analýza obohacení genové sady (Gene enrichment analysis)	34

5	Diskuze	38
5.1	Izolace RNA a sekvenační data <i>M. corti</i>	38
5.2	Anotace referenčního transkriptomu	38
5.3	Kvantifikace a analýza diferenciální genové exprese	39
5.4	Analýza obohacení genové sady (Gene enrichment analysis).....	41
6	Závěr	42
7	Přílohy	43
8	Zdroje	45
9	Online zdroje	54

Slovník použitých pojmů	
Adaptéry	Krátké sekvenčně specifické oligonukleotidové sekvence, připojené k fragmentům nukleové kyseliny, která bude sekvenována. Poskytují vazebné místo pro primery při amplifikaci
Assembly	Složení sekvenačních readů do delších úseků na základě jejich překryvů
RNA-Seq read	Záznam sekvence nukleotidů získaných z RNA-Seq sekvenačních platforem • Single-end ready – sekvenační ready pochází z jednoho konce fragmentu • Paired-end ready – sekvenační ready pochází z obou konců stejného fragmentu
Referenční genom	Sekvence genomu, která může být použita pro mapování vstupních readů pro následnou analýzu
Referenční transkriptom	Celkový soubor syntetizované RNA (především mRNA a nekódující RNA) přítomné v dané buňce za určitých podmínek v určitém čase

1 Úvod a cíle práce

Parazitičtí helminti infikují miliony lidí po celém světě. Tito parazité mohou být původci několika typů rakoviny (karcinom močového měchýře, cholangiokarcinom, kolorektální hepatocelulární karcinom a spinocelulární karcinom). Novodobé studie však poukazují na to, že infekce parazitem mohou vést ke zpomalení některých druhů rakovinotvorných bujení či k úplné regresi rakoviny. Přesné mechanismy ovlivnění rakoviny však nejsou zcela známy. Parazit může vyvolat protirakovinný účinek sekrecí specifických molekul nebo aktivací imunitní odpovědi hostitele. Studium protinádorové aktivity může vést k novým způsobům léčby nádorových onemocnění.

Protinádorová aktivita byla prokázána u tasemnice *Echinococcus granulosus*, u které byl nalezen například protein *EgKl-1*, který je schopen přímo zabít rakovinné buňky (Ranasinghe a McManus 2018). Tato práce se zabývá transkriptomickou analýzou vývojového stádia tasemnice *Mesocostoides corti* a vlivem prostředí na genovou expresi. *Mesocostoides corti* na rozdíl od jiných druhů tasemnic není vysoce patogenní pro člověka, což umožňuje bezpečnou laboratorní manipulaci.

Literární úvod práce je zaměřen především na zpracování a analýzu RNA-Seq dat získaných ze sekvenační platformy Illumina, protože v porovnání s ostatními sekvenačními platformami má relativně nízkou míru chyb sekvenování a nabízí větší hloubku sekvenování, což umožňuje detekci málo exprimovaných transkriptů (Chu a Corey 2012). Další výhodou je i nižší cena za sekvenování oproti jiným sekvenačním platformám a pomocí této platformy byly sekvenovány data pro praktickou část práce.

Cíle diplomové práce:

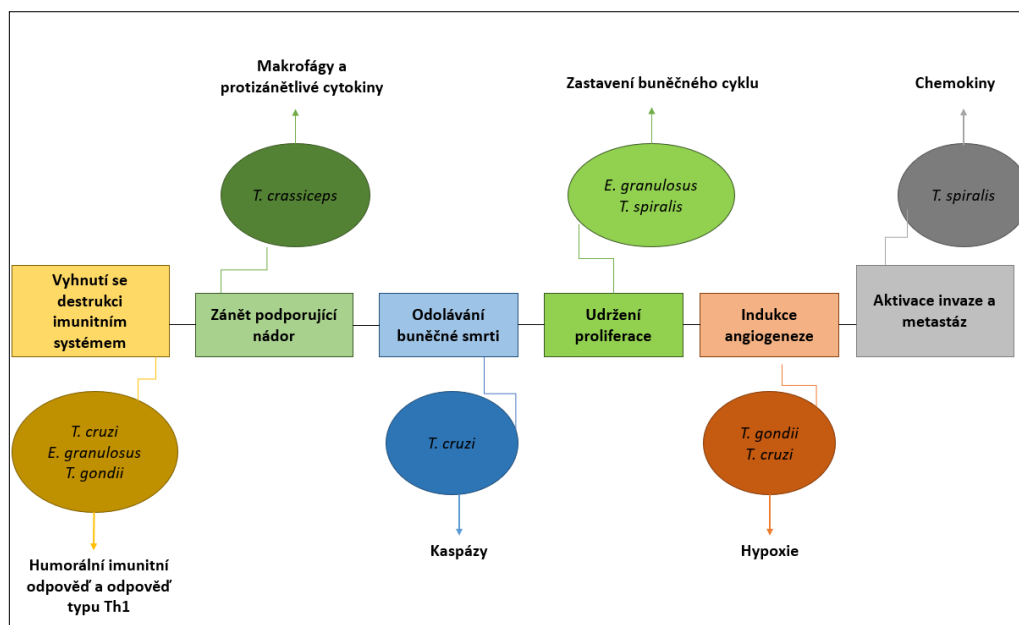
- Shrnutí jednotlivých kroků analýzy diferenciální genové exprese s příklady dostupných softwarů.
- Analýza transkriptomu vývojového stádia tasemnice *M. corti* kultivované *in vitro* a ve dvou kmenech myších hostitelů (inbrední, outbrední).
- Diferenciální genová analýza a s ohledem na vliv prostředí na genovou expresi.
- Analýza genového obohacení biologických drah (gene enrichment analysis).
- Analýza diferenciálně exprimovaných genů zaměřená na geny s možnou supresorovou aktivitou vůči rakovině.

2 Literární přehled

2.1 Parazit jako supresor rakoviny

Parazité a lidé spolu koexistují již miliony let. Řada druhů parazitů jsou známými původci rakoviny u savců. Novodobé studie však poukazují na to, že některé druhy parazitů včetně helmintů mohou vyvolat protinádorové účinky v hostitelském organismu (*Tab. 2*). Helminti jsou také součástí hygienické hypotézy. Ta v původní podobě poskytuje vysvětlení nárůstu alergických a autoimunitních onemocnění v ekonomicky vyspělých západních státech ve 20. století, kde nejsou děti v raném dětství tolik vystaveni určitým typům patogenů (Strachan 1989). Rozšíření této hypotézy spočívá v tom, že jedinci vystaveni infekcí helmintů mají nižší sklon k alergiím, astmatu, autoimunitním chorobám a rakovině (Finlay et al. 2014).

Rakovina je stav, kdy se buňky abnormálně dělí a mohou napadat i jiné tkáně. Tyto buňky mají vyvinuté mechanismy, pomocí kterých se vyhýbají supresorům růstu, odolávají buněčné smrti, udržují proliferativní signalizaci, mají replikační nesmrtelnost a další (Callejas et al. 2018). Různé znaky rakoviny mohou být ovlivněny parazity a jejich produkty (**Obr. 1**). Protirakovinné mechanismy indukované parazity, nejsou zcela prozkoumány, ale mohlo by se jednat o prezentaci běžných antigenů, zvýšení vrozené nebo získané imunity nebo indukce angiogeneze (Darani a Yousefi 2012).



Obr. 1: Paraziti a jejich potenciální účinek na charakteristický mechanismus rakoviny (Callejas et al. 2018, upraveno). *Toxoplasma gondii* a *Trypanosoma cruzi* mají protinádorový účinek na některé typy rakovinných buněk prostřednictvím indukce angiogeneze, reaktivace imunitní odpovědi a indukce apoptózy. *Taenia crassiceps* má schopnost regulovat zánětlivou reakci podporující nádor. *Echinococcus granulosus* má více protirakovinných mechanismů jako je reaktivace imunitní odpovědi a antiproliferativní účinek na nádorové buňky a společně s *Trichinellou spiralis* regulují aktivaci invaze a metastáze.

Helminti se řadí mezi skupinu organismů, které mohou karcinogenezi inhibovat, nebo vyvolávat. Potvrzené druhy karcinogenních helmintů (Tab. 1) vyvolávají rakovinu pomocí ne zcela objasněných mechanismů, jejichž hlavním rysem je chronická zánětlivá odpověď hostitele (Vennervald a Polman 2009). V tuto chvíli je známo několik helmintů s prokázanými protinádorovými aktivitami (Tab. 2). Protinádorová aktivita může být přímá pomocí helmintem uvolňovaných molekul nebo nepřímá, která je výsledkem adaptivní imunity generované proti specifickým parazitickým antigenům. Tyto nové poznatky mohou pomoci k vyvinutí nových způsobů léčby rakoviny pomocí imunoterapie.

Tab. 1: Příklady jednoznačných karcinogenních helmintů podle IARC (Benamrouz et al. 2012, upraveno).

Parazit	Druh rakoviny	Mechanismus působení	Zdroj
<i>Clonorchis sinensis</i>	Cholangiokarcinom	Přímé fyzické poškození a chemické podráždění žlučovodu a okolních jaterních tkání	(Parkin 2006)
<i>Opisthorchis viverrini</i>	Cholangiokarcinom	Přímé fyzické poškození žlučovodu a chemické podráždění, imunopatologie v důsledku zánětu	(Parkin 2006)
<i>Schistosoma haematobium</i>	Karcinom močového měchýře	Nejasný (možné chronické podráždění močového měchýře)	(Vennervald a Polman 2009)

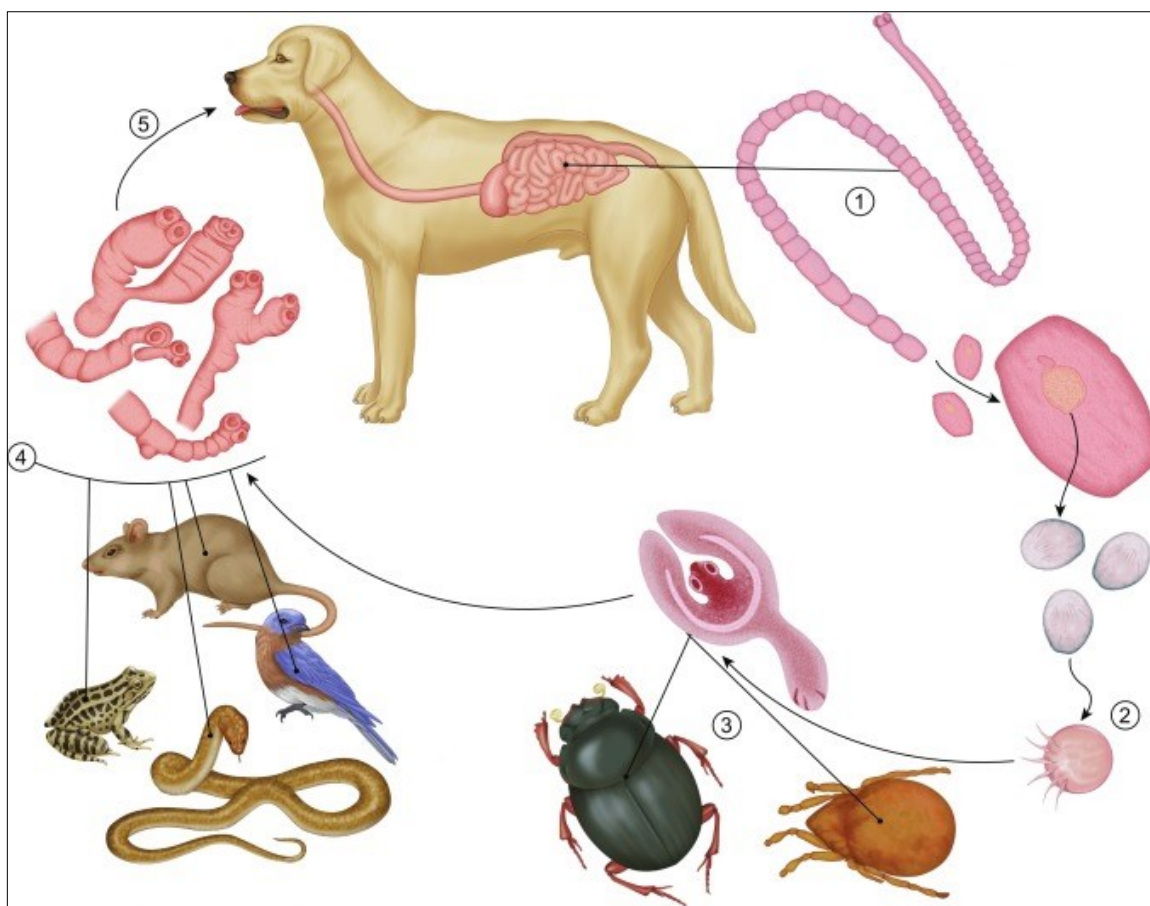
Tab. 2: Paraziti a jejich pravděpodobná protinádorová aktivita. (Callejas et al. 2018, upraveno)

Parazit	Druh rakoviny	Mechanismus působení	Zdroj
<i>Echinococcus granulosus</i>	Karcinom prsu a kolorektální karcinom	Produkce protilátek pro rozpoznávání nádorových buněk	(Berriel et al. 2013), (Noya et al. 2013)
	Fibrosarkom	Není známo	(Yousofi Darani et al. 2012)
<i>Taenia crassiceps</i>	Kolorektální karcinom spojený s nespecifickými střevními záněty	Snížení rekrutování zánětlivých monocytů a zánětu v tlustém střevě	(León-Cabrera et al. 2014)
<i>Toxoplasma gondii</i>	Melanom	Aktivace CD8 ⁺ , NK buněk a exprese MHC-I a MHC-II v APC	(Baird et al. 2013)
	Fibrosarkom	Zvýšení aktivity cytotoxických T buněk	(Motamedi et al. 2009)
	Melanom a karcinom plic	Potlačení neovaskularizace indukci hypoxie avaskulární nekrózy	(Hunter et al. 2001), (Kim et al. 2007)
<i>Trichinella spiralis</i>	Melanom	Redukce plicních metastáz prostřednictvím CXCL9, CXCL10, IL-4, CXCL1 a CXCL13	(Kang et al. 2013)
	Buněčná linie lidského hepatomu (HT402) a lidské chronické myeloidní leukémie (K562)	Zastavení buněčného cyklu v G1 nebo v S fázi	(Wang et al. 2009), (Wang et al. 2013)

<i>Trypanosoma cruzi</i>	Karcinom prsu a kolorektální karcinom	Aktivace CD4 ⁺ a CD8 ⁺ buněk a produkce protilátek proti rakovinným buňkám	(Ubillos et al. 2016)
	Experimentální adenokarcinom prsu	Parazit pomáhá odhalit přítomnost nádorových buněk v imunitním systému	(Abello-Cáceres et al. 2016)
	Karcinom prsu	Inhibice proliferace a migrace endoteliálních buněk	(López et al. 2010)
	Melanom	Rekombinantní protein J18 indukuje apoptózu prostřednictvím kaspázy 3	(Atayde et al. 2008)

2.1.1 *Mesocetoides corti* (syn. *M. vogae*)

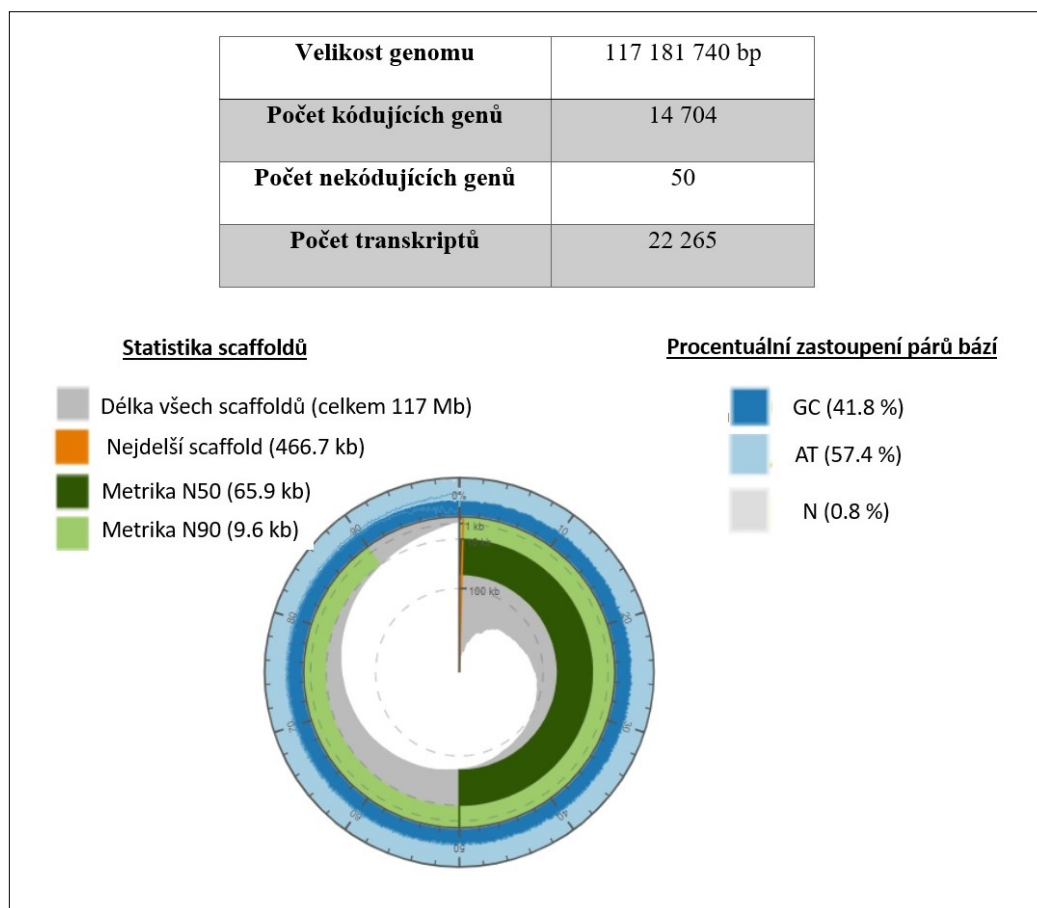
Mesocetoides corti je druh tasemnice patřící do třídy Cestoda, která obsahuje přes 3000 druhů střevních parazitů infikujících zejména savce včetně člověka. Životní cyklus (**Obr. 2**) zahrnuje dvě larvální stádia a dva meziphostitele. Prvním meziphostitelem jsou pravděpodobně zástupci členovců, kteří pozřou vajíčka s onkosférami a v jejich těle probíhá vývoj prvního larválního stádia cysticerkoidu. Druhé larvální stádium tetrathyridium se vyvíjí v druhém meziphostiteli, kterým může být malý savec, plaz, obojživelník či pták. Tetrathyridia jsou významné svým nepohlavním rozmnožováním pomocí buněčného dělení, díky němuž lze snadno laboratorně studovat infekci larválními stádii u myší (Siles-lucas a Hemphill 2002). Definitním hostitelem jsou zástupci masožravců, kde probíhá proces strobilace neboli vývoj larvy v dospělce. Vyvinutá tasemnice se pomocí scolexu přichytí k tenkému střevu (Saari et al. 2019).



