Univerzita Karlova v Praze

Přírodovědecká fakulta

Studijní program:

Molekulární biologie, genetika a virologie

Bc. Jan Röslein

Mutační a substituční tempo u sexuálních a klonáních forem: možný klíč k vysvětlení persistence sexu u modelové skupiny sekavců

Mutation AND substitution rates in sexual and asexual forms: a clue to the persistence of sex in a model group of Cobitis?

Typ závěrečné práce

Diplomová

Vedoucí závěrečné práce: Mgr. Karel Janko, Ph.D.

Praha, 2015

Velký dík náleží mému školiteli Mgr. Karlu Jankovi, Ph.D. za velmi nápomocné, direktivní vedení práce. Též bych rád poděkoval panu Mgr. Janu Pačesovi, Ph.D. za více než vzdělávací rozměr v oblasti bioinformatické analýzy a Mgr. Ladislavu Pekárikovi, Ph.D., Mgr. Janu Kočímu za pomoc při analýze vybraných kapitol. Také bych rád poděkoval rodině za podporu. Všem participantům na této diplomové práci se hluboce omlouvám za způsobenou psychickou újmu.

Prohlášení:

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

V Praze dne 12. 8. 2015

Podpis:

Abstrakt

Klíčová slova:

Abstract

Key words:

Obsah

Úvod

Diplomová práce se zabývá historicky starou, leč stále zcela nevyřešenou problematikou evoluce sexuální reprodukce, speciace a hybridizace. Pro studium asexuality existuje excelentní modelový komplex relativně blízce příbuzných druhů kostnatých ryb rodu Sekavec produkujících viabilní potomstvo vzniklého z mezidruhového křížení, které je nicméně gynogenetické (samec nemusí geneticky přispět k tvorbě potomstva, je pouze aktivátorem diploidního, či polyploidního oocytu, z něhož vzniká embryo). Výhoda tohoto modelového komplexu tkví nejen v detailně prostudované fylogenezi, populační historii, ekologických preferencích, ale i fyziologii druhů *Cobitis taenia, Cobitis elongatoides* a *Cobitis tanaitica*. Navíc mnohé hybridní kombinace těchto druhů vyskytující se běžně v přírodě byly artificiálně vytvořeny *in vitro*.

Od úsvitu dějin je lidem jasné, že biologická rozmanitost mezi organismy není kontinuální, ale spíše má tendenci shlukovitě klastrovat. Ponechme stranou, zda má smysl hledat obecně platnou definici druhu, či zda pro různé typy organismů platí různé definice (Dubois 2011), ale samotná existence této shlukovitosti se dá považovat za důkaz objektivní existence druhů (Fontaneto et al. 2007). Druhy mohou vznikat různými způsoby, ale přinejmenším pro sexuální organismy proces speciace vyžaduje existenci mechanismů omezujících genový tok mezi druhy. Ty bývají typicky kategorizovány jako pre a post zygotické. Výzkum speciace ukázal, že mezidruhová diferenciace může být způsobena několika dibergovanými geny velkého účinku (e.g. (Mihola et al. 2009)) zatímco zbytek genomu může podléhat výrazné admixii (e.g. (Nadeau et al. 2012)). Možná však je také účast mnoha genů s malým, ale epistatickým účinkem, které mohou být loializovány do specifických oblastí v genomu, či výrazně roztroušeny (rev. in (Seehausen et al. 2014))(see e.g. (Parchman et al. 2013)).

Ať již však ke speciaci dochází jakkoliv, zdá se, že speciační proces má charakter kontinua, kdy je míra reprodukční izolace mezi druhy úměrná jejich genegické ditanci – toto bývá také nazýváno tzv. ‘speciation clock’ (Bolnick and Near 2005). Komparativní studie ukázaly, že s tím, jak se zvyšuje distance mezi hybridizujícími druhy, snižuje se nejen fertilita hybridů, ale i typ jejich postižení. Například u ryb (Russell 2003) hybridizující blízké páry druhů vytvářejí oboustranně plodné potomstvo, ale se zvyšující se distancí mezi druhy vzrůstá pravděpodobnost postižení fertility jednoho, posléze obou pohlaví a nakonec se u nejvzdálenějších párů projevuje hybridní neživotaschopnost.

Hybridizace je však úzce spjata i s fenoménem asexuality a produkcí hybridních klonálních linií (e.g. (Choleva et al. 2012). Jak kauzálně hybridizace a asexualita spolu souvisí není známo, ale má se za to, že dva nesourodé genomy v jedinci nedokáží úspěšně kontrolovat složitý meiotický aparát, což může vést k produkci neredukovaných gamet. ((Schultz 1969); (CARMAN 1997); (California 2008)). (Moritz et al. 1992) si všiml, že proporce neredukovaných gamet u hybridů stoupá s divergencí jejich rodičů a navrhnul tzv. *Balance Hypothesis*, která predikuje, že ke vzniku trvale asexuální linie může dojít jen tehdy, když jsou parentální genomy dostatečně vzdáleny, aby jejich kombinace vedla k disrupci meiosy, ale ne tak vzdáleny, aby byla výrazně postižena fitness hybrida. Tato hypotéza koreluje s pozorováními, že známé hybridní klonální formy pocházejí z hybridizací druhů, které nejsou sesterské, ale vždy je mezi nimi jistá větší genetická distance (Moritz et al. 1992). Je tedy docela možné, že tak jako tvorba reprodukčně izolačních mechanismů vyžaduje postupnou akumulaci epistatických mutací mezi divergujícími druhy (Dobzhansky-Müller model speciace), tak i vznik asexuality pomocí hybridizace je výsledkem obdobné akumulace epistatických interakcí postihující proces meiosy.

1.1 Asexualita

Termín asexualita je používán především ve spojení s eukaryotními organismy, jelikož výlučně eukaryota přešla ve svém vývoji k "pravému" sexuálnímu rozmnožování. Asexualitu vnímáme jako stav reversovaný, kdy organismy přešly od sexuální formy ke klonální reprodukci zabraňujíce rekombinaci sesterských alel a změnám redukčního dělení, ať již endoduplikací genové sádky nebo přeskočením redukční fáze meiózy endoduplikací po redukční fázi. Jelikož je meióza velmi komplexním procesem, nemůžeme říci, že by existovala sada jedinečných změn podmiňující zvrat ke klonální reprodukci (Bengtsson 2009). Termín asexualita je ale tradičně používán i ve spojitosti s reprodukcí bakterií a archea, proto byl nahrazen termínem apomixie (Kondrashov 1993). Ačkoliv samotné zařazení bakterií a archeí mezi asexuální jedince je zvláštní, uvážíme-li, že rekombinace, inkorporace i původem xenogenní DNA je zcela běžné. Nicméně v odborné literatuře se setkáváme s oběma termíny, budu tedy v této práci nadále užívat termínu asexuální reprodukce.

Asexualitu můžeme definovat krátce jako zavržení sexu a přeskočení redukce gamet také často vedoucí ke změnám ploidie. Korelace mezi polyploidií a asexualitou je

velmi těsná. Polyploidie je pro meiotický cyklus velmi nevýhodný stav, protože způsobuje početní aberace v genomu, častým produktem je vznik aneuploidií, zejména pak v případě orthoploidie. Až na výjimky vznik polyploidní sádky genomu není spojen s redukcí genomu, ale právě naopak, úroveň ploidie buněk zůstává většinou konstantní (Bengtsson 2009). S asexuální reprodukcí je rovněž úzce asociována hybridizace, protože hybridizace může vést opět k narušení meiózy (Johnson and Bragg 1999). Hybridizace a polyplidizace nejčastěji nastávají v jeden okamžik. V případě hybrida s ortoploidní sádkou chromozomů mohou vznikat gamety schopné dát vzniku jak sexuálním, tak asexuálním gametám.

2.1.1 Mechanismy vzniku neredukovaných gamet

Změny v meióze vedoucí k asexuálnímu rozmnožování byly studovány a to velmi detailně u rostlin, proto se v této části zaměřím především na ně. Principy alterací meiózy jsou ale univerzální a nevztahují se exklusivně na říši rostlin.

Jak je výše zmíněno, pro přechod k asexuální reprodukci je nezbytné zabránit redukci gamet. Neredukované gamety vznikají u sexuálních rostlin spontánně s četností menší než 0.5 %. Genetické defekty meiózy odráží především fáze, ve které vznikly.

Meiotická fáze I je charakterizovanou párováním a rekombinací v místech chiasmat. Změny v proteinech zásadních pro profázi I vedou nejčastěji poruchám párování sesterských chromatid a tedy vzniků univalentů, což vede k disbalancované segregaci anafázi I a disbalancované segregaci v anafázi II, především pokud také dojde k předčasné ztrátě kohezinového kompexu mezi sesterskými chromatidami, např. díky mutaci v rekombináze 8 (De Muyt et al. 2009). Tato situace byla mnohokrát popsána jako následek mutace v komplexu dyad/SWI1 (chromosomové organizátory) (Ravi et al. 2008). Problémy obecnějšího rázu v buněčném cyklu mohou hrát také významnou roli. Jak je notoricky známo, mezi hlavní regulátory buněčného cyklu patří cyklin dependentní kinázy. Zcela zásadním rozdílem mitózy o meiózy jsou dvě konsekventní dělení bez stádia replikace DNA. Byť malá změna v správné regulaci hladiny cyklinů mezi fázemi I a II meiózy, může vést k syntéze DNA – vložení S fáze a tvorbě diploidní gamety. Za replikaci mezi meiózou I a II je např. zodpovědný CYCA1 a CYCA2, či OSD1 (neznámá funkce, pravděpodobně moduluje funkci CDK skrze aktivaci APC kompexu) bránicí vstupu do druhé fáze meiotického cyklu. Vznikají tedy dvě diploidní, nikoli čtyři haploidní gamety (Wang et al. 2010). Další cestou vedoucí k neredukovaným gametám je mechanismus

spojený s orientací chromosomů na vřeténku během meiózy II, kdy je nutno fyzicky oddělit dvě dělící vřeténka (Ramanna and Jacobsen 2003); tento mechanismus se nicméně mezi živočichy a rostlinami mírně liší. Krytosemenné rostliny po první telofázi I zůstávají ve společené cytoplasmě až do druhé cytokineze. Organizace, orientace dělících vřetének musí být proto přísně kontrolována. Zde můžeme například uvést mutanty genů *Atps1* a *jason* způsobující přeskupování chromozomů oddělených po telofázi I, kdy každá diploidní buňka může obsahovat i chromatidy homologních chromozomů. Vznikají dyády, triády balancovaných, či disbalancovaných konstitucí díky fúzím mikrotubulů, či jiným přeskupením dělících vřetének (d'Erfurth et al. 2008). Nicméně toto se děje pouze u samčích gamet, samičí meióza II se vyznačují jinou třídimenzionální organizací. Další problémy mohou nastávat také během cytokineze, ale ty nebudou rozvedeny, protože se exkluzivně vztahují pouze na samčí pohlaví.

2.1.2 Gynogenetická forma reprodukce

Gynogenese, jak vyplývá z názvu, označuje materiální původ genomu (antonymem by byla androgeneze – eliminace maternálního genomu a splynutí dvou spermií). Gynogenetická reprodukce (na spermiích závislá partenogeneze, botaniky nazývaná pseudogamie) je označení jedné z forem klonální reprodukce dependentní na spermiích pro iniciaci dělení oocytu. Genetická informace opačného pohlaví se až na výjimky žádným způsobem nepodílí svou genetickou informací, nepředává ji do další generace. Samice se z tohoto pohledu chovají paraziticky, jelikož samec z oplozování vajíček prakticky nemá žádný benefit, naopak "plýtvá" energií. Donor spermií může být i hermafrodický jedinec. K syngamii, fúzi buněk, zpravidla nedochází, pokud ano, zygota zaniká, nebo může dát vzniku jedinci polyploidního genomu, přičemž paternální genotyp bývá transkripčně umlčen. Jak bylo řečeno, meiotických alterací vedoucích k zachování ploidie pohlavních buněk, je mnoho. Gynogeneze bývá spojena z polyploidií a to linií tvořenou gpouze samicemi nebo hermafrodity. Gynogeneze byla nalezena u kmenů Chordata, Mollusca, Annelida, Arthropoda, Rotifera a Platyhelmites (Beukeboom and Vrijenhoek 1998).