Obr. 2: Životní cyklus *Mesocestoides corti* (Saari et al. 2019). 1. Dospělce v tenkém střevě definitního hostitele vylučující zralé proglotidy, které jsou vylučovány do prostředí; 2. Vyloučené proglotidy s vajíčky, vývoj onkosféry; 3. První meziphostitelé (členovci), vývoj cysticerkoidu; 4. Vývoj tetrathyridia (nepohlavní rozmnožování) v druhém meziphostiteli (plazi, obojživelníci, drobní savci, ptáci); 5. Definitní hostitel (pes), který se nakazí pozřením druhého meziphostitele, proces strobilace (vývoj larvy v dospělce).

Ve vhodných kultivačních podmínkách lze pozorovat vývoj larvy až do dospělce díky čemuž lze studovat komplexní životní cyklus, který je důležitý pro studium vztahů mezi hostitelem a parazitem (Markoski et al. 2003). Další výhodou je i bezpečná laboratorní manipulace, neboť je zde nízké riziko přenosu parazita na člověka (Camargo de Lima et al. 2020).

K dispozici je osekvenovaný referenční genom *M. corti* (Online zdroj č. 1), který byl vytvořen jako součást projektu 50 Helminth Genomes Project (Online zdroj č. 2). Z genomu bylo odvozeno 22 265 transkriptů. Kvalita referenční genomové assembly (**Obr. 3**) je dostatečně dobrá pro provedení procesu anotace. Na tomto organismu byl již studován transkriptomický profil dvou vývojových stádií (tetrathyridia a dospělce) pomocí RNA-Seq (Basika et al. 2019) a identifikace a profilace mikroRNA dvou vývojových stádií (tetrathyridia a dospělce) (Basika et al. 2016).



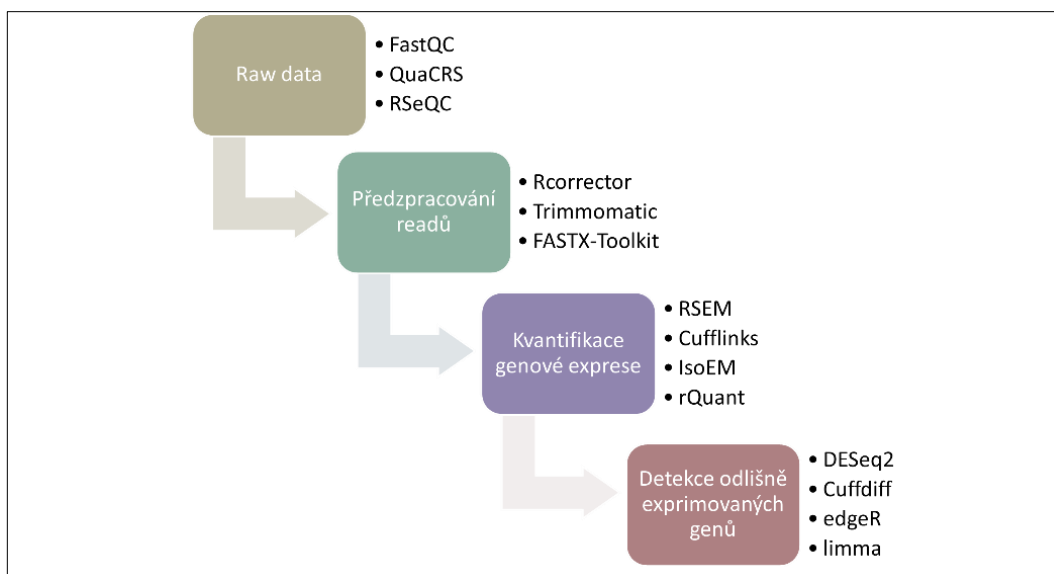
Obr. 3: Popis referenčního genomu organismu *Mesocestoides corti* (Online zdroj č. 1).

2.2 Bioinformatické zpracování RNA-Seq dat

Sekvenování RNA-Seq je metoda generující velké množství RNA sekvenačních dat, kterou na rozdíl od dříve hojně využívané metody microarrays (Murphy 2002) lze použít i pro detekci nových genů a izoforem pro jakýkoliv organismus. Využívá se v celé řadě analýz například analýza genové exprese, analýza alternativního sestřihu, detekce genových fúzí a mnoho dalších. Díky velkému množství aplikací neexistuje univerzální postup zpracování dat a univerzální nástroje, které by byly využitelné pro všechny druhy analýz a pro všechny druhy organismů.

Jednou z aplikací RNA-Seq je analýza diferenciální genové exprese, která se používá k identifikaci odlišně exprimovaných genů. Jedná se o klíčový proces pro pochopení změn fenotypů a vlastností buněk při onemocnění či při odlišných podmínkách. Tato analýza zahrnuje proces mapování readů na referenční sekvenci (genom, transkriptom), kvantifikaci namapovaných readů a detekci odlišně exprimovaných genů na základě statistických testů (Obr. 4). K dispozici je řada softwarů a volba nejvhodnějšího softwaru závisí především

na počtu replikátů, použité sekvenační technologii a složitosti sekvenovaného organismu. V následujících podkapitolách jsou popsány jednotlivé kroky analýzy a je kladen důraz především na softwary, které byly následně použity v praktické části práce.



Obr. 4: Postup analýzy genové exprese s příklady nejčastěji používaných softwarů. Raw data (nezpracovaná) – kontrola kvality osekvenovaných readů; Předzpracování readů – odstranění adaptérů a filtrování sekvenačních chyb, které by mohly zkreslit celou analýzu; Kvantifikace genové exprese – na základě procesu mapování readů na referenční sekvence jsou identifikovány četnosti jednotlivých genů nebo izoforem; Diferenciální genová exprese – provedení statistických testů založených na kvantifikaci genové exprese.

2.2.1 Kontrola kvality vstupních dat

Kvalita nezpracovaných sekvenačních dat odráží spolehlivost výsledků analýzy. Pro kontrolu kvality sekvenačních dat z platformy Illumina se nejčastěji používá program FastQC (Online zdroj č. 3), který provádí kontroly kvality sekvencí, kontroluje obsah GC párů, přítomnost adaptérů a další (Online zdroj č. 3). Program RSeQC (Wang et al. 2012) je rozšířen o kontroly kvality specifické pro RNA-Seq data, jako je např. kontrola hloubky sekvenování, distribuce readů a jednotnost pokrytí (coverage uniformity)¹. RNA-QC-chain (Zhou et al. 2018) je software, který na rozdíl od jiných nástrojů pro kontrolu kvality RNA-Seq provádí i trimování readů, detekci kontaminace ribozomální RNA (rRNA) a automatickou identifikaci readů pocházejících z kontaminací jiných organismů. Jednotlivé

¹ Jednotnost pokrytí – metrika, která kontroluje rovnost distribuce readů v cílových sekvenčních oblastech, pokud se ready primárně hromadí na 3' konci transkriptů, znamená to nízkou kvalitu readů

kroky kontroly provádí buď postupně za sebou, nebo nezávisle na sobě, což umožňuje možnost paralelizace, čímž je výrazně zvýšena rychlost celého procesu (Zhou et al. 2018).

2.2.2 Předzpracování sekvenačních readů (Preprocessing)

Přítomnost chyb sekvenování, nekvalitních a adaptérových sekvencí zkresluje celkový průběh analýzy a může vést k nesprávným výsledkům. Kvalitu jednotlivých osekvenovaných bází v platformě Illumina vyjadřuje „phred quality score“ (Phred, Q skóre), které nabývá hodnot 0 až 50. Vyjadřuje pravděpodobnost, že daná báze byla chybně přečtena ($Q = -10 \log_{10}P$, kde P je pravděpodobnost, že daná báze byla chybně přečtena). Čím vyšší hodnota, tím vyšší je kvalita osekvenované báze (Online zdroj č. 4). Odstranění adaptérů probíhá na základě shody sekvence s databází nejčastěji používaných adaptérových sekvencí.

Součástí analýzy diferenciální genové exprese je proces mapování, při kterém jsou vstupní trimované ready mapovány na referenční sekvenci. Bylo ukázáno, že velmi krátké ready, které mohou být výsledkem trimování, se mohou mapovat na více míst a tím výrazně ztížit analýzu. Proto je vhodné při trimování definovat minimální délku readů (Williams et al. 2016). Některé mapovací softwary mají funkci soft-clipping², která se provádí paralelně během procesu mapování. Byla provedena studie (Liao a Shi 2020), která se zaměřila na přesnost kvantifikace genové exprese s otrimovanými ready a ready bez trimování s použitím mapovacího softwaru s funkcí soft-clipping. Tato studie ukázala, že proces trimování nezlepšuje přesnost kvantifikace (ověřeno pomocí TaqMan RT-PCR). Při vynechání kroku trimování dat došlo k řádovému zrychlení celé analýzy (Liao a Shi 2020).

V rámci předzpracování sekvenačních readů se provádí oprava chybně osekvenovaných bází. Vzhledem k rozdílné povaze genomových a transkriptomových sekvenačních dat se doporučuje používat software navržený přímo na RNA-Seq data (MacManes 2015).

Jedním z dostupných softwarů je Rcorrector (Song a Florea 2015), který je navržený pro opravu sekvenačních chyb sekvencí získaných z platformy Illumina. Výhodou je časová a paměťová složitost, neboť nevyžaduje značné hardwarové zdroje a lze jej spustit prakticky

² Soft-clipping – při procesu mapování jsou ignorovány části readů, které se na obou koncích neshodují s referencí

na jakémkoliv stolním počítači. Proto je vhodný pro analýzu druhů organismů s velkými a komplexními transkripty a genomy (Song a Florea 2015). Trimmomatic (Bolger et al. 2014) a ngsShoRT (next-generation sequencing Short Reads Trimmer) (Chen et al. 2014) jsou trimovací softwary, které slouží k odstranění adaptérových sekvencí a k filtrování chybných bází sekvenčních dat z platformy Illumina. Implementují různé druhy algoritmů a jejich výhodou je paralelizace, která výrazně snižuje časovou i paměťovou složitost celého procesu. Nástroj ngsShoRT je lepší, co se týče jak časové složitosti, tak kvality trimovaných readů (Chen et al. 2014).

2.2.3 Kvantifikace genové exprese

Při kvantifikaci genové exprese se provádí mapování sekvenčních readů na referenční genom nebo transkriptom. Při studiu exprese transkriptů lze mapovat ready na referenční transkriptom, čímž se výrazně sníží časová složitost (neboť velikost referenčního genomu je daleko větší). Nicméně mapováním readů na referenční genom lze studovat expresi daleko širší škály transkriptů, díky alternativnímu splicingu. Pokud není k dispozici referenční sekvence, lze pomocí readů ze všech replikátů za všech zkoumaných podmínek sestavit *de novo* „referenční“ transkriptom (Raghavan et al. 2022). Metoda Illumina RNA-Seq generuje relativně krátké sekvenční ready v případě pair-end sekvenování 2 x 151bp – 2 x 301bp (Online zdroj č. 5), což mapování výrazně komplikuje. Ready se mohou mapovat na více genů či izoform. Kvantifikační softwary tyto nejednoznačně namapované ready buď do analýzy nezahrnují, nebo se je snaží identifikovat pomocí metody Maximum likelihood či Bayéské metody (Garber et al. 2011).

S vývojem sekvenčních metod souvisí i rozvoj mapovacích softwarů (Tab. 3). Volba nejvhodnějšího softwaru pro konkrétní analýzu spočívá v kompromisu mezi přesností a rychlostí procesu. Přesnost mapování lze částečně určit podle procenta namapovaných readů či podle odhadovaného genového pokrytí. Rychlost procesu je vedlejším problémem, nicméně s rostoucím množstvím dat je časová složitost důležitý faktor analýzy (Musich et al. 2021). Mapovací softwary jsou rozděleny do 2 kategorií: „**splice-unaware**“ (zarovnávají k referenčnímu transkriptomu, nemožnost rozdělit ready na hranicích intron-exon), „**splice-aware**“ (zarovnávají k referenčnímu genomu, možnost rozdělit ready na hranicích intron-exon).

Tab. 3: Příklady mapovacích softwarů (Online zdroj č. 6).

Mapovací software	Reference	Popis	Aktuální verze (2022)	Citace
Bowtie	Transkriptom	Ultrarychlý, paměťově efektivní, zaměřený na velké soubory krátkých readů s velkými transkriptomy	1.3.1	(Langmead et al. 2009)
Bowtie2	Transkriptom	Ultrarychlý, paměťově efektivní, vhodný pro zarovnání readů k dlouhým referenčním sekvencím (např. savčím genomům)	2.4.4	(Langmead a Salzberg 2012)
BWA	Transkriptom	Vhodný pro mapování málo divergentních sekvencí proti velkému referenčnímu genomu (např. lidský)	0.7.17	(Li a Durbin 2009)
HISAT2	Genom	Rychlý software pro mapování readů sekvenování nové generace na referenční genomy, nástupce TOPHAT	2.2.1	(Kim et al. 2015)
STAR	Genom	Ultrarychlý a univerzální mapovač, schopnost hledat fúzní geny ³	2.7.10a	(Dobin et al. 2013)
TOPHAT	Genom	Vhodný pro mapování RNA-Seq readů s velkými referenčními genomy	2.1.1	(Trapnell et al. 2009)

Při porovnání softwarů pro mapování na referenční genom jsou softwary TOPHAT a HISAT2 vhodnější spíše pro organismy s nízkou složitostí genomu⁴, zatímco software STAR je vhodný pro složitější genomy (Waranid Sutthacharoenthad et al. 2018).

Příkladem dříve hojně využívaného spojení softwarů je spojení mapovacího softwaru TopHat (Trapnell et al. 2009) a balíčku softwarů Cufflinks (Trapnell et al. 2010). Program TopHat provádí mapování s referenční sekvencí a jeho výstupní soubor zpracovává program Cufflinks, který provádí kvantifikační analýzu a na výstupu používá normalizační metodu FPKM (viz 1.1.4 Normalizace). Součástí balíčku Cufflinks je i program Cuffdiff2, který provádí diferenciální analýzu (viz 1.1.5 Diferenciální genová exprese). Tento protokol je

³ Fúzní gen – hybridní gen vytvořený ze dvou dříve nezávislých genů

⁴ Složitost genomu – odvíjí se od relativního množství unikátních a repetitivních sekvencí

omezen na organismy s existujícím referenčním genomem a sekvenační data z platform Illumina nebo SOLiD (Trapnell et al. 2012).

Programy IsoEM (Nicolae et al. 2011) a RSEM (Li a Dewey 2011) jsou srovnatelné z hlediska přesnosti při kvantifikačních analýzách s použitím paired-end readů, ale při použití single-end readů je RSEM o něco přesnější (Li a Dewey 2011). RSEM standardně využívá mapovací software Bowtie (Langmead et al. 2009), naproti tomu program IsoEM vyžaduje na vstupu již namapovaný soubor ve formátu SAM. Výsledné odhady exprese genů jsou u RSEM normalizovány v metrikách TPM (viz 1.1.4 Normalizace) a FPKM (viz 1.1.4 Normalizace) a u softwaru IsoEM pomocí FPKM (viz 1.1.4 Normalizace). RSEM nevyžaduje existenci referenční genomové sekvence a ve spojení s *de novo* sestavování transkriptomu umožňuje analyzovat všechny druhy organismů (Li a Dewey 2011).

2.2.4 Normalizace počtu readů

Normalizace je jedním z nejdůležitějších kroků analýzy diferenciální genové exprese. Úroveň exprese jednotlivých genů může být zkreslena hloubkou sekvenování (Robinson a Oshlack 2010), velikostí sekvenační knihovny, délkou jednotlivých genů a obsahem GC párů (Risso et al. 2011). Cílem normalizace je, aby rozdíly v normalizovaných počtech readů odpovídaly skutečným biologickým rozdílům v expresi. Žádná nebo chybná normalizace může mít za následek nepřesné výsledky diferenciální genové exprese. K dispozici je několik normalizačních metod, které mají své předpoklady, při jejichž porušení dochází k chybným závěrům analýz (Evans et al. 2018).

Gen je diferenciálně exprimovaný (DE), pokud je úroveň jeho exprese závislá na různých podmínkách experimentu. Pomocí normalizace jsou vyrovnány úrovně exprese nediferenciálně exprimovaných genů (neDE) napříč experimenty. Velikost sekvenační knihovny a hloubka sekvenování jsou hlavními faktory ovlivňující srovnávací analýzy mezi vzorky. Při větší hloubce sekvenování je generováno větší množství readů. Bez provedení procesu normalizace by byly neDE geny ve vzorku s větším počtem readů označeny jako DE (Robinson a Oshlack 2010).