2.1 Teorie úspěchu sexuální formy reprodukce

K čemu je sexuální reprodukce vůbec zapotřebí, když je možné zvolit cestu náročné klonální reprodukce? A to z mnoha hledisek, sexuální reprodukce vyžaduje mnoho

energieg a času na vyhledání partnera, zvyšuje riziko napadení predátory, přenosu parazitů. Vývoj reprodukčních orgánů je sám o sobě energeticky náročný, nemluvě o nákladech vydaných na atrakci přenašečů gamet, atrakce partnera a soupeření. Fitness obou pohlaví musí být stabilizován i přes často bizarní morfologické rozdíly. Sexuální druhy musí udržovat jistou *densitu* jedinců populace pro úspěšné párování. Také vyvstává otázka konfliktu mezi pohlavími, samec často investuje do vývoje potomka méně energie. Hlavním argumentem je ale fakt, že fitness klonální reprodukce je 2x vyšší nežli u sexuální reprodukce – hypotéza nazvaná "two-fold cost of sex" (Flegr 2007). Selekční koeficienty, se kterými je běžně pracováno v populační genetice, se málokdy blíží hodnotě 0.5. Tedy proč došlo u bezmála 98 % eukaryot k přechodu k sexuální reprodukci?

Hledisko krátkodobých přínosů vysvětluje hypotéza synergistické epistáze. V případě, že na jednom chromosomu koexistující mutace mající různé fenotypové projevy vzhledem k fitness, bude výsledek fitness jedince vždy sumou jednotlivých mutací. V případě klonální reprodukce nelze tyto vazbové skupiny rozbít a vyloučit je z populace. Proč by měla být rychlejší adaptabilita výhodnější? Na tuto otázku odpovídá teorie červené královny – "Aby ses dostala někam jinam, musíš běžet dvakrát tak rychleji!" (Lewis C.). Teorie červené královny popisuje vztah host-parazit v evoluci (Hamilton 1980). Rychlé změny, adaptabilita na nové prostředí dávájí hostiteli velkou výhodu v obraně před parazity (Flegr 2007)(Salathe et al. 2008). Naopak u hybridů, asexuálů by se dal očekávat díky heteroznímu efektu, odlišnému fenotypu od obou rodičů dočasně opačný efekt. Odbočíme-li mírně od generalizovaného tématu; u studovaného komplexu gynogenetických ryb s rodičovskými druhy rodu *Cobitis* nebyla tato korelace detekovatelná, i když byla potvrzena rozdílná preference mikrohabitatů (Kotusz et al. 2014), morfologie, fyziologie a exprese na úrovni celého transkriptomu odlišná od obou rodičů.

Z pohledu dlouhodobých výhod sexuální reprodukce hovoříme o DNA reparaci: u určitých druhů společná výchova potomků, sexuální selekce ve prospěch nejúspěšnějšího samce. DNA reparace, možnost využití nadpočetné kopie DNA k reparaci HEJ, NHEJ je spjata spíše s polyploidizací haploidního genomu, která je ale také spojená s nástupem sexuální reprodukce. Hlavní výhody sexuální reprodukce popsal Fisher-Müller. Představme si situaci, kdy máme dva loci a čtyři alely, např. dvě dominantní a dvě recesivní alely, přičemž dominantní alely leží na jiném loci než recesivní alely. Řekněme, že pro adaptaci v novém prostředí je vyžadována přítomnost obou recesivních alel v

jedinci, nejpravděpodobnější situací je tedy stochastický vznik mezi členy populace. V klonální linii je nezbytné "vyčkat" na přítomnost recesivních alel loci v jedinci, zatímco sexuální reprodukce dává možnost spojení a rekombinace loci mezi jedinci, a tak i rychlejší adaptabilitě. Müller navrhl další přelomovou hypotézu ve studiu asexuality: Müllerova rohatka. Ta říká, že rekombinace není jen schopna kombinovat výhodné mutace pro urychlení adaptace, ale může zbavit genotyp nevýhodných, škodlivých mutací. Zjednodušme situaci následujícím způsobem: mutace mají nezávislou fitness, ke zpětným mutacím téměř nedochází a populace je má nekonečnou efektivní velikost populace (drift nehraje roli), pak průměrný počet mutací na chromosomu bude nepřímo úměrný mutační rychlosti. U klonálních glinií tedy s každou generací musí zákonitě klesat fitness a není cesty zpět, reverzní mutace a výskyt selekčně výhodné mutace je spíše ojedinělý jev. Hlavní roli v procesu degenerace tedy hraje efektivní velikost populace, drift. Naproti tomu u sexuální reprodukce přeskupením vazebných skupin může dojít k rychlé purifikaci mutací s nepříznivým vlivem na fitness. Müller (1964) se díval na evoluci perzistence sexu spíše z pohledu asexuálních linií; existuje též hypotéza Kondrashovy sekerky (Kondrashov 1993), která se naopak na problém akumulace nesynonymních mutací dívá z pohledu vývoje sexuálních druhů. Každopádně společným jmenovatelem teorií Müllerovy rohatky a Kondrashovy sekerky je fakt, že nesynonymní mutace vedou k snížení fitness (Kondrashov 1988) (Kondrashov 1993). Kondrashov především poukázal na význam nezávislosti mutací, protože v případě s akumulací nesynonymních mutací může vzrůstat také synergie mezi mutacemi vzhledem fitness a vliv Müllerovy rohatky může být značně zpomalen (Kondrashov 1994). Ačkoliv se asexuální linie zdají být evolučně mrtvé, bez významu, opak je pravdou.

Asexuální linie mohou díky snížení efektivní velikosti populace (selekce na pozadí) a driftu přežít opravdu dlouhou dobu, takovým příkladem je čeleď *rotifera* rod *bdelloidea*, kde známe asexuální linie staré 35 – 40 miliónů let (Waggoner and Jr 1993), jejich stáří může být ale až dvojnásobné (Mark Welch and Meselson 2000). Asexuální linie vznikaly v evoluci druhů nezávisle, mnohokrát a často velmi výrazně ovlivňovaly vývoj druhů, ze kterých vznikly; ať již urychlily separací druhů, nebo se naopak staly konkurenty, sexuálními parazity snižující celkovou fitness obou, či jednoho druhu. Vznik klonálních linií se v evoluci opakoval nesčetněkrát. "Mírná zátěž parazitů spolu s rozumnou mírou mutační rychlosti může poskytnout sexu obranu proti opakovaným invazím klonů."

(Howard 1994). Asexuální hybridy můžeme také označit za jeden z mezistupňů evoluce druhů.

Sexuální reprodukci můžeme ale také označit za evoluční past, protože vedla k vývoji genomického imprintingu mezi pohlavími, který brání alternativním formám vývoje. Na význam imprintingu, jeho příčině u sexuálních organismů existují dvě teorie, první říká, že imprinting vznikl jako následek konfliktu mezi pohlavími díky rozdílným investicím do vývoje a výchovy nové generace. Příkladem mohou být geny pro růst placenty, např. *Igf2* (Moore and Haig 1991), nebo notoricky známý gen *medea*. Druhá teorie vyzvedává význam evoluční, prezence alel v genomu rozdílného fenotypu, přičemž pouze jedna alela je exprimována, může hrát zásadní roli adaptabilitě, plasticitě organismu – se změnou prostředí může dojít ke změně imprintingu alel (Beaudet and Jiang 2002).

3.1 Evoluce rodu *Cobitis*

Ve své práci jsem se zaměřil na genomické studium konsekvencí hybridizace, polyploidizace a asexuality u hybridního komplexu *Cobitis taenia*. Tato skupina sladkovodních ryb vznikla patrně během terciéru a jako jediná linie rodu *Cobitis* kolonizovala ne-Mediterránní Evropu (Bohlen et al. 2006). Během tohoto procesu se rozrůznila do několika druhů obývajících široké oblasti od Atlantických povodí až po Volhu a od Skandinávie až po Černé Moře (Janko et al. 2007). Konkrétně se jedná o následující druhy (jelikož se jedná o větší počet druhů a jejich následných kombinací v hybridních liniích, uvedu za druhovým jménem i zkratku, pomocí níž budu genom daného druhu označovat): C. taenia (TT; T značí genom tohoto druhu, tudíž čistá diploidní forma má toto označení – u dalších druhů tomu bude analogicky), *C. tanaitica (nn), C. pontica (pp), C. taurica (cc)* a *C. elongatoides (ee)*. Jak ukázáno, tyto druhy se velice ochotně kříží a do dnešní doby bylo popsáno mnoho typů hybridů v různých kombinacích včetně polyploidních: *et* (tzn. *e* x *t*), *en, ec, eet, ett, een, enn,* a dokonce i trihzbridních kombinací *etn, etp* (Janko et al. 2007).

O evoluční historii tohoto komplexu je známo, že druh *C. elongatoides* patrně divergoval od ostatních v Pliocénu a obsadil Dunajskou oblast, zatímco zbývající druhy mají rozšíření Pontokaspické (Bohlen et al. 2006). Během klimatických změn, především v Pleistocénu, docházelo k sekundárním kontaktům mezi víceméně alopatrickými druhy a jejich vzájemnému křížení. Takto právě vznikaly zmíněné hybridní linie, z nichž nejstarší

má kolem 350 tisíc let (Janko et al. 2005). Evolučně velice zajímavé jest to, že všechny dosud známé hybridní linie se rozmnožují klonálně a to gynogeneticky (Janko et al. 2007), (Choleva et al. 2012), přičemž dochází ke tvorbě klonálních vajíček, která jsou posléze oplodněna spermiemi samců rodičovských druhů, avšak genom spermie je obvykle zničen a oplození pouze iniciuje dělení a vývoj klonálního vajíčka. Tyto a podobné asexuální linie jsou v literatuře také nazývány „pseudogamní" anebo sexuální paraziti (Bengtsson 2009). Ve vzácnějších případech genom spermie s vajíčkem splyne a vytvoří polyploidní zygotu a založí tím nový polyploidní klon. Tím vlastně dochází u sekavců k vývoji tzv. „leaky gynogenesis" (Janko et al. 2007), kdy je rozmnožování klonální, avšak může docházet k jednosměrnému genovému toku z rodičovských druhů a inkorporacím jeho genomů. Tak došlo k tomu, že nejstarší známá asexuální hybridní linie patrně vznikla před cca 350ti tis. lety jako diploidní EN forma, ale postupem času dala vznik mnoha nezávislým triploidním klonálním liniím o genomové kompozici *een*, *enn* i *etn*. Proces polyploidizace nekončí na triploidní úrovni, ale pokračuje dále k tetraploidizaci; nicméně z doposud neznámých důvodů tyto tetraploidní linie nejsou úspěšné a až na výjimky netvoří perzistentní klonální linie (Janko et al. 2012). Ve skutečnosti tetraploidní zygoty vykazují řádově vyšší úmrtnost než zygoty triploidní (Juchno and Boroń 2006).

Ačkoliv o cytologických mechanismech klonality není u evropských sekavců mnoho známo, lze se na základě s jejich vzdálenými asijskými příbuznými (asexuální linie japonských Misgurnus anguilicaudatus; (Zhang et al. 1998)) domnívat, že ke klonalitě dochází pomocí tzv. „premeiotické endoduplikace". Oogonie se před vstupem do meiozy endoduplikují – z diploidních, nebo triploidních oogonií se stanou tetra- nebo hexaploidní oogonie - a ty pak vstoupí do „normální" meiosy s rekombinací a segregací. Avšak k tvorbě bivalentů dojde jen mezi sesterskými chromosomy vzniklými endoduplikací, proto rekombinace nevnáší do potomstva žádnou variabilitu a výsledkem je vzhledem k somatické tkáni neredukovaná klonální gameta.

Na rozdíl od klasických případů hybridizace se sekavčí hybridi nezdržují jen v úzkých hybridních zónách, kde dochází k reprodukčnímu kontaktu rodičovských druhů, ale expandují do zázemí jednotlivých druhů tak úspěšně, že v podstatě okupují celé jejich dnešní areály. Teoreticky by se mohlo zdát, že sekavčí hybridi sice mohou mít místně a dočasně velký význam – jsou schopni úspěšně kompetovat s rodiči, užírat jim zdroje, měnit jejich populační hustoty, a dokonce, jak matematicky dokázáno, jsou schopni i výrazně ovlivňovat biogeografii rodičovských druhů tím, že omezují jejich počty a tím i

šanci expandovat (Janko and Eisner 2009). Nicméně z dlouhodobého hlediska slouží jen jako evoluční „žumpa“ pro genomy, které se do nich dostanou při jejich vzniku. To proto, že není znám zatím žádný způsob, jak by mohlo dojít ke zpětnému genovému toku z klonálních hybridů zpět do sexuálních druhů. Takže pokud skutečně klony časem podlehnou zmíněným procesům jako Müllerova rohatka, vezmou s sebou do hrobu celou svoji genetickou výbavu aniž ji mohly někomu předat.

Ukázalo se ale, že tomu tak vždy být nemohlo. (Choleva et al. 2014) ukázal, že *C. tanaitica*, ač jaderně velmi blízký druhu *C. taenia*, má mitochondriální DNA velice blízce příbuznou druhu *C. elongatoides*. Pomocí matematické analýzy autoři dokázali, že k takovému mosaicismu mohlo dojít jedině hybridizací, což ukazuje obrovský paradox. Na jednu stranu máme komplex druhů, které se spolu kříží, ale hybridi jsou pouze klonální, což teoreticky vylučuje jakoukoliv výměnu genů mezi druhy, na druhou stranu zde máme vzhledem k velkým areálům jeden z největších známých případů fixace cizorodé mitochondriální genealogie v živočišné říši.