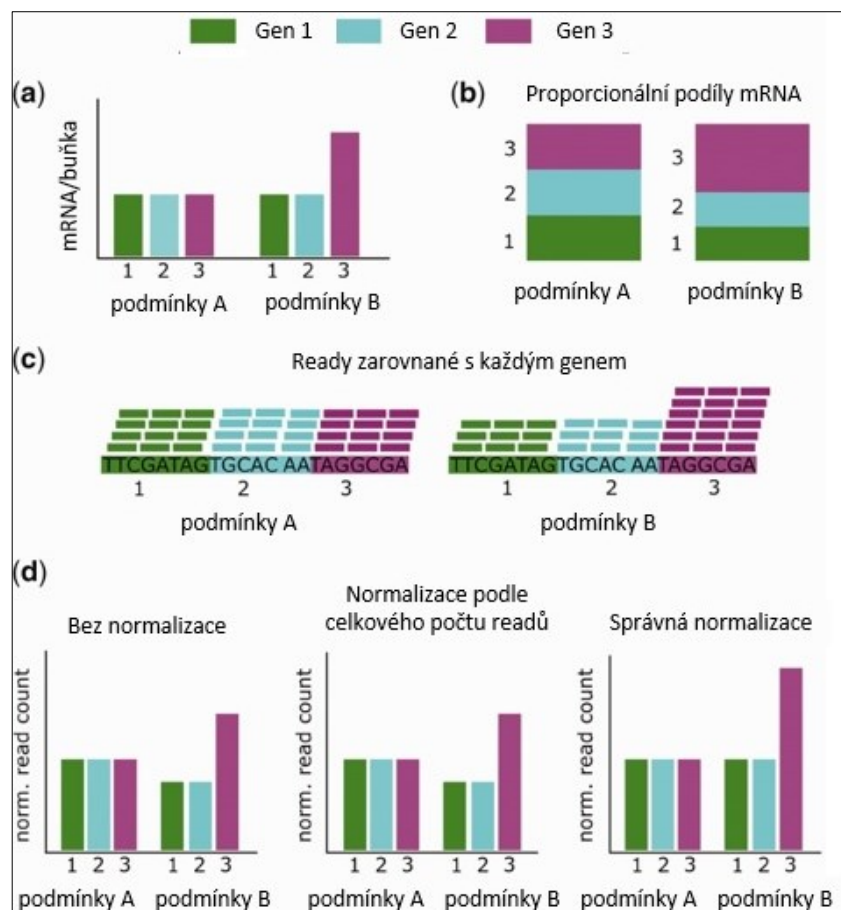
K dispozici je řada normalizačních metrik, které normalizují rozdíly v hloubkách sekvenování mezi vzorky (Tab. 4). CPM (counts per million) je metrika, která normalizuje namapované ready pouze na základě velikosti sekvenační knihovny, čímž umožňuje pouze srovnávání počtu genů mezi replikáty vzorků stejné skupiny (Robinson et al. 2010). Metrika

RPKM/FPKM (reads/fragments per kilobase of transcript per million) nejprve normalizuje namapované ready podle velikosti sekvenační knihovny a následně podle délky genů (Mortazavi et al. 2008). Obecně platí, že čím vyšší RPKM/FPKM genu, tím vyšší je exprese daného genu. Hodnoty této metriky se však mezi vzorky významně liší, a proto není vhodná pro srovnávání exprese genů napříč vzorky (Wagner et al. 2012). Proto byla navržena alternativní metrika TPM (transcripts per million), která nejprve provede normalizaci podle délky genu a podle těchto nových hodnot provede normalizaci podle velikosti sekvenační knihovny. Součet normalizovaných hodnot readů je pro všechny replikáty stejný, proto je možné tyto hodnoty porovnávat napříč vzorky.

Tab. 4: Příklady normalizačních metod normalizujících rozdíly v hloubkách sekvenování.

Normalizační metoda	Normalizace	Doporučené použití
CPM	Velikost sekvenační knihovny	Srovnávání počtu genů mezi replikáty stejné skupiny vzorků
RPKM/FPKM	Velikost sekvenační knihovny, délka genů	Srovnání exprese genů mezi geny ve vzorku
TPM	Délka genů, velikost sekvenační knihovny	Srovnání počtu genů ve vzorku a srovnání exprese genů ve vzorku a napříč vzorky

Kromě velikosti sekvenační knihovny a délky genů může mít na počty namapovaných readů vliv i procentuální zastoupení konkrétního genu v sekvenační knihovně, neboť sekvenační kapacita je omezená. Pokud je v jednom ze vzorků několik vysoce exprimovaných genů, tak procentuální zastoupení neDE genů bude nižší, než ve vzorku s menším množstvím vysoce exprimovaných genů (**Obr. 5**) (Robinson a Oshlack 2010). Tento problém lze normalizovat pomocí metod normalizace na základě distribuce/testem nebo pomocí kontrol (housekeeping geny, spike-in kontroly) (Evans et al. 2018). Počty readů jsou normalizovány pomocí vypočteného normalizačního faktoru. Mezi tyto metody se řadí například normalizace DESeq2 a TMM normalizace, která je součástí nástrojů pro diferenciální genovou expresi edgeR (Robinson et al. 2010) a limma+voom (Law et al. 2014).



Obr. 5: Ovlivnění diferenciální genové exprese vysoce upregulovaným genem (Evans et al. 2018, upraveno). (a) Gen 3 je za podmínek B dvakrát více exprimovaný (DE), zatímco geny 1 a 2 nejsou DE (neDE) (exprimují v obou podmínkách stejné množství mRNA); (b) všechny vzorky mají stejný počet osekvenovaných readů, distribuce jednotlivých genů v sekvenčních datech bude pro všechny geny za podmínek A 1/3, ale za podmínek B budou mít geny 1 a 2 procentuální zastoupení 1/4 a gen 3 bude zastoupen v 1/2 sekvenčních dat; (c) Tyto rozdíly v distribucích mezi neDE geny (1 a 2) se projeví při procesu mapování (za podmínek A se na gen 1 a 2 namapuje více readů než za podmínek B, přičemž jsou neDE); (d) bez provedení normalizace a i při provedení normalizace podle sekvenční knihovny jsou označeny všechny geny (1,2,3) jako DE, správná normalizační metoda normalizuje ready tak, aby neDE geny (1 a 2) měli po normalizaci stejné hodnoty exprese a gen 3 byl dvakrát tak exprimovaný za podmínek B.

2.2.5 Diferenciální genová exprese

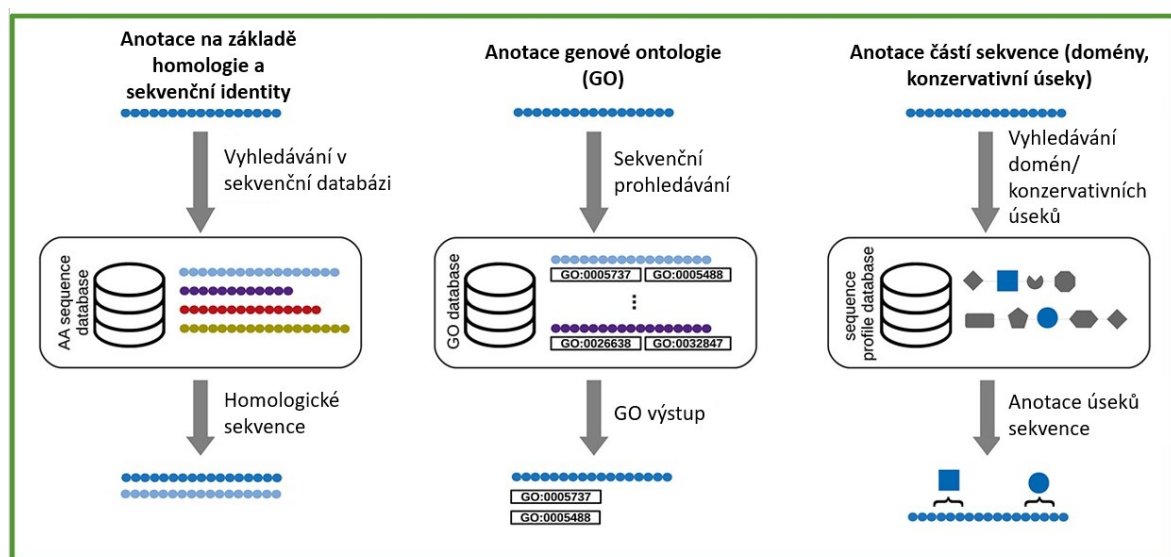
Analýza diferenciální genové exprese je proces, jehož cílem je nalezení genů, které mají statisticky významný rozdíl v expresi napříč různými podmínkami. Pomocí diferenciálně exprimovaných genů lze studovat příslušné fenotypy a genetické mechanismy v organismech. Diferenciálně exprimované geny jsou důležité například v rámci výzkumu vakcín, kde mohou být terčem terapeutických cílů. Nástroje pro diferenciální genovou expresi provádí statistické testy, které jsou založené na kvantifikaci exprese genů. Míra exprese je vyjádřena pomocí p-hodnoty, která vyjadřuje pravděpodobnost diferenciální genové exprese konkrétního genu v důsledku náhody. Velikost rozdílu exprese konkrétního

genu napříč experimenty je vyjádřena pomocí metriky log2foldchange. Jedná se o hodnotu log₂ poměru exprese konkrétního genu za různých podmínek.

Nástroje DESeq2 (Love et al. 2014) a edgeR (Robinson et al. 2010) patří k nejpoužívanějším softwarům určeným pro analýzu diferenciální genové exprese (určeno podle počtu citací). Software edgeR je nejlepší kompromis mezi přesností a rychlostí. Pokud je v experimentu nízký počet replikátů (<12), edgeR vykazuje nejlepší přesnost určení diferenciálně exprimovaných genů. Pro experimenty, kde je vyšší počet replikátů (> 12) se doporučuje software DESeq2 (Schurch et al. 2016). Při porovnání softwaru DESeq2 a softwaru limma+voom (Law et al. 2014) bylo zjištěno, že se více než 90 % detekovaných diferenciálně exprimovaných genů překrývalo mezi oběma metodami. Oba tyto softwary jsou spolehlivé a vhodné pro diferenciální analýzy (Tong 2021). Všechny zde zmíněné softwary jsou implementovány v rámci projektu Bioconductor (Online zdroj č. 7), takže se snáze instalují pomocí nástroje pro manipulaci s balíčky v programu R. Limma+voom byla vyhodnocena jako nejlepší software, co se týče rychlosti, obecnosti použití, dobré dokumentaci a robustnosti (Seyednasrollah et al. 2015).

2.2.6 Anotace transkriptomu

Anotace transkriptomu je proces přiřazování jména a/nebo biologické funkce předkládanému transkriptu. Vyžaduje integraci více biologických databází a několika výpočetních nástrojů. Proces funkční anotace transkriptů obvykle zahrnuje 3 kroky: sekvenční vyhledávání na základě homologie a sekvenční identity, anotace části sekvence (domény, konzervativní úseky) a anotace genové ontologie (GO) a biochemické dráhy (**Obr. 6**) (Raghavan et al. 2022). Tyto kroky nemusí být na sobě závislé a nemusí být implementovány ve všech anotačních softwarech. Kvalita anotace je závislá především na kvalitě a obsahu použitých databází. Při procesu anotace se doporučuje využívat více referenčních databází, které jsou zaměřené na konkrétní organismy (specializované), tak i databáze zahrnující velké množství organismů (univerzální) (Raghavan et al. 2022). Po provedení anotačního procesu je vhodné anotaci zveřejnit. Existují tři základní způsoby, jak zpřístupnit anotaci veřejnosti: vytvořit vlastní databázi genomu a umístit ji online, odeslat anotace do GenBank (Sayers et al. 2019) nebo Ensembl (Howe et al. 2021) nebo odeslat do kterékoliv z tematických databází genomů (Yandell a Ence 2012).



Obr. 6: Jednotlivé kroky funkční anotace transkriptomu (Raghavan et al. 2022, upraveno). Funkční anotace zahrnuje 3 kroky: anotaci na základě homologie a sekvenční identity, anotaci genové ontologie (GO) a anotaci částí sekvencí (domény, konzervativní úseky).

2.2.6.1 NCBI (The National Center for Biotechnology Information) databáze

NCBI je velká databáze biologických dat zahrnující i databázi citací PubMed a abstraktů publikovaných v přírodovědeckých časopisech (Sayers et al. 2021). Základním zdrojem dat je komplexní veřejná databáze GenBank (Sayers et al. 2019), která obsahuje více než 6,25 bilionů párů bází z více než 1,6 miliardy nukleotidových sekvencí pro 450 000 formálně popsaných druhů (Sayers et al. 2019). Vyhledávání v databázích v rámci NCBI zajišťuje integrovaný systém Entrez (Schuler et al. 1996). Jedná se o molekulárně-biologický databázový a vyhledávací systém, který poskytuje přístup k 34 databázím (odkazy na jednotlivé databáze jsou uvedeny na stránce globálního vyhledávání Entrez, Online zdroj č. 8), které dohromady obsahují 3 miliardy záznamů (Sayers et al. 2021).

Proces vyhledávání v sekvenční databázi je umožněn pomocí integrovaného programu BLAST (Basic Local Search Alignment) (Boratyn et al. 2013), který na základě vstupní sekvence provádí vyhledávání podobných sekvencí z datových souborů (Sayers et al. 2021). Nicméně rychlost tohoto softwaru není optimální pro dlouhé dotazovací nebo databázové sekvence. Proto byla vyvinuta aplikace příkazového řádku BLAST+ (Camacho et al. 2009), která tyto sekvence rozdělí do bloků, což výrazně zkrátí dobu vyhledávání (Camacho et al. 2009).

Výsledkem vyhledávání jsou názvy sekvencí seřazené podle statistické významnosti nalezené shody (bit score a e-value). Bit score vyjadřuje statistickou významnost alignmentu

vstupní sekvence se sekvencí v databázi. Čím vyšší hodnota, tím větší podobnost těchto sekvencí. E-value je informativnější parametr, který navíc zahrnuje velikost prohledávané databáze. Vyjadřuje pravděpodobnost, že náhodná sekvence v databázi bude mít stejnou nebo vyšší podobnost ke vstupní sekvenci než nalezená sekvence v databázi dané velikosti. E-value bývá obvykle použita jako první filtr kvality alignmentu pro výsledek hledání pomocí BLAST (blastn – nukleotidová sekvence proti nukleotidové databázi, blastp – proteinová sekvence proti proteinové databázi, blastx – nukleotidová sekvence proti proteinové databázi, tblastn – proteinová sekvence proti nukleotidové databázi). Čím nižší hodnota e-value tím lepší shoda (Online zdroj č. 9).

Kromě názvu sekvence je důležité znát další informace například o funkci proteinu, zařazení do proteinových skupin a metabolických drah. Pro přiřazení funkční anotace jednotlivým sekvencím byl vytvořen nástroj InterproScan (Jones et al. 2014), jehož výstup zahrnuje informace z databáze InterPro (Blum et al. 2021). Databáze InterPro sjednocuje informace o proteinových rodinách, doménách a funkčních místech proteinů ze 13 databází, které klasifikují proteiny pomocí řady různých metod (Online zdroj č. 10). InterproScan klasifikuje části vstupní sekvence na základě podobnosti v databázi InterPro (domény, funkční místa). Každá shoda se záznamem v databázi obsahuje hypertextový odkaz na hlavní zdroj databáze InterPro a na příslušnou databázi (Quevillon et al. 2005).

2.2.6.2 KEGG (*Kyoto Encyclopedia of Genes and Genomes*) databáze

Databáze KEGG je japonská databáze, která byla vytvořena k přiřazování funkčních významů proteinů na molekulární úrovni (sítě molekulárních interakcí, reakcí a vztahů). V současné době KEGG obsahuje 18 správce spravovaných databází, které jsou rozděleny do 4 funkčních kategorií, kde každá databáze má svůj specifický identifikátor (Kanehisa et al. 2021) (**Obr. 7**). Součástí je i databáze KEGG Orthology (KO), kde jsou uloženy biologické funkce. Každý záznam KO je definován jako funkční ortolog genů a proteinů.

Category	Database name	Content	Prefix
Systems Information	KEGG PATHWAY	KEGG pathway maps	map, hsa, etc.
	KEGG BRITE	BRITE functional hierarchies and tables	br, ko, etc.
	KEGG MODULE	KEGG modules Reaction modules	M RM
Genomic Information	KEGG ORTHOLOGY (KO)	KO groups for functional orthologs	K
	KEGG GENOME	KEGG organisms and viruses	T
	KEGG GENES	Genes and proteins	
	KEGG SSDB	Sequence similarity among GENES entries	
Chemical Information	KEGG COMPOUND	Metabolites and other small molecules	C
	KEGG GLYCAN	Glycans	G
	KEGG REACTION	Biochemical reactions	R
	KEGG RCLASS	Reaction class	RC
	KEGG ENZYME	Enzyme nomenclature	
Health Information	KEGG NETWORK	Disease-related network elements Network variation maps	N nt
	KEGG VARIANT	Human gene variants	
	KEGG DISEASE	Human diseases	H
	KEGG DRUG	Drugs	D
	KEGG DGROU	Drug groups	DG
	KEGG ENVIRON	Crude drugs and health-related substances	E
	JAPIC	Japanese drug labels	
	DailyMed	FDA drug labels (links only)	

Obr. 7: Seznam 18 správce spravovaných databází, které jsou součástí KEGG databáze (Kanehisa et al. 2021). KEGG databáze se skládá z 18 správce spravovaných databází, které jsou rozděleny do 4 kategorií. Každá databáze má svůj specifický KEGG identifikátor (Prefix), za kterým následuje pětimístné číslo.

Většina sekvenčních dat v KEGG databázi je uložena v kategorii KEGG organismy (KEGG organisms) a KEGG viry (KEGG viruses). Kategorie KEGG organismy je hlavní částí integrované databáze GENES, která obsahuje osekvenované genomy z databáze RefSeq (O’Leary et al. 2016) a GenBank (Clark et al. 2016). Celkem obsahuje 759 zástupců eukaryot, 7027 zástupců bakterií a 389 zástupců archeí (k červnu 2022). Každý KEGG organismus má identifikátor, který je reprezentován tří/čtyřpísmenným kódem (např. „eco“ pro *Escherichia coli*). Kategorie KEGG viry obsahuje data z RefSeq (O’Leary et al. 2016), která je identifikována dvoupísmenným kódem ‘vg’ (vg:geneid) a nejsou zde rozlišeny jednotlivé virové genomy (Kanehisa et al. 2017). V současné době obsahuje 359 druhů virových genomů (červen 2022).