Sekavec se tedy jeví jako zcela excelentní modelový taxon umožňující studovat, jak spolu souvisí speciace, hybridizace, asexualita i polyploidie.

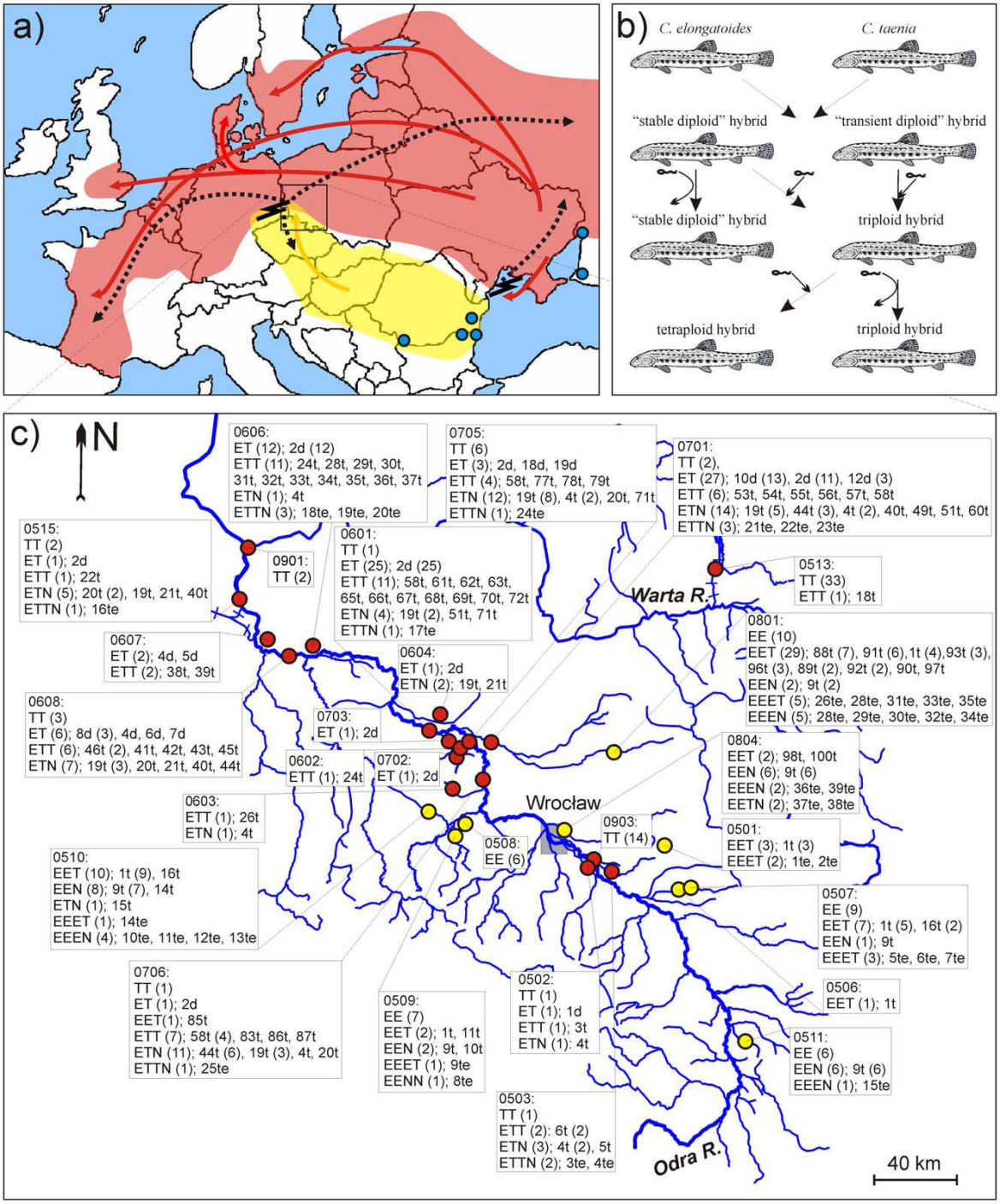


Schéma č. 0: Na obrázku a) je znázorněno geografické holocenní šíření druhů *C. elongatoides* a *C. taenia*. Na obrázku b) je schématicky vyobrazeno, jakým způsobem dochází k hybridizaci mezi druhy *C. elongatoides* , *C. taenia* a *C. elongatoides*. Schémc) je mapa středního toku Odry a jejich přítoků na území Polska, hybridní zóny mezi druhy *C. elongatoides* a *C. taenia*, body vyznačeny červeně vyznačují habitaty druhu *C. taenia*, zatímco žluté body označují druh *C. elongatoides.* Převzato z (Janko et al. 2012).

4.1 Vliv hybridizace a polyploidizace na transkripci

Polyploidie je dědičný stav vyznačující se změnou celkového stavu počtu chromosomů. Polyploidie je velmi běžná mezi rostlinami, kde se předpokládá, že přibližně 80 %

krytosemenných rostlin vzniklo duplikací, ať již z mezidruhového křížení nebo autogamií (Otto and Whitton 2000). Polyploidie je neprávem vnímána jako stav evolučně, selekčně nevýhodný přímo vedoucí pouze k extinkci, a tedy bez evolučního významu. Naopak ukázalo se, že polyploidie zejména tetraploidní stav se udál v evoluci mnohokrát a často vedl k neofunkcionalizaci genů, rapidní změně komplexity organismu, jelikož jedna z genových kopií se stala redundantní, vedl k úniku před specializovanými parazity (Weiss-Schneeweiss et al. 2013).

Hybridizace je zejména, a nejen u rostlin, velmi významný evoluční podnět. V případě kompatibility genomů, transkriptomů dochází často k jevu heterózy (nové výhodné interakce alel) a hybrid je po určitou dobu, počet reprodukčních cyklů úspěšnější než oba rodiče. Tento jev má zcela zásadní hospodářský význam (Xing and Zhang 2010). Hlavní hypotéza tázající se na efekt transgrese hybridů vidí příčinu v modifikaci genové exprese. Molekulární mechanismy podtrhující tyto změny hybridních genomů poukazují na genetickou povahu dominance parentálních alel. Obecně může dojít k nárustu přínosné gentické variability nových kombinací alel (Birchler et al. 2010).

Geny, které se projevují domiantně trangresivně, mohou být důležité pro hybrida z hlediska drastického vlivu na fenotyp (Chen 2010).

V případě triploidizace, jak se například děje vnesením třetího haplotypu do genomu diploidního hybrida při procesu gynogenetické reprodukce, mohou nastávat vážnější problémy a to především z hlediska epigenetických změn a regulace transkripce. Z toho důvodu jsou pro viabilitou takových jedinců nezbytné zásahy na úrovni genu ve formě imprintingu, hetrochromatizace jednoho z genomů a zachování původní regulace jednoho z rodičovských druhů. V opačném případě, který není ojedinělý, může triploidní jedinec získat vyšší fitness, a to nejen z důvodu energeticky méně náročné výhodnější klonální reprodukci, ale i heteróznímu efektu.

Pokud je nově polyploidní genom viabilní, nastává zásadní problém při meiotickém dělení redukční fáze – vysoká frekvence vzniku aneuploidních gamet v závislosti na typu ploidie. Situace u autopolypoidů se odlišuje v tom, že jsou schopni formovat chromozomové multivalenty v metafázi I. V anafázi I je jejich separace obtížnější a to i u tetravalentního stavu – mohou vznikat abnormální segregační profily (Comai 2005). Při tvorbě gamet anortoploidního jedince nelze získat balancovaný stav gamet žádným známým mechanismem. Produktem anafáze I jsou nejčastěji anuploidní buňky.

V případě gynogenetického rozmnožování triploidních jedinců existuje více scénářů meiózy. Mají ale společné rysy, kdy chromosomy jednoho druhu mezi sebou preferenčně párují a vytvářejí bivalenty, protože jsou si sekvenčně podobnější. Jeden z chromozomů tedy zůstává nespárovaný a v některých případech může dojít i k jeho ztrátě. Podobný systém existuje i u vyšších obratlovců druhu *Rana* (Morishima et al. 2008).

Gynogentické rozmnožování hybridního jedince P1 generace vyžaduje několik podmínek. Sesterské chromatidy nesmí být během anafáze spojeny – ztráta koheze, rekombinace – narušením funkce kinetochoru, rekombinázy ad. (Qi et al. 2006). Nesmí dojít v anafázi I k redukčnímu dělení.

Jakým způsobem hybridizace spouští klonální, gynogenetické rozmnožování u živočichů, není příliš známo. U druhu *Daphnia pulex* byly nalezeny pohlavně ovlivněné dominantní geny, které mohou stát za kontrolou vzniku neredukovaných gamet. Konkrétně se jedná o komplex nezávisle segregujících čtyř epistatických dominantních genů (Lynch et al. 2008).

5.1 Hybridizace, polyploidizace a imprinting

6.1 Metody studia transkriptomu a imprintingu

Recentní vývoj masivně paralelizovaných sekvenačních technologií posunul směr kvantitativní transkriptomiky o velký skok kupředu. Naprostá většina produkovaných dat dnešních dnů je generována převážně na principu syntézy z důvodu nejvýhodnějšího poměru ceny za sekvenovanou bázi spolu s rozumnou chybovostí (illumina), kterou lze statisticky snadno "podchytit". Jeden typ dat převažuje především díky uživatelům, kteří preferují analýzu identicky získaných dat, aplikujíce generické postupy bez domyšlení konsekvencí.

Dnes je možné sekvenovat RNA neuvěřitelných komplexit, od absolutní kvantifikace obsahu cDNA jednotlivých buněk, cDNA asociovanou s určitým typem RNA vázajících proteinů, analýza sestřihových variant, malých RNA, a další. Díky nepřeberné možnosti designů experimentů, vysoké reprodukovatelnosti a ceně, vytlačilo masivní sekvenování nové generace microarray technologii do propadliště dějin. V případě analýzy

transkriptomu není ani potřeba referenčního genomu, existuje tedy možnost pracovat i s nemodelovými organismy. Nicméně *de novo assembly* cDNA je nutno věnovat i přesto trochu pozornosti, vlastní *assembly* je dnes sice plně automatizovaný algoritmicky (převážně aplikovaný algoritmus de Bruijn grafů) a velmi pokročilý proces schopný dobrých výsledků i s *ready* okolo 50 bp porovnáním s výsledky *assembly* podle referenčního templátu genomu (Grabherr et al. 2011). Dynamický rozsah měření RNAseq (sekvenování RNA) se odvíjí především od hloubky sekvenování - kolik "prostoru" má fragment k hybridizaci na sekvenační destičku, tzn., odvíjí se od kapacity destičky a molární koncentrace fragmentu cDNA – počet *readů* / kapacita destičky.

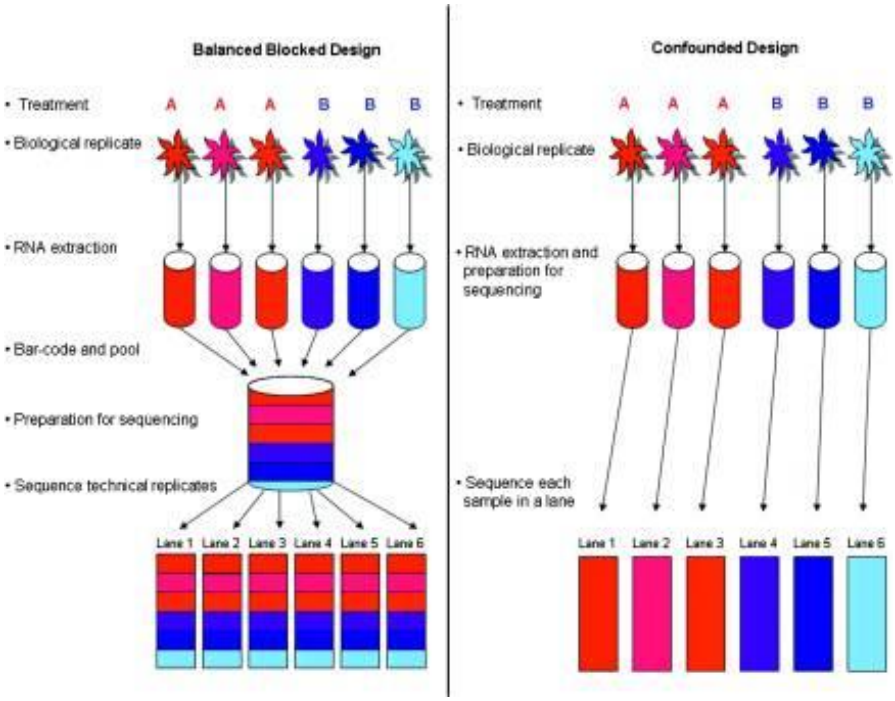
Technický proces získání transkriptomu bych shrnul krátce a obecně asi takto. Izolovaná RNA (adekvátní integrity) je tepelným šokem na odpovídající distribuce fragmentů, které jsou přepsány na základě polyA řetězce nebo náhodných hexamerů do cDNA (často se *spike* kontrolami). cDNA je opatřena sekvenčními kódy (*barcode*) tak, že na reparované konce fragmentů je přidán adenin terminální polymerázou, která zvyšuje účinnost ligace s barcode dsDNA. *Barcode* slouží k rozeznání vzorku v případě, kdy na sekvenační destičku aplikujeme směsný vzorek. Po sekvenaci je nutno vzorky rozdělit do samostatných souborů, podle *barcode* a směru sekvenace a posléze je těchto kódů zbavit. V případě poklesu kvality na koncích *readů* jsou takové báze umazány. Nyní může být provedeno *de novo assembly*, či mapování na již připravenou referenci. Sestavené cDNA je ale nutno mnoha rozdílnými přístupy kontrolovat, např. definovat transkribovaou oblast, anotovat je, predikovat nekódující RNA, identifikovat rRNA (pokud nebyla aplikována adekvátní metoda snížení kompexity RNA) a další (Wilhelm and Landry 2009). Příklady zmíněných kontrol jsou mnohdy opomíjeny a mohou vést k zcela špatným interpretacím výsledků (Lovén et al. 2012).