V rámci KEGG databáze je k dispozici řada softwarů, jako např. KEGG Mapper tools, které mapují proti databázím KEGG PATHWAY (grafické mapy představující např. molekulární dráhy pro metabolismus, zpracování genetických informací, další buněčné procesy, lidské nemoci a vývoj léků), KEGG BRITE (hierarchické klasifikační systémy popisující zařazení proteinů do proteinových skupin) a KEGG MODULE (sady

genů charakterizující metabolické dráhy, fenotypové znaky a reakční kroky v metabolických drahách). Tyto nástroje jsou sloučeny do 3 obecných mapovacích nástrojů: Reconstruct Pathway, Search Pathway a Search & Color Pathway, kde každý nástroj umožňuje mapování proti více KEGG databázím najednou (Kanehisa a Sato 2020). Cílem těchto nástrojů je propojení molekulárních objektů (genů, proteinů, metabolitů a glykanů) s objekty vyšší biologické úrovně (cesty, moduly, hierarchie, taxonomie a nemoci).

Nejedná se o veřejnou databázi, proto je zapotřebí si zakoupit licenci. Licencovaná verze KEGG databáze umožňuje přístup k KEGG FTP serveru, který umožňuje hromadné stahování dat, což výrazně zrychlí práci. Pro akademické pracovníky jsou volně zpřístupněny KEGG webové stránky a za poplatek přístupová práva k KEGG FTP. Pro akademické pracovníky mimo Japonsko je cena této licence 2 000 USD na rok pro jednoho koncového uživatele a 5 000 USD pro více koncových uživatelů (červen 2022).

2.2.7 Analýza obohacení genové sady (Gene enrichment analysis)

Analýza obohacení genové sady slouží k nalezení statisticky významně upregulovaných či downregulovaných biologických drah či proteinových rodin. Diferenciální analýza genové exprese identifikuje pouze diferencially exprimované geny/transkripty, ale nikoliv rozdíly v biologických procesech či v proteinových rodinách, které jsou typicky tvořené více proteiny. Je velmi obtížné určit geny, jejichž proteinové produkty spolu vzájemně interagují a jak ovlivňují zkoumané biologické procesy (Tipney a Hunter 2010).

Jedna z možných metod výpočtu genového obohacení se snaží identifikovat statisticky významně obohacené biologické dráhy nebo proteinové rodiny (upregulované, downregulované) pomocí Fisherova exaktního testu (Fisher 1935), ze kterého jsou získány p-hodnoty. Tyto p-hodnoty vyjadřují nejmenší hladinu významnosti testu, při níž zamítneme nulovou hypotézu (daná biologická dráha není statisticky významně genově obohacena) na zkoumaných datech. Čím nižší p-hodnota, tím nižší pravděpodobnost platnosti nulové hypotézy. K výpočtu Fisherova exaktního testu je zapotřebí vytvořit kontingenční tabulku (Tab. 5), která přehledně zobrazuje vzájemné vztahy mezi zkoumanými daty. Z takto vytvořené tabulky je pomocí následujícího vzorce vypočtena výsledná p-hodnota:

$$p = \frac{\binom{k+(n-k)}{k} \binom{(m-k)+(N+k-n-m)}{(m-k)}}{\binom{N}{k+(m-k)}} =$$

$$= \frac{(k + (m - k))! (k + (n - k))! ((n - k) + (N + k - n - m))! ((m - k) + (N + k - n - m))!}{N! k! (m - k)! (n - k)! (N + k - n - m)!}$$

Tab. 5: **Kontingenční tabulka 2x2** (Hung 2013, upraveno). Jednotlivé proměnné reprezentují: k – počet up/downregulovaných genů v daném vzorku v konkrétní biologické dráze, m – celkový počet up/downregulovaných genů v daném vzorku v konkrétní biologické dráze, n – celkový počet up/downregulovaných genů v daném vzorku, N – celkový počet up/downregulovaných genů v daném vzorku.

	UP/DOWN regulované geny	neUP/neDOWN regulované geny	Celkový počet genů
Geny v konkrétní biologické dráze	k	$m-k$	m
Všechny geny, které nejsou v konkrétní biologické dráze	$n-k$	$N+k-n-m$	$N-m$
Celkový počet genů	n	$N-n$	N

Jednotlivé biologické dráhy jsou označeny za statisticky významně obohacené, pokud výsledná p-hodnota je nižší nebo rovna zvolené hladině významnosti testu (většinou 0,05, což je 5 % šance, že dané obohacení je způsobeno náhodou) (Huang et al. 2009a).

Tento test je však proveden na všechny biologické dráhy, kde se vyskytují up/downregulované geny zkoumaného vzorku. Při zvolení hladiny významnosti 0,05, tak při provedení 100 testů bude průměrně 5 testů špatně (5 % pravděpodobnost chyby 1. stupně). Proto je nutné řešit problém násobného testování hypotéz. K dispozici je několik typů korekčních metod např. Bonferroniho korekce (Bland a Altman 1995), která upravuje hladinu významnosti testu tak, že ji dělí počtem provedených testů (např. 0,05/100 při provedení 100 testů na hladině významnosti 0,05). Nicméně tato korekce je velmi přísná a při větším počtu provedených testů je relativně těžké dosáhnout statistické významnosti. Korekčních metod je velké množství např. Tukeyova korekce, Benjamini-Hochbergova korekce, Holmova korekce a další (de Leeuw et al. 2016).

Pro výpočet genového obohacení je k dispozici řada nástrojů, nicméně ty fungují jen pro vybrané organismy. Software DAVID (Huang et al. 2009b) má v sobě integrované databáze a řadu analytických nástrojů. Na vstupu přijímá seznam zkoumaných genů

s libovolným počtem společných genových identifikátorů (Huang et al. 2009b). V nově aktualizované verzi pokrývá 55 464 organismů a je v současnosti nejpoužívanějším nástrojem (Sherman et al. 2022). Další software BiNGO (Maere et al. 2005) na vstupu přijímá buď seznam genů vložený jako text, nebo podgrafy biologických sítí vizualizované v softwaru Cytoscape (Shannon et al. 2003). GOrilla (Eden et al. 2009) na rozdíl od ostatních nástrojů zahrnuje přesnější statistické výpočty, efektivní grafické znázornění a dobrou časovou složitost (Eden et al. 2009). Nástroj Blast2GO (Gotz et al. 2008) kombinuje různé anotační strategie a nástroje související s anotací a obsahuje grafické funkce (GO-graf na výstupu) (Gotz et al. 2008).

3 Materiál a metody

3.1 Sběr materiálu, izolace RNA a sekvenace

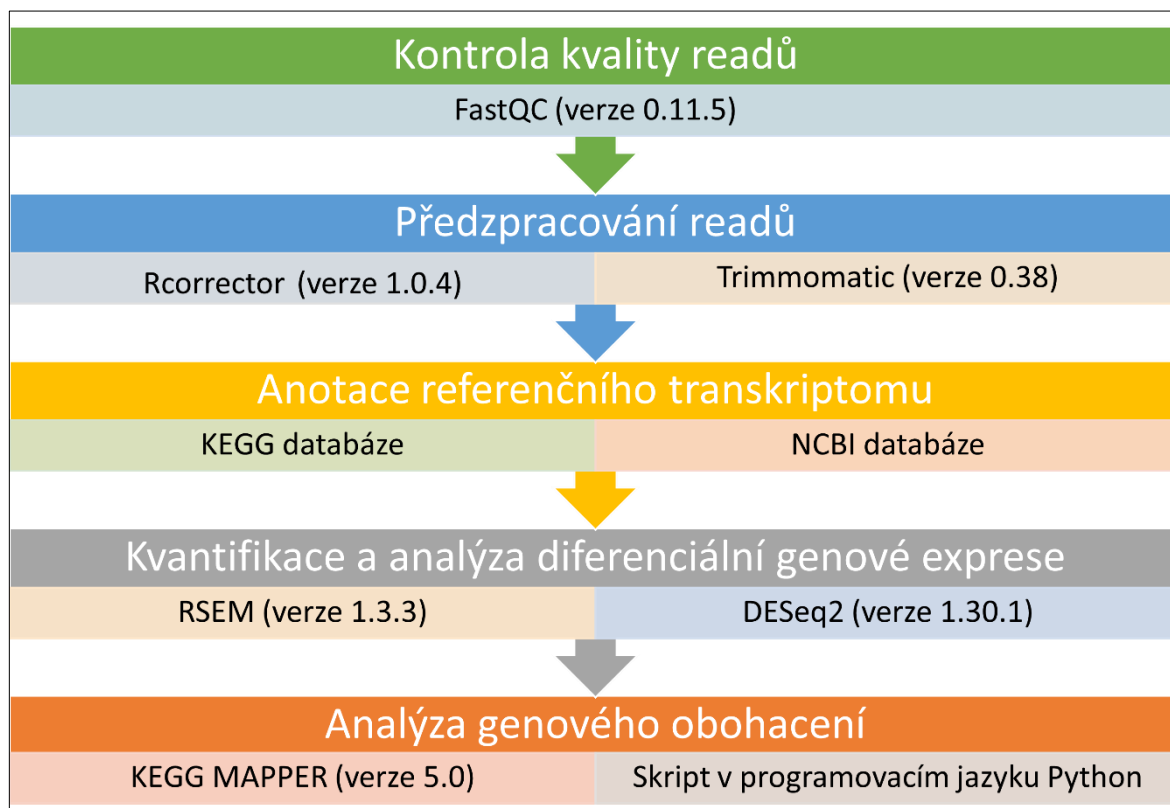
Tetrathyridia tasemnice *Mesocestoides corti* (syn. *M. vogae*) byly získány nákazou outbredního kmene myši Hsd:ICR (CD-1) (ICR) a inbredního kmene C57BL/6OlaHsdd (BLACK) (B). Každá myš byla infikována šesti sty larvami a nákaza trvala po dobu 14 dnů. Po 14 dnech byly tasemnice (n=1000) vypitvány z peritoneální dutiny a 3x promyty sterilním fyziologickým roztokem. Po krátké centrifugaci byl odstraněn zbytek fyziologického roztoku a vzorky byly stabilizovány TRIzolem (Invitrogen) a uloženy v -80°C.

Dále byla provedena *in vitro* (IV) kultivace, kde byla použita tetrathyridia vypitvané z myšního kmene BALB/c. Tasemnice byly kultivovány po dobu 14 dnů v DMEM mediu s přidáním penicilin-streptomycinem (100 U/ml). Výměna média probíhala každé dva dny.

Izolace total RNA ze všech vzorků proběhla podle protokolu TRIzol RNA Isolation Reagent (Invitrogen). Reziduální DNA byla odstraněna pomocí TURBO DNA-freeKit (Invitrogen). Kvalita a integrita RNA byla měřena přístrojem Bioanalyzer 2100 (Agilent). Knihovny pro sekvenaci byly konstruovány pomocí protokolu NEBNext® Poly(A) mRNA Magnetic Isolation Module (Online zdroj č. 11) a NEBNext® Ultra II Directional RNA Library Prep Kit for Illumina (Online zdroj č. 12) a sekvenace proběhla na platformě NovaSeq (Illumina) pomocí pair-end sekvenování. Každý vzorek byl analyzován ve 4 biologických replikátech, celkově tedy bylo osekvenováno 12 transkriptomů *M. corti*.

3.2 Bioinformatická analýza

Výpočetní kapacita byla využita v rámci infrastruktury Metacentrum. Bioinformatická analýza zahrnovala několik kroků, kde v každém kroku byl zvolen vhodný software, nebo napsán skript v programovacím jazyku Python (**Obr. 8**). Tyto kroky jsou popsány níže.



Obr. 8: Schéma bioinformatické analýzy s použitými softwary. Nejprve byla provedena kontrola kvality vstupních readů pomocí softwaru FastQC (v. 0.11.5). Dále bylo provedeno předzpracování readů, které zahrnovalo korekce sekvenčních chyb pomocí softwaru Rcorrector (v. 1.0.4) a odstranění nekvalitních konců readů pomocí softwaru Trimmomatic (v. 0.38). Následoval krok anotace referenčního transkriptu pomocí KEGG a NCBI databáze. Následně byla provedena kvantifikace genové exprese pomocí softwaru RSEM (verze 1.3.3) a analýza diferenciální genové exprese pomocí softwaru DESeq2 (verze 1.30.1). Na konec byla provedena analýza genového obohacení, kde na základě mapování transkriptů do biologických drah (KEGG Mapper verze 5.0) byly zjištěné statisticky významně upregulované biologické dráhy. Pro tuto analýzu byl vytvořen skript v programovacím jazyku Pythonu (Přílohy *gene_enrichment.py*).

3.2.1 Sekvenační data, předzpracování dat a kontrola kvality

U dat získaných ze sekvenátoru byla nejprve provedena kontrola kvality pomocí softwaru FastQC (verze 0.11.5) (Online zdroj č. 3) s defaultním nastavením. Následně byla provedena korekce sekvenčních chyb pomocí softwaru Rcorrector (verze 1.0.4) (Song a Florea 2015) s defaultním nastavením. Pomocí softwaru Trimmomatic (verze 0.38) (Bolger et al. 2014) byly odstraněny nekvalitní konce readů. Software byl spuštěn s následujícím nastavením parametrů: PE; SLIDINGWINDOW:4:5; LEADING:5; TRAILING:5;

MINLEN:25. Následně byla znovu provedena kontrola kvality pomocí softwaru FastQC (verze 0.11.5) (Online zdroj č. 3).

3.2.2 Anotace referenčního transkriptomu

Anotace referenčního transkriptomu *M. corti* PRJEB510.WBPS15 (Online zdroj č. 1) (kapitola 2.1.1) byla provedena pomocí dvou různých databází: KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa et al. 2021) a NCBI (Sayers et al. 2021). Referenční transkriptom byl anotován pomocí automatického anotačního serveru GhostKoala (verze 2.2) (Kanehisa et al. 2016) s využitím prokaryotické a eukaryotické databáze genů. Anotace pomocí NCBI databáze (verze březen 2020) byla provedena pomocí softwaru blast-plus (verze 2.10.1) (Camacho et al. 2009) s následujícími parametry: -db nr; -evalue 1e-05; -max_target_seqs 5.

Výstupní soubor NCBI anotovaného transkriptomu obsahoval ke každému transkriptu 5 různých anotací. U každého transkriptu byla vybrána jedna konkrétní anotace primárně podle funkční anotace genu a sekundárně podle nejvyšší hodnoty bit_score (udává kvalitu zarovnání). Na tuto filtraci byl vytvořen skript v programovacím jazyku Python (Přílohy *filtering_ncbi.py*). Pro výsledné anotace transkriptomů jednotlivých vzorků byly vybrány transkripty, jejichž hodnota „expected counts“ byla vyšší či rovna 10 ve všech replikátech (Přílohy *transcriptoms_annotation.py*), ale pro diferenciální genovou analýzu a analýzu genového obohacení byly využity všechny transkripty v každém replikátu.

3.2.3 Kvantifikace a analýza diferenciální genové exprese

Očištěné a anotované ready byly mapovány na referenční anotovaný transkriptom a byla provedena kvantifikace genové exprese pomocí softwaru RSEM (verze 1.3.3) (Li a Dewey 2011). Mapování a kvantifikace byla provedena pro každý vzorek analýzy zvlášť. Detekce odlišně exprimovaných genů byla zjištěna pomocí softwaru DESeq2 (1.30.1) (Love et al. 2014) pro každou dvojici vzorků (IV&B, IV&ICR, ICR&B).

3.2.4 Sloučení výstupů a analýza obohacení genové sady

Jednotlivé výstupy ze softwaru DESeq2 (tabulky s diferenciálně exprimovanými transkripty) byly sloučeny s KEGG a NCBI anotacemi na základě unikátního identifikátoru

každého transkriptu. Jednotlivé transkripty s hodnotou padj. menší než 0,5 byly rozděleny podle hodnoty $\log_2\text{foldchange}$ do několika kategorií od mírně up/down regulovaných až po velmi up/down regulované (Tab. 6). Pro analýzu genového obohacení byly tyto kategorie nekriticky sloučeny do dvou souborů: upregulované $<0, \infty$) a downregulované $(-\infty, 0)$ vzhledem k Vzorku1. Pro tyto kroky byl vytvořen skript v jazyku Python (Přílohy *diffexpr_annotation.py*).

Tab. 6: Rozdělení diferenciálně exprimovaných transkriptů podle hodnoty $\log_2\text{foldchange}$.

Hodnota $\log_2\text{foldchange}$	Vzorek 1	Vzorek 2
< -2	Vysoce downregulované	Vysoce upregulované
$[-2, -1]$	Downregulované	Upregulované
$[-1, 0]$	Mírně downregulované	Mírně upregulované
$[0, 1]$	Mírně upregulované	Mírně downregulované
$[1, 2]$	Upregulované	Downregulované
> 2	Vysoce upregulované	Vysoce downregulované

Pomocí nástroje KEGG Mapper (verze 5.0) (Kanehisa a Sato 2020) bylo provedeno mapování up/downregulovaných transkriptů na konkrétní metabolické dráhy. Na výsledky tohoto mapování byla provedena analýza genového obohacení na základě Fisherova exaktního testu pomocí vytvořeného skriptu v jazyku Python (Přílohy *gene_enrichment.py*). KEGG biologické dráhy byly označeny za obohacené, pokud jejich p -hodnota (získaná z Fisherova exaktního testu) byla nižší nebo rovna 0,05 (hladina významnosti testu).