Bohužel i RNAseq má své limitace, na které je nutné brát při analýze a interpretaci ohled. Obsah GC se liší v sekvenci, může dramaticky měnit a způsobovat problémy při sekvenační reakci. Délka exonu ne vždy musí odpovídat délce cDNA. Při přípravě knihovny mohou vznikat PCR amplifikační artefakty. Bias může v konečné fázi způsobit i volba nevhodného softwaru pro provedení alignemntu. Některé programy – ať už využívající referenci genomu, nebo ne - mohou stanovit počet sestřihových variant a skupin, jde o matematicky i výpočetně složitý proces začínající s bipartitní grafy; cílem je najít řetězec s maximálním počtem bipartit, překrývajících se fragmentů řešením Dilworth,

či König matematickými teorémemy (Dilworth 1950). Reprezentovatelnost septřihových variant není velká, navíc je silně ovlivněná délkou sekvence (Rehrauer et al. 2013).

Další otázkou je, zda ponechat *ready mapující* se na referenci pouze parciálně – např. nemáme sestaveny kompletní sekvence cDNA, nebo *ready* s konkrétní hranicí rozdílů vůči referenci, které vykazují dobrou shodu i s jinými místy v referenci. Jednoznačná odpověď neexistuje, zde má význam spíše konzistence analýzy mezi vzorky.

Dříve než se pokusím o popis statistických metod, upozornil bych rád čtenáře bez povrchové znalosti tématu na nutnost správné volby designu experimentu. V žádném případě není akceptovatelné analyzovat směs dat rozdílných designů společně, protože statistické metody detekce DE genů se zaměřují na konkrétní design, ve většině případů blokový, neboť dnešní sekvenační platformy disponují velkou kapacitou. Schéma dvou možných RNAseq designů je znázorněno na Obr. č. 2.



Obr. č. 2: Znázornění používaných designů RNAseq experimentů; převzato z (Auer and Doerge 2010).

Prvním krokem analýzy diferenciální exprese je získání samotné informace počtu *namapovaných* readů na referenci. Je ale nutno si uvědomit, že mezi vzorky neexistuje rovnost v sekvenační hloubce a geny rozhodně nemají stejnou délku. Z tohoto důvodu je data nutno normalizovat, abychom zajistili srovnatelnost mezi knihovnami geny a sestřihovými variantami (Pickrell et al. 2010). Nejtriviálnějším přístupem, jak zajistit normalizaci vzhledem k délce genů a sekvenační hloubce, je RPKM (počet *readů* na

kilobázi na milión *mapovaných readů*). Bohužel RPKM není vhodná normalizační metoda např. pro cDNA získanou metodou náhodných hexamerů (*random primimg*), protože je částečně závislá na sekvenčním složení, a tedy ani pokrytí cDNA reference nemusí být uniformní (Hansen et al. 2010). Problémem je také mylný předpoklad homogenity mezi vzorky. Příkladem mohou být geny silně exprimované u jednoho vzorku, které "brání" sekvenování cDNA jiných nízce exprimovaných genů, přičemž situace u jiného vzorku může být jiná, jinými slovy, ačkoliv známe celkový počet *readů* v knihovně, množství celkové RNA mezi vzorky se může lišit v závislosti na složení RNA. Z toho důvodu byly vytvořeny sofistikovanější přístupy. Prvním z nich je TMM (*trimmed mean of medians*). Pro každý testovací vzorek je spočítán vážený průměr logaritmovaných poměrů mezi referencí a testem po vyloučení genů s největším poměrem reference / test. Referenční hodnotou je počet *readů* na sekvenační destičce, test značí počet *readů* ve vzorku. Slabě exprimované geny by měly mít TMM blíže 1 (Robinson and Oshlack 2010). Na podobném proncipu funguje normalizační metoda DESeq, která ale počítá normalizační faktor jako podíl geometrického průměr poměru všech vzorků a mediánu vzorku a opět získáváme poměr blížící se 1 s předpokladem, že většina genů není diferenciálně exprimovaná (Anders and Huber 2010). Ačkoliv existuje velké množství normalizačních metod, TMM a DESeq jsou nejvíce preferovány pro svou robustnost a univerzálnost.

Statistické analýzy diferenciální exprese vycházejí z předloh pro analýzu microarray dat, protože normalizovaná data obou metod vypadají velmi podobně, pokud nejsou dále transformovány. Prvním přístupem, který je dodnes součástí používaných statistických programů zaměřených na RNAseq je "nafitování" dat na poissonův model, který dobře vystihuje trend dat, nicméně není schopen příliš dobře "zachytit" biologickou varianci dat (Dillies et al. 2013). Díky tomu vnáší poissonův model více chyb I. řádu – falešně pozitivních výsledků podhodnocením variability mezi biologickými rerplikáty (Langmead et al. 2010). Recentně jsou data "fitována" na model negativně binominální, který byl získán aproximací poissonovského modelu. Pro využití statistické síly biologických replikátů byl navržen model společné disperze (Auer and Doerge 2010). Dalšími možnostmi "zpřesňování" výsledků je aplikace maximum likelyhood metod.