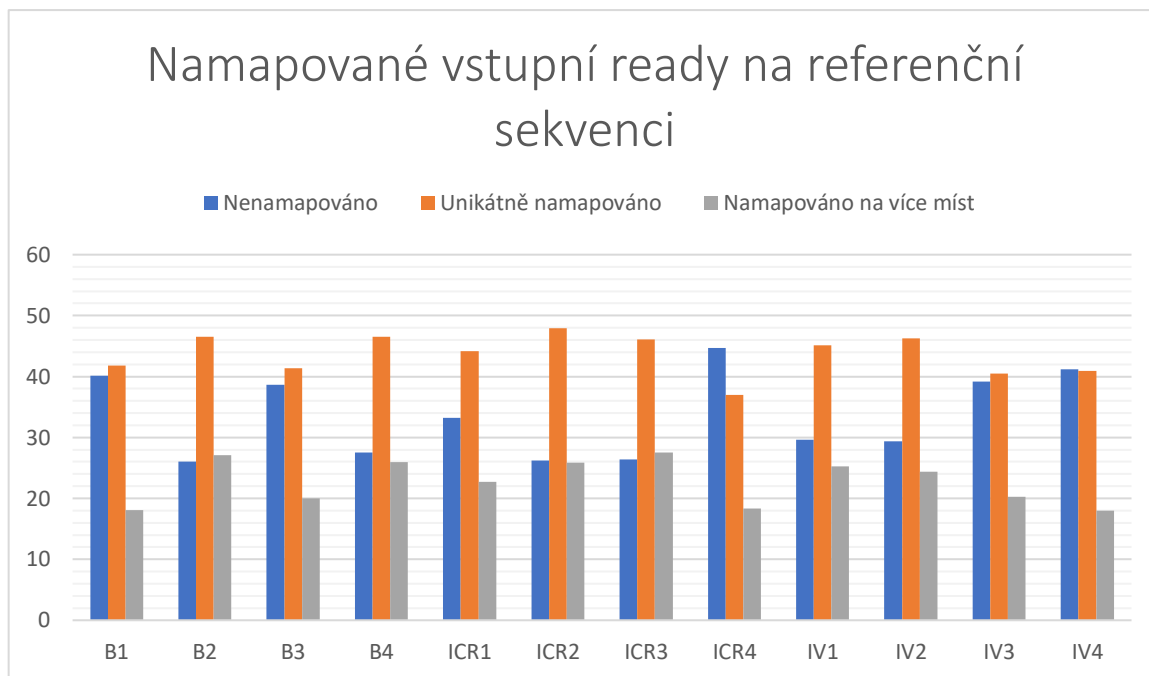
4 Výsledky

4.1 Izolace RNA a sekvenační data *M. corti*

RNA pro tvorbu knihoven pro sekvenaci byla získána z tetrathyridií *M. corti*, které byly kultivovány *in vitro* (IV) a v myších hostitelích inbredního druhu C57BL/6OlaHsd (BLACK) (B) a outbredního druhu Hsd:ICR (CD-1) (ICR). Každý vzorek byl tvořen čtyřmi replikáty. Izolovaná RNA byla podrobena kontrole kvality pomocí mikrokapilární elektroforézy na přístroji Bioanalyzer s následujícími RIN (RNA integrity number) hodnotami: 8,60 (B1); 9,50 (B2); 9,20 (B3); 9,50 (B4); 8,50 (IV1); 7,40 (IV2); 9,30 (IV3); 9,20 (IV4); 9,20 (ICR1); 9,30 (ICR2); 9,50 (ICR3) a 8,90 (ICR4). Po sekvenaci bylo celkem získáno 337,9 milionů vysoce kvalitních (otrimovaných) pair-end readů se srovnatelným zastoupením pro vzorky získané z myší (B – 117,9 milionů; ICR – 116,1 milionů). Vzorky z *in vitro* kultivace byly zastoupeny menším počtem readů (103,9 milionů). Procento readů namapovaných na referenční transkriptom *M. corti* bylo u replikátů B 59,85–73,63 %, u replikátů ICR 55,32–73,82 % a u replikátů IV 58,86–70,59 % (Tab. 7). Z toho unikátně namapovaných readů bylo 41,35–46,54 % u B, 36,95–47,93 % u ICR a 40,5–46,25 % u IV (Obr. 9).

Tab. 7: Informace o sekvenačních datech tetrathyridií *Mesocostoides corti* kultivované v myších hostitelích inbredního druhu BLACK (B), outbredního druhu Hsd:ICR (CD-1) (ICR) a *in vitro* (IV).

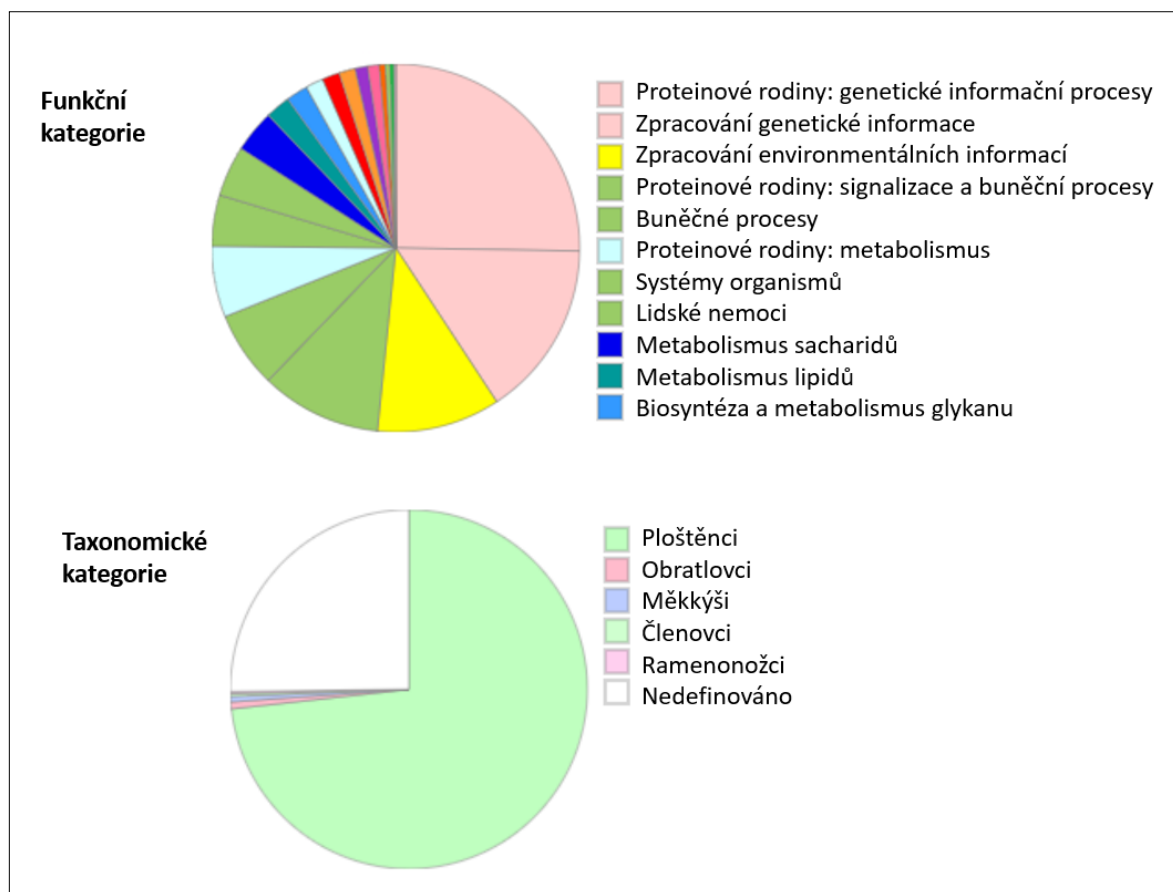
Název vzorku	Počet pair-end readů	Počet otrimovaných readů	Počet namapovaných readů (v %)	Počet unikátně namapovaných readů (v %)
B1	34 229 424	33 123 094	19 822 032 (59,85)	13 837 344 (41,78)
B2	33 971 130	33 183 640	24 434 480 (73,63)	15 444 252 (46,54)
B3	22 513 708	21 935 170	13 461 980 (61,38)	9 069 392 (41,35)
B4	30 352 722	29 655 130	21 484 780 (72,45)	13 796 532 (46,52)
ICR1	29 003 682	28 263 210	18 885 526 (66,83)	12 476 828 (44,15)
ICR2	32 673 958	31 739 626	23 429 216 (73,82)	15 211 658 (47,93)
ICR3	30 410 378	29 503 534	21 744 344 (73,63)	13 617 886 (46,11)
ICR4	27 349 016	26 637 484	14 734 968 (55,32)	9 842 900 (36,95)
IV1	25 916 924	25 325 050	17 820 674 (70,37)	11 426 902 (45,12)
IV2	26 967 034	26 341 514	18 595 438 (70,59)	12 183 828 (46,25)
IV3	24 685 180	23 729 776	14 424 192 (60,79)	9 611 154 (40,50)
IV4	29 207 524	28 482 224	16 764 096 (58,86)	11 642 664 (40,88)



Obr. 9: Sloupcový graf reprezentující výsledky mapování (v %) vstupních readů u všech replikátů každého vzorku.

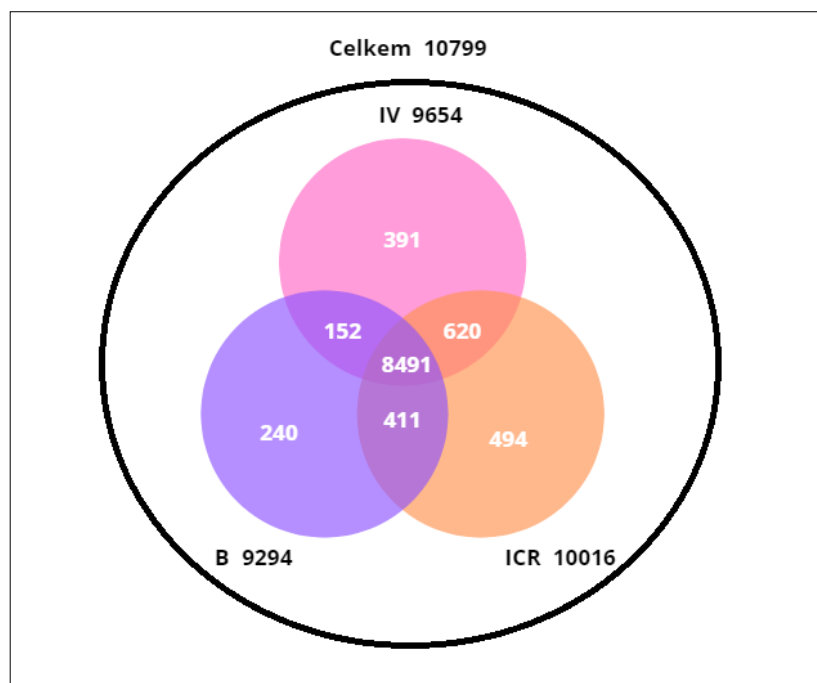
4.2 Anotace referenčního transkriptomu

Referenční transkriptom obsahuje 22 265 transkriptů. Pomocí NCBI databáze bylo anotováno 20 333 transkriptů (~ 91,3 %) a pomocí KEGG databáze 10 214 transkriptů (~ 45,9 %). Téměř polovina anotovaných transkriptů je podle KEGG databáze součástí proteinové rodiny genetických informačních procesů. Většina transkriptů byla z taxonomické kategorie ploštěnci a zhruba čtvrtina transkriptů nebyla přiřazena k žádné známé taxonomické kategorii (Obr. 10).



Obr. 10: Anotace referenčního transkriptomu pomocí KEGG databáze (Kanehisa et al. 2021, upraveno).

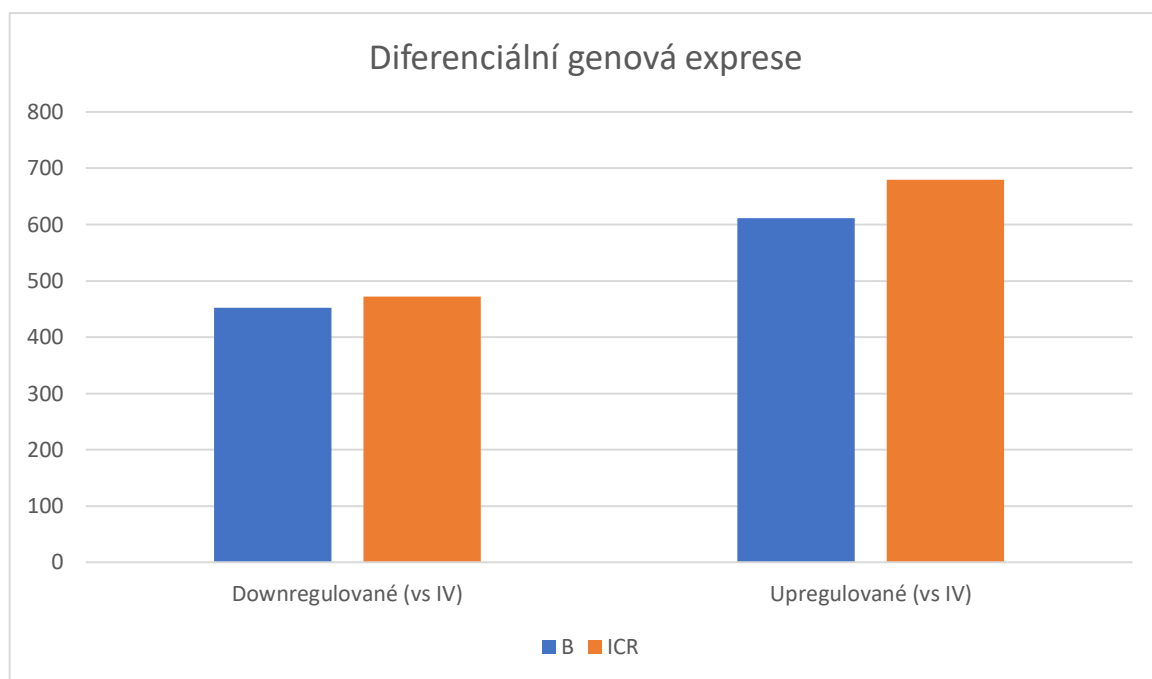
Z celkového počtu 22 265 referenčních transkriptů zkoumané vzorky exprimovaly 10 799 transkriptů, z čehož 8 491 sdílely všechny zkoumané replikáty. Nejméně specifických transkriptů bylo ve vzorcích získaných z inbredních myší (B – 240). Zatímco ve vzorcích získaných z outbredních myší (ICR – 494) a z *in vitro* kultivace (IV – 391) bylo množství specifických transkriptů v srovnatelném zastoupení (**Obr. 11**). V rámci práce byl pro každý vzorek (B, ICR, IV) sestaven anotovaný transkriptom, který obsahoval všechny transkripty, které měly hodnotu „expected counts“ větší nebo rovnou 10 (Přílohy S1).



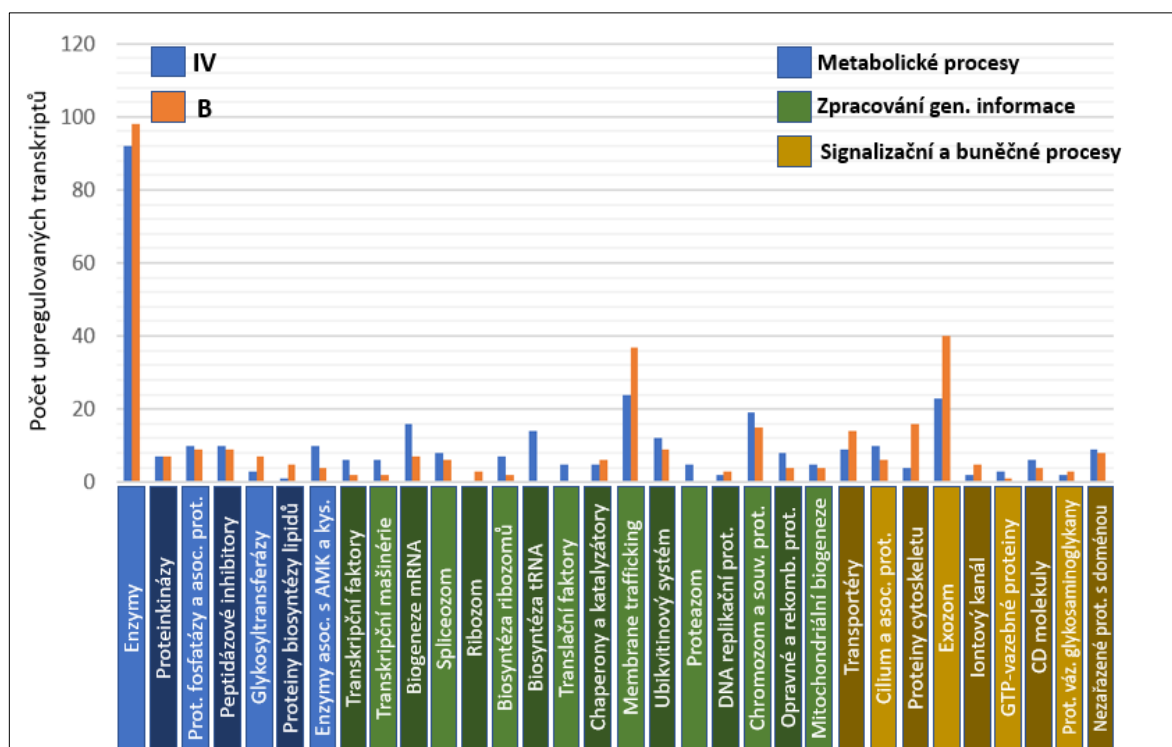
Obr. 11: Vennův diagram zobrazující srovnání exprimovaných transkriptů u vzorků B, ICR a IV. Graf obsahuje počty transkriptů, které měly hodnotu „expected counts“ větší nebo rovnou 10 ve všech replikátech.

4.3 Kvantifikace a analýza diferenciální exprese

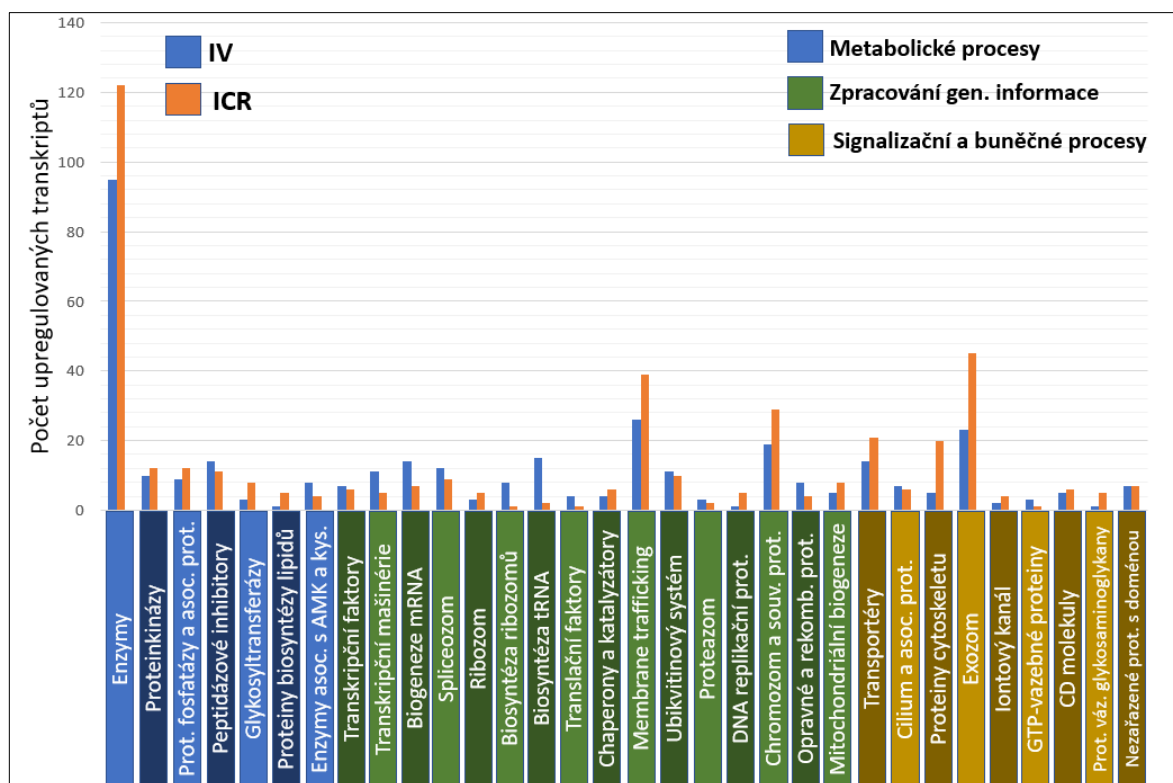
Analýza diferenciální genové exprese byla provedena pro každou kombinaci dvojic druhů vzorků (B & IV, ICR & IV, B & ICR) (**Obr. 12**) (Přílohy **S2**). Při vzájemných porovnáních bylo upregulováno 611:452 (B: IV), 680:472 (ICR: IV) a 9:17 (B: ICR) transkriptů. Kvůli nízkému počtu diferenciálně exprimovaných genů, není provedena u srovnání B & ICR analýza genového obohacení. Nejvíce upregulovaných transkriptů ve všech porovnávaných replikátech je podle KEGG BRITE databáze enzymatického charakteru (**Obr. 13 a 14**). Největší rozdíly mezi upregulovanými transkripty u tasemnic pocházejících z myších hostitelů (B, ICR) a *in vitro* kultivovanými tasemnicemi (IV) byly v kategoriích enzymy, membrane trafficking a v proteinech souvisejících s exozomem. Naopak více upregulovaných transkriptů v IV (vs B a ICR) bylo ve většině proteinů souvisejících s translací a transkripcí (kategorie zpracování genetické informace). Deset nejvíce upregulovaných transkriptů u replikátů B a ICR ve srovnání s IV replikáty kódovalo chromodomain helicase DNA binding protein 1, glioma pathogenesis protein 1, cysteine-rich secreted protein 2 precursor, DH domain containing protein a unnamed proteins (*Tab. 8, Tab. 9*).



Obr. 12: Upregulované a downregulované transkripty replikátů B a ICR ve srovnání s IV replikáty.



Obr. 13: Sloupcové grafy znázorňující upregulované transkripty ve srovnání IV a B replikátů (Kyoto Encyclopedia of Genes and Genomes (KEGG), BRITe anotace).



Obr. 14: Sloupcové grafy znázorňující upregulované transkripty ve srovnání IV a ICR replikátů (Kyoto Encyclopedia of Genes and Genomes (KEGG), BRITE anotace).

Tab. 8: Deset nejvíce upregulovaných transkriptů v B replikátech ve srovnání s IV replikáty.

ID transkriptu	Expected counts (4 replikátů)				Log2fold change	NCBI anotace
MCU_005129-RA	0	678	0	383	-24,8	chromodomain helicase DNA binding protein 1 (<i>Echinococcus multilocularis</i>)
MCU_012543-RA	229	302	370	162	-10,5	glioma pathogenesis protein 1 (<i>Echinococcus multilocularis</i>)
MCU_004110-RA	1755	2475	1165	2172	-9,7	cysteine-rich secreted protein 2 precursor (<i>Mesocestoides vogae</i>)
MCU_006924-RD	329	0	0	122	-9,2	Immunoglobulin E set (<i>Echinococcus granulosus</i>)
MCU_000888-RC	264	18	105	17	-9,1	Transducin beta-like protein 3 (<i>Echinococcus granulosus</i>)
MCU_012433-RA	73	88	74	110	-8,8	cysteine-rich secreted protein 2 precursor (<i>Mesocestoides vogae</i>)
MCU_012577-RB	156	44	47	41	-8,6	unnamed protein product (<i>Mesocestoides corti</i>)
MCU_000072-RE	234	18	0	23	-8,5	active breakpoint cluster region (<i>Echinococcus multilocularis</i>)
MCU_006389-RA	83	28	86	0	-8,0	DH domain containing protein (<i>Echinococcus multilocularis</i>)
MCU_014000-RA	51	17	67	35	-7,8	cysteine-rich secreted protein 2 precursor (<i>Mesocestoides vogae</i>)

Tab. 9: Deset nejvíce upregulovaných transkriptů v ICR replikátech ve srovnání s IV replikáty.

ID transkriptu	Expected counts (4 replikátů)				Log2fold change	NCBI anotace
MCU_012543-RA	351	423	511	686	-11,3	glioma pathogenesis protein 1 (<i>Echinococcus multilocularis</i>)
MCU_005129-RA	41	420	625	0	-10,4	chromodomain helicase DNA binding protein 1 (<i>Echinococcus multilocularis</i>)
MCU_012433-RA	59	69	203	354	-9,8	cysteine-rich secreted protein 2 precursor (<i>Mesocestoides vogae</i>)
MCU_000297-RB	39	275	145	0	-9,2	oligomeric Golgi complex subunit 8 (<i>Echinococcus multilocularis</i>)
MCU_004908-RB	36	361	0	38	-9,1	sec1 family domain containing protein 1 (<i>Echinococcus multilocularis</i>)
MCU_010279-RA	95	38	266	103	-8,6	unnamed protein product (<i>Hydatigera taeniaformis</i>)
MCU_006984-RD	11	54	31	117	-8,1	titin (<i>Echinococcus multilocularis</i>)
MCU_011868-RA	421	441	699	613	-8,0	unnamed protein product, partial (<i>Mesocestoides corti</i>)
MCU_014000-RA	22	73	64	38	-8,0	cysteine-rich secreted protein 2 precursor (<i>Mesocestoides vogae</i>)
MCU_010278-RA	231	184	537	350	-7,9	unnamed protein product, partial (<i>Mesocestoides corti</i>)

4.4 Analýza obohacení genové sady (Gene enrichment analysis)

U replikátů B (tasemnice v inbredních myších hostitelích) představovalo 147 upregulovaných transkriptů (ve srovnání s IV replikáty) 6 obohacených KEGG biologických drah (Tab. 10), většina z nich souvisí s metabolismem (n = 127) [metabolic pathways (n = 69), biosynthesis of secondary metabolites (n = 27), microbial metabolism in diverse environments (n = 15)], metabolismem sacharidů [pentose phosphate pathway] (n = 7) a energetickým metabolismem [oxidative phosphorylation] (n = 9). Dále byly statisticky významně označeny i dráhy související s lidskými chorobami [chemical carcinogenesis (n = 11), yersinia infection (n = 9)], nicméně tyto dráhy nejsou v rámci tohoto výzkumu biologicky relevantní a jedná se o náhodné namapování genů do více biologických drah. Naproti tomu u IV replikátů (tasemnicích kultivovaných *in vitro*) představovalo 58 upregulovaných transkriptů (ve srovnání s B replikáty) 8 obohacených KEGG biologických drah (Tab. 9). Většina upregulovaných transkriptů souvisí s environmentálními informačními procesy (n = 22) [MAPK signaling pathway (n = 8), Ras signaling pathway (n = 8), Notch signaling pathway (n = 4)]. Další obohacená dráha souvisí s genetickými

informačními procesy (n = 15), zejména s translací [aminoacyl-tRNA biosynthesis (n = 10), mRNA surveillance pathway (n = 5)]. Dále jsou obohaceny dráhy související s buněčnými procesy [lysosome] (n = 10). Další signifikantně obohacená dráha souvisí s lidskými chorobami [breast cancer (n = 9), herpes simplex virus 1 infection (n = 4)], nicméně tato dráha není v rámci tohoto výzkumu biologicky relevantní a jedná se o náhodné namapování genů do více biologických drah (Přílohy S3).

U replikátů ICR (tasemnice v outbredních myších hostitelích) představovalo 181 upregulovaných transkriptů (ve srovnání s IV replikáty) 8 obohacených KEGG biologických drah (Tab. 11), většina z nich souvisí s metabolismem (n = 145) [metabolic pathways (n = 80), biosynthesis of secondary metabolites (n = 34), microbial metabolism in diverse environments (n = 21)] a energetickým metabolismem [oxidative phosphorylation] (n = 10). Další obohacené dráhy souvisí s oběhovým systémem [cardiac muscle contraction] (n = 7) a s lidskými chorobami [chemical carcinogenesis – reactive oxygen species (n = 10), yersinia infection (n = 6), dilated cardiomyopathy (n = 5)], nicméně tyto dráhy nás nebudou zajímat, neboť nejsou biologicky relevantní a jedná se o náhodné namapování genů do více biologických drah. Naproti tomu u IV replikátů (tasemnicích kultivovaných *in vitro*) představovalo 84 upregulovaných transkriptů (ve srovnání s ICR replikáty) 13 obohacených KEGG biologických drah (Tab. 9). Většina upregulovaných transkriptů (n = 30) souvisí s buněčnými procesy s transportem a katabolismem [lysosome (n = 11), autophagy – animal (n = 7)] a s buněčným růstem a smrtí [apoptosis] (n = 7). Další obohacená dráha souvisí s environmentálními informačními procesy (n = 25) [MAPK signaling pathway (n = 7), MAPK signaling pathway (n = 6), Ras signaling pathway (n = 8), Rap 1 signaling pathway (n = 4)]. Obohacena je také dráha související s genetickými informačními procesy, konkrétně s translací (n = 19) [aminoacyl-tRNA biosynthesis (n = 8), mRNA surveillance pathway (n = 7), ribosome biogenesis in eukaryotes (n = 4)]. Dále jsou obohaceny dráhy související s endokrinním systémem [oxytocin signaling pathway] (n = 5) a s lidskými chorobami [endocrine resistance (n = 5)] (Přílohy S3).

Tab. 10: Přehled obohacených metabolických drah v B a ICR replikátech ve srovnání s IV replikáty na základě klasifikace KEGG pathway (šedě označeny hodnoty, které nedosáhly statistické významnosti genového obohacení).

	Celkový počet genů v dráze (B & IV)	Počet upregulovaných genů u B (vs IV)	Celkový počet genů v dráze (ICR & IV)	Počet upregulovaných genů u ICR (vs IV)
Metabolism				
Metabolic pathways	102	69	112	80
Biosynthesis of secondary metabolites	40	27	46	34
Microbial metabolism in diverse environments	19	15	26	21
Citrate cycle (TCA cycle)	7	7	9	8
Oxidative phosphorylation	10	9	10	10

Tab. 11: Přehled obohacených metabolických drah v IV replikátech ve srovnání s B a ICR replikáty na základě klasifikace KEGG pathway (šedě označeny hodnoty, které nedosáhly statistické významnosti genového obohacení).

	Celkový počet genů v dráze (B & IV)	Počet upregulovaných genů u IV (vs B)	Celkový počet genů v dráze (ICR & IV)	Počet upregulovaných genů u IV (vs ICR)
Genetic Information Processing				
Aminoacyl-tRNA biosynthesis	10	10	9	8
mRNA surveillance pathway	5	5	8	7
Ribosome biogenesis in eukaryotes	3	1	4	4
Environmental Information				
MAPK signaling pathways	11	8	8	7
MAPK signaling pathway - fly	6	3	7	6
Ras signaling pathway	10	8	9	8
Notch signaling pathway	4	4	5	1
Rap1 signaling pathway	6	3	4	4
Cellular Processes				
Lysosome	11	10	12	11
Autophagy - animal	6	2	7	7
Apoptosis	8	3	9	7
Cellular senescence	5	2	5	5

5 Diskuze

Tato práce se zabývala porovnáním transkriptomických profilů vývojového stádia tasemnice *M. corti* kultivované *in vitro* (IV) a v inbredních (B) a outbredních (ICR) myších kmenech. V této studii byla provedena analýza diferenciální genové exprese mezi všemi dvojicemi vzorků a analýza genového obohacení (gene enrichment analysis), která identifikovala statisticky významně up/downregulované biologické dráhy ve srovnávaných vzorcích. Dosud se transkriptomické studie *M. corti* zaměřovaly především na identifikaci a profilaci mikroRNA dvou vývojových stádií (tetrathyridia, dospělec) (Basika et al. 2016) a jejich transkriptomický profil (Basika et al. 2019).

5.1 Izolace RNA a sekvenační data *M. corti*

Izolovaná RNA nabývala po kontrole kvality pomocí mikrokapilární elektroforézy vysokých RIN hodnot (průměrně 9,0). To vypovídá o vysoké kvalitě RNA, která byla následně sekvenována. Ze sekvenační platformy bylo získáno celkem 347 280 680 pair-end readů. Po filtrování a trimování zůstalo 337 919 452 pair-end readů, což je zhruba 97 %. To značí vysokou kvalitu sekvenačních dat. Na rozdíl od dřívějších transkriptomických studií, kde bylo ze sekvenační platformy získáno 70,2 milionů pair-end readů a po oříznutí zbylo jen 78 % pair-end readů (Basika et al. 2019). Z vysoce kvalitních (otrimovaných) readů bylo namapováno na referenční transkriptom zhruba 67 %, což je podobné, jako u nedávné studie transkriptomického profilu dvou vývojových stádií *M. corti* (Basika et al. 2019), kde bylo provedeno mapování na referenční genom. Nicméně v této práci bylo pouze 39 % readů namapováno unikátně. To může být způsobeno neúplností referenčního transkriptomu, který neobsahuje všechny varianty alternativního sestřihu transkriptů.

5.2 Anotace referenčního transkriptomu

Při anotaci referenčního transkriptomu bylo anotováno cca 91,3 % transkriptů pomocí NCBI databáze a pouze 45,9 % transkriptů pomocí KEGG databáze. Tento výrazný rozdíl může být způsoben faktem, že NCBI databáze obsahuje cca 237 520 318 sekvencí (databáze Genbank) a 1 781 374 217 celogenomových sekvencí (duben 2022). Zatímco databáze KEGG obsahuje 41 397 244 genových sekvencí pro tzv. „KEGG organisms“ (červen 2022). Tasemnice *M. corti* není součástí „KEGG organisms“, což může být také jeden z důvodů,

tak nízkého počtu anotovaných transkriptů. Příbuzné organismy (Platyhelminthes), které jsou v „KEGG organisms“ jsou: *Echinococcus granulosus*, *Opisthorchis viverrini*, *Schistosoma mansoni* a *Schistosoma haematobium*.

5.3 Kvantifikace a analýza diferenciální genové exprese

Na základě exprese transkriptů (min. 10 „expected counts“ v každém replikátu) všechny vzorky sdílí 8491 transkriptů. Tasemnice v outbredních myších hostitelích ICR sdílí větší podobnost s IV kultivovanými tasemnicemi (620 transkriptů) než s tasemnicemi v inbredních myších hostitelích B (411 transkriptů). Což je opak našeho očekávání, neboť z logického hlediska by měly mít tasemnice v myších hostitelích podobnější transkriptomický profil. Tento fakt může způsobit nízké procento unikátně namapovaných readů na referenční transkriptom u všech replikátů, což komplikuje kvantifikaci genové exprese, neboť není jasné přesné místo namapování readu na referenci. Navíc pro tuto statistiku museli být transkripty exprimovány ve všech replikátech s hodnotou „expected counts“ větší rovné 10, takže v některých B replikátech mohl daný transkript chybět.

Dle výsledků analýzy diferenciální genové exprese je zřejmé, že tasemnice v myších hostitelích (B, ICR) mají více upregulovaných genů ve srovnání s IV tasemnicemi. Diferenciální analýza tasemnic kultivovaných v myších hostitelích B a ICR ukázala, že je pouze 9 upregulovaných transkriptů u replikátů B a 17 upregulovaných transkriptů u replikátu ICR. Z toho vyplývá, že genová exprese je u myších replikátů velmi podobná. Z důvodu nízkého počtu diferenciálně exprimovaných genů nebyly výsledky analýzy genového obohacení relevantní.

Na základě zařazení upregulovaných transkriptů získaných z diferenciální analýzy do proteinových rodin pomocí KEGG BRITE databáze je vidět, že transkriptomický profil je pro B a ICR replikáty ve srovnání s IV replikáty velmi podobný. Z čehož vyplývá, že pro studium diferenciální genové exprese tasemnice kultivované *in vitro* a v myším hostiteli nemá zásadní vliv, zda se jedná o outbrední či inbrední myší kmen. Je zde také vidět, že více upregulovaných transkriptů v IV replikátech (ve srovnání s B a ICR) souvisí s transkripčními a translačními procesy, což může znamenat syntézu chybějících proteinů v prostředí mimo definitivního hostitele.

Při následné analýze vysoce upregulovaných transkriptů tasemnic v myších hostitelích (vs IV) ($\log_2\text{foldchange} > 2$) bylo nalezeno několik genů, které by mohly být potenciálními

supresory rakoviny (Tab. 12). Jedním z potenciálních genů je *GLIPRI*. Studie z roku 2010 označila tento gen za kandidátní supresor rakoviny (Thompson 2010). Snížená exprese tohoto genu je znakem několika nádorů (Yan et al. 2018; Sheng et al. 2016). Navíc bylo prokázáno, že je supresorem karcinomu prostaty *in vivo* (Sato et al. 2003). Nicméně v roce 2020 byla provedena další studie (Friend et al. 2020), která se zabývala analýzou mnohočetného myelomu. Studie ukázala, že zvýšená exprese *GLIPRI* výrazně nezměnila růst tumorových buněk *in vitro* ani *in vivo* (Friend et al. 2020). Z toho vyplývá, že tento gen nemusí být supresorem na různé typy rakoviny.

Dalším potenciálním rakovinným supresorem by mohly být geny *KATNA1* (vysoká exprese souvisí s lepším průběhem nemocných žen s rakovinou vaječníků) (Schiewek et al. 2018), *PTPRJ* (receptor protein tyrosin fosfatázy, negativní modulace několika proteinů z proliferačních signálů) (D'Agostino et al. 2018), *ZC3H14* (u hepatocelulárního karcinomu se vyskytují časté delece počtu kopií a downregulace tohoto proteinu, zatímco upregulace významně potlačuje růst karcinogenních buněk *in vitro* a metastáze *in vivo*) (Zhang et al. 2019) a *CD148* (součást proteinových rodin tyrosin fosfatáz receptorového typu kódované genem *PTPRJ*, potlačená exprese indukuje patologické rysy rakoviny a horší stav pacientů, naopak upregulaci zabraňuje proliferaci nádoru a jeho invazi) (Sun et al. 2020).

Tab. 12: Seznam kandidátních genů s jejich pravděpodobnými supresorovými aktivitami, které byly nadměrně upregulovány ($\log_2\text{foldchange} > 2$) v tasemnicích v myších hostitelích (B, ICR).

Název genu	Druh rakoviny	Popis	Zdroj
<i>GLIPRI</i>	Karcinom prostaty	Suprese růstu a proapoptické aktivity ve více rakovinných buněčných liniích Nadměrná exprese stimuluje proapoptotické aktivity pomocí trvalé signalizace ROS-JK	(Thompson 2010)
<i>KATNA1</i>	Karcinom ovaria	Vysoké hladiny exprese korelují s prodloužením života nemocných pacientek	(Schiewek et al. 2018)
<i>PTPRJ</i>	Karcinom plic	Tumor supresorová aktivita negativní modulací několika proteinů zapojených do proliferačních signálů	(D'Agostino et al. 2018)
<i>ZC3H14</i>	Hepatocelulární karcinom (HCC)	Nadměrná exprese významně potlačila růst buněk HCC <i>in vitro</i> a metastáze <i>in vivo</i>	(Zhang et al. 2019)
<i>CD148</i>	Karcinom žaludku	Nadměrná exprese brání proliferaci a invazi tumoru	(Sun et al. 2020)

5.4 Analýza obohacení genové sady (gene enrichment analysis)

Analýza genového obohacení byla provedena na základě Fisherova exaktního testu. Nicméně provádělo se několik stovek testování (pro každou KEGG biologickou dráhu jedno) s hladinou významnosti 0,05, proto je nutné provést korekci mnohočetného testování, aby se zabránilo falešně pozitivním výsledkům. Nicméně, po provedení Bonferroniho korekce nebyly označeny skoro žádné biologické dráhy jako signifikantně up/downregulovány. Důvodem může být, že se jedná o silnou korekci a s narůstajícím počtem provedených testů je relativně těžké dosáhnout statistické významnosti. Proto v provedené analýze tato korekce nebyla použita a všechny statisticky významné biologické dráhy byly brány s určitou opatrností.

Podle upregulovaných drah u tasemnic kultivovaných v myších hostitelích (B, ICR) je vidět, že většina drah souvisí s metabolismem, zejména s biosyntézou sekundárních metabolitů, citrátovým cyklem a oxidativní fosforylací. To může být způsobeno faktem, že tasemnice v myších hostitelích (bez ohledu na kmen) mají energeticky náročnější prostředí, neboť zde například probíhá interakce s imunitním systémem hostitele. Tasemnice kultivované *in vitro* (IV) mají statisticky významně upregulované biologické dráhy související s buněčnými procesy, environmentálními procesy a genetickými informačními procesy (hlavně translační a transkripční dráhy).

Jeden gen se však při procesu mapování může namapovat do více biologických drah a nelze určit, která dráha je ta správná či zda se gen skutečně vyskytuje ve všech namapovaných drahách. Na to je potřeba dát pozor při zpracování výsledků analýzy genového obohacení, neboť to může vést k zavádějícím závěrům. To je názorně vidět u statisticky signifikantně upregulované dráhy u IV (v porovnání s B), kdy je zde upregulovaná biologická dráha související s herpes virem 1 infekcí nebo s rakovinou prsu, což v daném kontextu postrádá biologickou relevanci.

6 Závěr

Cílem teoretické části práce bylo shrnutí jednotlivých kroků analýzy diferenciální genové exprese s příklady dostupných softwarů. V praktické části práce byla provedena analýza transkriptomického profilu vývojového stádia tasemnice *Mesocostoides corti* kultivované *in vitro* a ve dvou kmenech myších hostitelů (inbrední, outbrední) včetně diferenciální genové analýzy a analýzy genového obohacení biologických drah.

V této diplomové práci se podařilo:

- Shrnout jednotlivé kroky analýzy diferenciální genové exprese s příslušnými dostupnými softwary a doporučením, který software je dobré použít pro který typ experimentu.
- Anotovat referenční transkriptom *M. corti* pomocí NCBI a KEGG databáze.
- Kvantifikovat genovou expresi tasemnic kultivovaných *in vitro* (IV) a ve 2 kmenech myších hostitelů (inbrední (B), outbrední (ICR)) a určit, které transkripty jsou specifické pro daný vzorek a které jsou sdílené napříč vzorky.
- Analyzovat diferenciální genovou expresi mezi všemi dvojicemi vzorků (B&IV, ICR&IV, B&ICR) a analyzovat genové obohacení biologických drah.
- Bylo vytipováno několik genů s potenciální supresorovou aktivitou proti rakovině, které byly upregulovány v tasemnicích v myších hostitelích (B, ICR).

V rámci této práce bylo poprvé provedeno studium transkriptomického profilu vývojového stádia tasemnice *M. corti* a vliv prostředí na genovou expresi. Výsledky diferenciální analýzy a analýzy genového obohacení budou dále zpracovány v rámci projektu GAČR - 21-28946S Helmintologickou laboratoří katedry Parazitologie PřF UK a mohou pomoci v identifikaci genů s potenciální supresorovou aktivitou vůči rakovině.

7 Přílohy

Všechny použité skripty a výsledné soubory jsou dostupné v GitHub repozitáři:
https://github.com/korenal/Analyza_transkriptomu_DP

(vysvětlení zkratk: B – tasemnice z inbredního kmene myších hostitelů, ICR – tasemnice z outbredního kmene myších hostitelů, IV – tasemnice kultivované *in vitro*)

Seznam skriptů:

- *filtering_ncbi.py*
Provádí výběr jedné konkrétní anotace z 5 různých NCBI anotací u každého transkriptu primárně na základě funkční anotace genu a sekundárně podle nejvyšší hodnoty *bit_score*
- *transcriptoms_annotation.py*
Provádí anotaci všech transkriptů s hodnotou „expected counts“ větší nebo rovnou 10 všech replikátů konkrétního vzorku (B, ICR, IV)
- *diffexpr_annotation.py*
Provádí slučování tabulky s diferenciálně exprimovanými transkripty jednotlivých dvojic replikátů (B&IV, ICR&IV, B&ICR) s KEGG a NCBI anotacemi na základě unikátního identifikátoru každého replikátu
- *gene_enrichment.py*
Provádí analýzu genového obohacení (gene enrichment analysis) pomocí Fisherova exaktního testu

Seznam výsledných souborů:

- **S1:** Anotované transkripty (B, ICR, IV)
 - *B_transcriptome_annotation.xlsx*
 - *IV_transcriptome_annotation.xlsx*
 - *ICR_transcriptome_annotation.xlsx*

- **S2:** Diferenciální exprese anotovaných transkriptů (B&IV, ICR&IV, B&ICR)
 - *B&IV_diffexprese_anotace.xlsx*
 - *ICR&IV_diffexprese_anotace.xlsx*
 - *B&ICR_diffexprese_anotace.xlsx*

- **S3:** Analýza genového obohacení (Gene enrichment) (B&IV, ICR&IV, B&ICR)
 - *B&IV_gene_enrichment.xlsx*
 - *ICR&IV_gene_enrichment.xlsx*

8 Zdroje

Abello-Cáceres, Paula, Javier Pizarro-Bauerle, Carlos Rosas, Ismael Maldonado, Lorena Aguilar-Guzmán, Carlos González, Galia Ramírez, Jorge Ferreira a Arturo Ferreira, 2016. Does native *Trypanosoma cruzi* calreticulin mediate growth inhibition of a mammary tumor during infection? *BMC Cancer* [online]. **16**(1), 731.

Atayde, Vanessa D., Miriam G. Jasiulionis, Mauro Cortez a Nobuko Yoshida, 2008. A recombinant protein based on *Trypanosoma cruzi* surface molecule gp82 induces apoptotic cell death in melanoma cells. *Melanoma Research* [online]. **18**(3), 172–183.

Baird, Jason R., Katelyn T. Byrne, Patrick H. Lizotte, Seiko Toraya-Brown, Uciane K. Scarlett, Matthew P. Alexander, Mee Rie Sheen, Barbara A. Fox, David J. Bzik, Marcus Bosenberg, David W. Mullins, Mary Jo Turk a Steven Fiering, 2013. Immune-Mediated Regression of Established B16F10 Melanoma by Intratumoral Injection of Attenuated *Toxoplasma gondii* Protects against Rechallenge. *The Journal of Immunology* [online]. **190**(1), 469–478.

Basika, T., G.P. Paludo, F.M. Araujo, A.C. Salim, F. Pais, L. Maldonado, N. Macchiaroli, J. Camargo De Lima, M. Rosenzvit, G.C. Oliveira, L. Kamenetzky a H.B. Ferreira, 2019. Transcriptomic profile of two developmental stages of the cestode parasite *Mesocestoides corti*. *Molecular and Biochemical Parasitology* [online]. **229**, 35–46.

Basika, Tatiana, Natalia Macchiaroli, Marcela Cucher, Sergio Espínola, Laura Kamenetzky, Arnaldo Zaha, Mara Rosenzvit a Henrique B. Ferreira, 2016. Identification and profiling of microRNAs in two developmental stages of the model cestode parasite *Mesocestoides corti*. *Molecular and Biochemical Parasitology* [online]. **210**(1–2), 37–49.

Benamrouz, S., V. Conseil, C. Creusy, E. Calderon, E. Dei-Cas a G. Certad, 2012. Parasites and malignancies, a review, with emphasis on digestive cancer induced by *Cryptosporidium parvum* (Alveolata: Apicomplexa). *Parasite* [online]. **19**(2), 101–115.

Berriel, Edgardo, Sofía Russo, Leticia Monin, María Florencia Festari, Nora Berois, Gabriel Fernández, Teresa Freire a Eduardo Osinaga, 2013. Antitumor Activity of Human Hydatid Cyst Fluid in a Murine Model of Colon Cancer. *The Scientific World Journal* [online]. **2013**, 1–7.

Bland, J M. a D. G Altman, 1995. Statistics notes: Multiple significance tests: the Bonferroni method. *BMJ* [online]. **310**(6973), 170–170.

Blum, Matthias, Hsin-Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasaamy, Alex Mitchell, Gift Nuka, Typhaine Paysan-Lafosse, Matloob Qureshi, Shriya Raj, Lorna Richardson, Gustavo A Salazar, Lowri Williams, Peer Bork, Alan Bridge, Julian Gough, Daniel H Haft, Ivica Letunic, Aron Marchler-Bauer, Huaiyu MI, Darren A Natale, Marco Necci, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Alex Bateman a Robert D Finn, 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research* [online]. **49**(D1), D344–D354.

Bolger, Anthony M., Marc Lohse a Bjoern Usadel, 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* [online]. **30**(15), 2114–2120.

Boratyn, Grzegorz M., Christiam Camacho, Peter S. Cooper, George Coulouris, Amelia Fong, Ning MA, Thomas L. Madden, Wayne T. Matten, Scott D. McGinnis, Yuri Merezuk, Yan Raytselis, Eric W. Sayers, Tao Tao, Jian Ye a Irena Zaretskaya, 2013. BLAST: a more efficient report with usability improvements. *Nucleic Acids Research* [online]. **41**(W1), W29–W33.

Callejas, Blanca E., Diana Martínez-Saucedo a Luis I. Terrazas, 2018. Parasites as negative regulators of cancer. *Bioscience Reports* [online]. **38**(5), BSR20180935.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer a Thomas L Madden, 2009. BLAST+: architecture and applications. *BMC Bioinformatics* [online]. **10**(1), 421.

Camargo De Lima, J., M.E. Caracciolo, L.R. Lima, E.J. Lopes-Torres a H.B. Ferreira, 2020. Morphological and ultrastructural features of *Mesocostoides corti* during in vitro differentiation from larva to adult worm. *Clinical Microbiology and Infection* [online]. **26**(3), 325–326.

Clark, Karen, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell a Eric W. Sayers, 2016. GenBank. *Nucleic Acids Research* [online]. **44**(D1), D67–D72.

D’Agostino, Sabrina, Delia Lanzillotta, Mariaconcetta Varano, Cirino Botta, Antonio Baldrini, Anna Bilotta, Stefania Scalise, Vincenzo Dattilo, Rosario Amato, Eugenio Gaudio, Francesco Paduano, Camillo Palmieri, Rodolfo Iuliano, Nicola Perrotti, Cesare Indiveri, Alfredo Fusco, Marco Gaspari a Francesco Trapasso, 2018. The receptor protein tyrosine phosphatase PTPRJ negatively modulates the CD98hc oncoprotein in lung cancer cells. *Oncotarget* [online]. **9**(34), 23334–23348.

Darani, Hossein Yousofi a Morteza Yousefi, 2012. Parasites and cancers: parasite antigens as possible targets for cancer immunotherapy. *Future Oncology* [online]. **8**(12), 1529–1535.

De Leeuw, Christiaan A., Benjamin M. Neale, Tom Heskes a Danielle Posthuma, 2016. The statistical properties of gene-set analysis. *Nature Reviews Genetics* [online]. **17**(6), 353–364.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson a Thomas R. Gingeras, 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* [online]. **29**(1), 15–21.

Eden, Eran, Roy Navon, Israel Steinfeld, Doron Lipson a Zohar Yakhini, 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* [online]. **10**(1), 48.

Evans, Ciaran, Johanna Hardin a Daniel M Stoebel, 2018. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics* [online]. **19**(5), 776–792.

Finlay, Conor M., Kevin P. Walsh a Kingston H. G. Mills, 2014. Induction of regulatory cells by helminth parasites: exploitation for the treatment of inflammatory diseases. *Immunological Reviews* [online]. **259**(1), 206–230.

Fisher, R. A., 1935. The Logic of Inductive Inference. *Journal of the Royal Statistical Society* [online]. **98**(1), 39.

Friend, Natasha, Jacqueline E. Noll, Khatora S. Opperman, Kimberley C. Clark, Krzysztof M. Mrozik, Kate Vandyke, Duncan R. Hewett a Andrew C. W. Zannettino, 2020. GLIPR1 expression is reduced in multiple myeloma but is not a tumour suppressor in mice. *PLOS ONE* [online]. **15**(1), e0228408.

Garber, Manuel, Manfred G Grabherr, Mitchell Guttman a Cole Trapnell, 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods* [online]. **8**(6), 469–477.

Gotz, S., J. M. Garcia-Gomez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M. Robles, M. Talon, J. Dopazo a A. Conesa, 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* [online]. **36**(10), 3420–3435.

Howe, Kevin L, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azob, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, Mehrnaz Charkhchi, Carla Cummins, Luca Da Rin Fioretto, Claire Davidson, Kamalkumar Dodiya, Bilal El Houdaigui, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Tiago Grego, Cristina Guijarro-Clarke, Leanne Haggerty, Anmol Hemrom, Thibaut Hourlier, Osagie G Izuogu, Thomas Juettemann, Vinay Kaikala, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, José Carlos Marugán, Thomas Maurel, Aoife C McMahon, Shamika Mohanan, Benjamin Moore, Matthieu Muffato, Denye N Oheh, Dimitrios Paraschas, Anne Parker, Andrew Parton, Irina Prosovetskaia, Manoj P Sakthivel, Ahamed I Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, Emily Steed, Michal Szpak, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Brandon Walts, Andrea Winterbottom, Marc Chakiachvili, Ameya Chaubal, Nishadi De Silva, Bethany Flint, Adam Frankish, Sarah E Hunt, Garth R Iisley, Nick Langridge, Jane E Loveland, Fergal J Martin, Jonathan M Mudge, Joanella Morales, Emily Perry, Magali Ruffier, John Tate, David Thybert, Stephen J Trevanion, Fiona Cunningham, Andrew D Yates, Daniel R Zerbino a Paul Flicek, 2021. Ensembl 2021. *Nucleic Acids Research* [online]. **49**(D1), D884–D891.

Huang, Da Wei, Brad T. Sherman a Richard A. Lempicki, 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* [online]. **37**(1), 1–13.

Huang, Da Wei, Brad T Sherman a Richard A. Lempicki, 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* [online]. **4**(1), 44–57.

Hung, Jui-Hung, 2013. Gene Set/Pathway Enrichment Analysis. In: Hiroshi Mamitsuka, Charles Delisi a Minoru Kanehisa, ed. *Data Mining for Systems Biology* [online]. Totowa, NJ: Humana Press, Methods in Molecular Biology, s. 201–213 [vid. 2022-06-16]. ISBN 978-1-62703-106-6.

Hunter, Christopher A., Duonan Yu, Michael Gee, Cam V. Ngo, Cinzia Seignani, Michael Goldschmidt, Tatyana V. Golovkina, Sydney Evans, William F. Lee a Andrei Thomas-Tikhonenko, 2001. Cutting Edge: Systemic Inhibition of Angiogenesis Underlies Resistance to Tumors During Acute Toxoplasmosis. *The Journal of Immunology* [online]. **166**(10), 5878–5881.

- Chen, Chuming, Sari S Khaleel, Hongzhan Huang a Cathy H Wu, 2014. Software for pre-processing Illumina next-generation sequencing short read sequences. *Source Code for Biology and Medicine* [online]. **9**(1), 8.
- Chu, Yongjun a David R. Corey, 2012. RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Therapeutics* [online]. **22**(4), 271–274.
- Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez a S. Hunter, 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* [online]. **30**(9), 1236–1240.
- Kanehisa, Minoru, Miho Furumichi, Yoko Sato, Mari Ishiguro-Watanabe a Mao Tanabe, 2021. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research* [online]. **49**(D1), D545–D551.
- Kanehisa, Minoru, Miho Furumichi, Mao Tanabe, Yoko Sato a Kanae Morishima, 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* [online]. **45**(D1), D353–D361.
- Kanehisa, Minoru a Yoko Sato, 2020. KEGG Mapper for inferring cellular functions from protein sequences. *Protein Science* [online]. **29**(1), 28–35.
- Kanehisa, Minoru, Yoko Sato a Kanae Morishima, 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology* [online]. **428**(4), 726–731.
- Kang, Yun-Jeong, Jin-Ok Jo, Min-Kyoung Cho, Hak-Sun Yu, Sun-Hee Leem, Kyoung Seob Song, Mee Sun Ock a Hee-Jae Cha, 2013. Trichinella spiralis infection reduces tumor growth and metastasis of B16-F10 melanoma cells. *Veterinary Parasitology* [online]. **196**(1–2), 106–113.
- Kim, Daehwan, Ben Langmead a Steven L Salzberg, 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* [online]. **12**(4), 357–360.
- Kim, Ju-Ock, Sung-Soo Jung, Sun-Young Kim, Tae Yun Kim, Dae-Whan Shin, Jae-Ho Lee a Young-Ha Lee, 2007. Inhibition of Lewis Lung Carcinoma Growth by Toxoplasma gondii through Induction of Th1 Immune Responses and Inhibition of Angiogenesis. *Journal of Korean Medical Science* [online]. **22**(Suppl), S38.
- Langmead, Ben a Steven L Salzberg, 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* [online]. **9**(4), 357–359.
- Langmead, Ben, Cole Trapnell, Mihai Pop a Steven L Salzberg, 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* [online]. **10**(3), R25.
- Law, Charity W, Yunshun Chen, Wei Shi a Gordon K Smyth, 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* [online]. **15**(2), R29.

- León-Cabrera, Sonia, Blanca E. Callejas, Yadira Ledesma-Soto, Jossimar Coronel, Carlos Pérez-Plasencia, Emma B. Gutiérrez-Cirlos, Federico Ávila-Moreno, Miriam Rodríguez-Sosa, Rogelio Hernández-Pando, Brenda Marquina-Castillo, Yolanda I. Chirino a Luis I. Terrazas, 2014. Extraintestinal Helminth Infection Reduces the Development of Colitis-Associated Tumorigenesis. *International Journal of Biological Sciences* [online]. **10**(9), 948–956.
- Li, Bo a Colin N Dewey, 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* [online]. **12**(1), 323.
- Li, H. a R. Durbin, 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* [online]. **25**(14), 1754–1760.
- Liao, Yang a Wei Shi, 2020. Read trimming is not required for mapping and quantification of RNA-seq reads at the gene level. *NAR Genomics and Bioinformatics* [online]. **2**(3), lqaa068.
- López, Nandy C., Carolina Valck, Galia Ramírez, Margarita Rodríguez, Carolina Ribeiro, Juana Orellana, Ismael Maldonado, Adriana Albini, Daniel Anacona, David Lemus, Lorena Aguilar, Wilhelm Schwaebler a Arturo Ferreira, 2010. Antiangiogenic and Antitumor Effects of Trypanosoma cruzi Calreticulin. *PLoS Neglected Tropical Diseases* [online]. **4**(7), e730.
- Love, Michael I, Wolfgang Huber a Simon Anders, 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* [online]. **15**(12), 550.
- MacManes, Matthew D, 2015. *Optimizing error correction of RNAseq reads* [online]. preprint. B.m.: Bioinformatics [vid. 2021-12-06].
- Maere, S., K. Heymans a M. Kuiper, 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* [online]. **21**(16), 3448–3449.
- Markoski, Melissa M., Cristiano V. Bizarro, Sandra Farias, Ingrid Espinoza, Norbel Galanti, Arnaldo Zaha a Henrique B. Ferreira, 2003. IN VITRO SEGMENTATION INDUCTION OF MESOCESTOIDES CORTI (CESTODA) TETRATHYRIDIA. *Journal of Parasitology* [online]. **89**(1), 27–34.
- Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer a Barbara Wold, 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* [online]. **5**(7), 621–628.
- Motamedi, Masoumeh, Samaneh Arab, Seied Mohammad Moazzeni, Masoomeh Khamis Abadi a Jamshid Hadhati, 2009. Improvement of a Dendritic Cell-Based Therapeutic Cancer Vaccine with Components of *Toxoplasma gondii*. *Clinical and Vaccine Immunology* [online]. **16**(10), 1393–1398.
- Murphy, David, 2002. Gene expression studies using microarrays: principles, problems, and prospects. *Advances in Physiology Education* [online]. **26**(4), 256–270.

Musich, Ryan, Lance Cadle-Davidson a Michael V. Osier, 2021. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. *Frontiers in Plant Science* [online]. **12**, 657240.

Nicolae, Marius, Serghei Mangul, Ion I Măndoiu a Alex Zelinkovsky, 2011. Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms for Molecular Biology* [online]. **6**(1), 9. ISSN 1748-7188. Dostupné z: doi:10.1186/1748-7188-6-9

Noya, Verónica, Sylvie Bay, María Florencia Festari, Enrique P. García, Ernesto Rodriguez, Carolina Chiale, Christelle Ganneau, Françoise Baleux, Soledad Astrada, Mariela Bollati-Fogolín, Eduardo Osinaga a Teresa Freire, 2013. Mucin-like peptides from *Echinococcus granulosus* induce antitumor activity. *International Journal of Oncology* [online]. **43**(3), 775–784.

O’Leary, Nuala A., Mathew W. Weight, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Eemolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O’Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael Dicuccio, Paul Kitts, Terence D. Murphy a Kim D. Pruitt, 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* [online]. **44**(D1), D733–D745.

Parkin, Donald Maxwell, 2006. The global health burden of infection-associated cancers in the year 2002. *International Journal of Cancer* [online]. **118**(12), 3030–3044.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler a R. Lopez, 2005. InterProScan: protein domains identifier. *Nucleic Acids Research* [online]. **33**(Web Server), W116–W120.

Raghavan, Venket, Louis Kraft, Fantin Mesny a Linda Rigerte, 2022. A simple guide to *de novo* transcriptome assembly and annotation. *Briefings in Bioinformatics* [online]. **23**(2), bbab563.

Ranasinghe, Shiwanthi L. a Donald P. McManus, 2018. *Echinococcus granulosus*: Cure for Cancer Revisited. *Frontiers in Medicine* [online]. **5**, 60. ISSN 2296-858X.

Risso, Davide, Katja Schwartz, Gavin Sherlock a Sandrine Dudoit, 2011. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics* [online]. **12**(1), 480.

Robinson, M. D., D. J. McCarthy a G. K. Smyth, 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* [online]. **26**(1), 139–140.

Robinson, Mark D a Alicia Oshlack, 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* [online]. **11**(3), R25.

Saari, Seppo, Anu Näreaho a Sven Nikander, 2019. Cestoda (Tapeworms). In: *Canine Parasites and Parasitic Diseases* [online]. B.m.: Elsevier, s. 55–81 [vid. 2022-06-16].

Satoh, Takefumi, Terry L. Timme, Takashi Saika, Shin Ebara, Guang Yang, Jianxiang Wang, Chengzhen Ren, Nobuyuki Kusaka, Vladimir Mouraviev a Timothy C. Thompson, 2003. Adenoviral Vector-Mediated mRTVP-1 Gene Therapy for Prostate Cancer. *Human Gene Therapy* [online]. **14**(2), 91–101.

Sayers, Eric W, Jeffrey Beck, Evan E Bolton, Devon Bourexis, James R Brister, Kathi Canese, Donald C Comeau, Kathryn Funk, Sunghwan Kim, William Klimke, Aron Marchler-Bauer, Melissa Landrum, Stacy Lathrop, Zhiyong Lu, Thomas L Madden, Nuala O’Leary, Lon Phan, Sanjida H Rangwala, Valerie A Schneider, Yuri Skripchenko, Jiyao Wang, Jian Ye, Barton W Trawick, Kim D Pruitt a Stephen T Sherry, 2021. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* [online]. **49**(D1), D10–D17.

Sayers, Eric W, Mark Cavanaugh, Karen Clark, James Ostell, Kim D Pruitt a Ilene Karch-Mizrachi, 2019. GenBank. *Nucleic Acids Research* [online]. gkz956.

Syednasrollah, F., A. Laiho a L. L. Elo, 2015. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics* [online]. **16**(1), 59–70.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski a Trey Ideker, 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* [online]. **13**(11), 2498–2504.

Sheng, Xiumei, Nathan Bowen a Zhengxin Wang, 2016. GLI pathogenesis-related 1 functions as a tumor-suppressor in lung cancer. *Molecular Cancer* [online]. **15**(1), 25.

Sherman, Brad T, Ming Hao, Ju Qiu, Xiaoli Jiao, Michael W Baseler, H Clifford Lane, Tomozumi Imamichi a Weizhong Chang, 2022. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research* [online]. gkac194.

Schiewek, Johanna, Udo Schumacher, Tobias Lange, Simon A. Joosse, Harriet Wikman, Klaus Pantel, Marina Mikhaylova, Matthias Kneussel, Stefan Linder, Barbara Schmalfeldt, Leticia Oliveira-Ferrer a Sabine Windhorst, 2018. Clinical relevance of cytoskeleton associated proteins for ovarian cancer. *Journal of Cancer Research and Clinical Oncology* [online]. **144**(11), 2195–2205.

Schuler, Gregory D., Jonathan A. Epstein, Hitomi Ohkawa a Jonathan A. Kans, 1996. [10] Entrez: Molecular biology database and retrieval system. In: *Methods in Enzymology* [online]. B.m.: Elsevier, s. 141–162 [vid. 2022-06-16].

Schurch, Nicholas J., Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G. Simpson, Tom Owen-Hughes, Mark Blaxter a Geoffrey J. Barton, 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* [online]. **22**(6), 839–851.

- Siles-lucas, M a A Hemphill, 2002. Cestode parasites: Application of in vivo and in vitro models for studies on the host-parasite relationship. In: *Advances in Parasitology* [online]. B.m.: Elsevier, s. 133–230 [vid. 2021-11-10].
- Song, Li a Liliana Florea, 2015. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* [online]. **4**(1), 48.
- Strachan, D. P., 1989. Hay fever, hygiene, and household size. *BMJ* [online]. **299**(6710), 1259–1260.
- Sun, Yiting, Song Li, Wenbin Yu, Cheng Chen, Teng Liu, Lanbo Li, Di Zhang, Zeyi Zhao, Jing Gao, Xiao Wang, Duanbo Shi a Lian Liu 2020. CD148 Serves as a Prognostic Marker of Gastric Cancer and Hinders Tumor Progression by Dephosphorylating EGFR. *Journal of Cancer* [online]. **11**(9), 2667–2678.
- Thompson, Timothy C., 2010. Glioma Pathogenesis-Related Protein 1: Tumor-Suppressor Activities and Therapeutic Potential. *Yonsei Medical Journal* [online]. **51**(4), 479.
- Tipney, Hannah a Lawrence Hunter, 2010. An introduction to effective use of enrichment analysis software. *Human Genomics* [online]. **4**(3), 202.
- Tong, Yihan, 2021. The comparison of limma and DESeq2 in gene analysis. *E3S Web of Conferences* [online]. **271**, 03058.
- Trapnell, Cole, Lior Pachter a Steven L. Salzberg, 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* [online]. **25**(9), 1105–1111.
- Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn a Lior Pachter, 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* [online]. **7**(3), 562–578.
- Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold a Lior Pachter, 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* [online]. **28**(5), 511–515.
- Ubillos, Luis, Teresa Freire, Edgardo Berriel, María Laura Chiribao, Carolina Chiale, María Florencia Festari, Andrea Medeiros, Daniel Mazal, Mariella Rondán, Mariela Bollati-Fogolin, Gabriel A. Rabinovich, Carlos Robello a Eduardo Osinaga, 2016. *Trypanosoma cruzi* extracts elicit protective immune response against chemically induced colon and mammary cancers: *Trypanosoma Cruzi* Extracts Elicits Antitumor Immunity. *International Journal of Cancer* [online]. **138**(7), 1719–1731.
- Vennervald, B. J. a K. Polman, 2009. Helminths and malignancy: Helminths and malignancy. *Parasite Immunology* [online]. **31**(11), 686–696.
- Wagner, Günter P., Koryu Kin a Vincent J. Lynch, 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences* [online]. **131**(4), 281–285. ISSN 1431-7613, 1611-7530.

- Wang, Ligu, Shengqin Wang a Wei Li, 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* [online]. **28**(16), 2184–2185.
- Wang, X.L., B.Q. FU, S.J. YANG, X.P. Wu, G.Z. Cui, M.F. Liu, Y. Zhao, Y.L. Yu, X.Y. Liu, H.K. Deng, Q.J. Chen a M.Y. Liu, 2009. *Trichinella spiralis*—A potential anti-tumor agent. *Veterinary Parasitology* [online]. **159**(3–4), 249–252.
- Wang, X.L., M.Y. Liu, S.M. Sun, X.L. Liu, L. Yu, X.R. Wang, L.X. Chu, B. Rosenthal, H.N. Shi, P. Boireau, F. Wang, Y. Zhao a X.P. Wu, 2013. An anti-tumor protein produced by *Trichinella spiralis* induces apoptosis in human hepatoma H7402 cells. *Veterinary Parasitology* [online]. **194**(2–4), 186–188.
- Waranid Sutthacharoenthad, Nattavat Sukko a Jiramaetha Sinsuebchuea, 2018. Evaluation of Reference-based Mapping Tools for RNA-Seq Pipeline [online]. [vid. 2022-06-16].
- Williams, Claire R., Alyssa Baccarella, Jay Z. Parrish a Charles C. Kim, 2016. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics* [online]. **17**(1), 103.
- Wingett, Steven W. a Simon Andrews, 2018. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research* [online]. **7**, 1338.
- Yan, Liang, Qi Li, Juan Yang a Baoping Qiao, 2018. TPX2-p53-GLIPR1 regulatory circuitry in cell proliferation, invasion, and tumor growth of bladder cancer. *Journal of Cellular Biochemistry* [online]. **119**(2), 1791–1803.
- Yandell, Mark a Daniel Ence, 2012. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* [online]. **13**(5), 329–342.
- Yousofi Darani, Hossein, Narges Soozangar, Soliman Khorami, Fatomeh Taji, Mortaza Yousofi a Hedayatollah Shirzad, 2012. Hydatid Cyst Protoscolices Induce Cell Death in WEHI-164 Fibrosarcoma Cells and Inhibit the Proliferation of Baby Hamster Kidney Fibroblasts *In Vitro*. *Journal of Parasitology Research* [online]. **2012**, 1–4.
- Zhang, Chuxiao, Pengbo Cao, Aiqing Yang, Xia Xia, Yuanfeng Li, Mengting Shi, Ying Yang, Xiaojun Wei, Chun Yang a Gangqiao Zhou, 2019. Downregulation of ZC3H14 driven by chromosome 14q31 deletion promotes hepatocellular carcinoma progression by activating integrin signaling. *Carcinogenesis* [online]. **40**(3), 474–486.
- Zhou, Qian, Xiaoquan Su, Gongchao Jing, Songlin Chen a Kang Ning, 2018. RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data. *BMC Genomics* [online]. **19**(1), 144.

9 Online zdroje

1. WormBase ParaSite genome assembly & annotation of *Mesocestoides corti*, [cit. 10.1.2022]
https://parasite.wormbase.org/Mesocestoides_corti_prjeb510/Info/Index/
2. Wellcome Trust Sanger Institute. 2014. The 50 Helminth Genomes Initiative, [cit. 15.2.2022]
<https://www.sanger.ac.uk/collaboration/50hgp/>
3. Andrews S. FASTQC. A quality control tool for high throughput sequence data, [cit. 20.2.2022]
<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
4. Quality Scores for Next-Generation Sequencing, [cit. 15.5.2022]
https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf
5. Maximum read length for Illumina sequencing platforms, [cit. 20.5.2022]
<https://support.illumina.com/bulletins/2020/04/maximum-read-length-for-illumina-sequencing-platforms.html>
6. Read mapping to a reference genome/transcriptome, [cit. 26.5.2022]
https://biocorecrg.github.io/PHINDaccess_RNAseq_2020/alignment.html
7. Bioconductor, [cit. 1.6.2022]
<https://www.bioconductor.org/>
8. Globální vyhledávací stránka Entrez, [cit. 10.6.2022]
<https://www.ncbi.nlm.nih.gov/search/>
9. E-values and Bit-scores in BLAST, [cit. 15.6.2022]
[E-values and Bit-scores in BLAST - Ricardo Avila \(ravilabio.info\)](https://www.ravilabio.info/e-values-and-bit-scores-in-blast/)

10. InterPro consortium member databases, [cit. 20.6.2022]
<https://interpro-documentation.readthedocs.io/en/latest/databases.html>
11. Protocol for use with NEBNext Poly(A) mRNA Magnetic Isolation Module, [cit. 7.12.2020]
<https://international.neb.com/protocols/2017/02/05/protocol-for-use-with-nebnext-poly-a-mrna-magnetic-isolation-module-e7490-and-nebnext-ultra-ii-di>
12. Protocol for use with NEBNext rRNA Depletion Kit (Human/Mouse/Rat) and NEBNext Ultra II Directional RNA Library Prep Kit for Illumina, [cit. 7.12.2020]
<https://international.neb.com/protocols/2017/02/07/protocol-for-use-with-nebnext-rna-depletion-kit-human-mouse-rat-neb-e6310-and-nebnext-ultra-ii-directional-rna-library-prep-kit-for-illumina-neb-e7760-e7765>