Nyní se zaměřme na experimentální přístup získání polymorfismů porovnáním *namapovaných readů* vůči referenční sekvenci.

expektační minimalizační algoritmus.

Cíle práce

V původním zadání své práce jsem měl za úkol otestovat, zda u klonálních sekavčích linií dochází k akumulacím škodlivých nesynonymních substitucí (např. Müllerova rohatka) a to za pomocí genomických, či transkriptomických dat. Jak se ale ukázalo, tak přípravné práce pro provedení samotné studie byly natolik komplikované a obsáhlé, že jsem de facto se svými školiteli musel řešit několik úrovní a témat zároveň. Vzhledem k získaným datům se tedy cíle mé práce zaměřují na následující témata:

1) Sestavit a anotovat věrohodný transkriptom sekavce, který bude využitelný pro následné mapování RNA *readů* mnoha jedinců.
2) Získat mapu pozic (SNP), ve kterých se jednotlivé druhy sekavců liší.
3) Využít RNAseq dat pro studium genové exprese s cílem najít loci, které mohou souviset s hybridním či polyploidizačním genomickým šokem, jakož i loci, které mohou souviset s iniciací asexuality.
4) Využít druhově specifických SNP a RNAseq dat k testování genomového umlčení, nebo imprintingu, tj. testovat, zda některé druhově specifické alelické varianty genů jsou v hybridech exprimovány více než jiné.
5) Konečně pomocí získaných SNP pozic u rodičovských druhů i různě ploidních hybridů testovat, zda u asexuálních linií dochází k vyššímu tempu nesynonymních mutací, což by mohlo nasvědčovat roli Müllerovy rohatky v evoluci klonů.

Materiál a metody

Pro analýzy transkriptomu a jednonukleotidových polymorfismů byly užity níže uvedené vzorky cDNA - tab. č.1, 2. a 3. popisující základní charakteristiku vzorků, jejich druhový, geografický původ, včetně typu sekvenování a přípravy RNA.

Vzorek	Pohlaví	Biotyp	Geografický původ	Počet readů	SNP
co01	F	*pp*	Bulgaria	43981964	164228
co02	F	*ee*	Odra R. (E4), Poland	14557443	124640
co03	F	*ss*	Zagortsi, Bulgaria	8393163	37104
co04	M	*ss*	Zagortsi, Bulgaria	3749533	37969
co05	F	*tt*	Odra R. (0903), Poland	6938768	59196
co06	F	*tt*	Odra R. (0903), Poland	4621640	37156
co07	M	*pp*	Kachul, Bulgaria	21875011	44519
co08	M	*ee*	Odra R. (E4), Poland	22168759	46167
co09	M	*nn*	Oltenitza, Danube River, Romania	30777098	99779
co10	F	*nn*	Oltenitza, Danube River, Romania	26481992	83443
cab04L	F	*tt*	NorthEastern poland	21639250	59786
cab05L	F	*tt*	NorthEastern poland	15804728	64753
cab05o	F	*tt*	NorthEastern poland	28451417	83197
cab06L	F	*tt*	NorthEastern poland	11709835	36965
cab10L	F	*tt*	NorthEastern poland	19427035	64842

Tab. č. 1: Vzroky 454 sekvenování normalizované cDNA, ze kterých byla vytvořena prvotní databáze jednonukleotidových polymorfismů a druhově specifických pozic díky srovnání několika druhů v rámci rodu *Cobitis*; vzorky končíci L, či o jsou vzorky normalizované cDNA specificky jedné tkáně, pokud toto označení chybí, jedná se o směsný vzorek normalizoavné cDNA

Vzorek	pohlaví	Biotyp	Geografický původ	Počet readů	SNP (N)	SNP (tt)	mtDNA
cab02L	F	*etn*	neznámý	55579372	46445	133532	*tt*
cab03L	F	*etn*	neznámý	40701503	45980	158026	*ee*
cab04L	F	*tt*	NorthEastern poland	16243592	32255	25782	*tt*
cab05L	F	*tt*	NorthEastern poland	27129726	47751	41246	*tt*
cab06L	F	*tt*	NorthEastern poland	32475299	48885	41929	*tt*
cab07L	F	*eet*	Polska Woda	31435509	41545	128984	*tt*
cab08L	F	*eet*	Polska Woda	39565517	43315	141146	*tt*
cab09L	F	*eet*	Polska Woda	35280898	39791	140404	*ee*
cab10L	F	*tt*	NorthEastern poland	44951055	50854	46030	*tt*
cab11L	F	*et*	Barycz	21624757	35641	84950	*tt*
cab13L	F	*een*	Polska Woda	14520839	31824	88592	*ee*
cab14L	F	*een*	Polska Woda	50005977	45409	165254	*ee*
cab15L	F	*etn*	Barycz	37778308	45866	150395	*tt*
cab16L	F	*ett*	Barycz	14540303	30924	74199	*tt*

cab17L	F	*ett*	Barycz	56825040	48248	136921	*tt*
cab18L	F	*ee*	Budkowiczanka R, Odra basin	26537273	44136	116454	*ee*
cab19L	F	*ee*	Budkowiczanka R, Odra basin	29507575	46601	131635	*ee*
cab20L	F	*ee*	Budkowiczanka R, Odra basin	31694148	44683	126245	*ee*
cab21L	F	*ee*	Budkowiczanka R, Odra basin	39984649	47557	135404	*ee*
cab22L	F	*et*	Barycz	19681257	34308	111344	*tt*
cab23L	F	*et*	Swedrnia	23467650	32273	76005	*tt*
cab24L	F	*ett*	Swedrnia	32301761	38119	104742	*tt*
cab25L	F	*ett*	Barycz	27813983	36144	90485	*tt*

Tab. č. 2: Vzorky jater cDNA sekvenování nenormalizovaných dat platformou illumina – RNAseq jater

Vzorek	**Pohlaví**	**Biotyp**	**Geografický původ**	**Poč. *readů***	**SNP (*tt*)**	**SNP (N)**	**mtDNA**
cab01o	F	*etn*	neznámý	47564562	158160	45163	*tt*
cab02o	F	*etn*	neznámý	67643876	160054	43327	*tt*
cab03o	F	*etn*	neznámý	28588893	135847	42080	*tt*
cab04o	F	*tt*	NorthEastern poland	12238231	21100	37838	*tt*
cab05o	F	*tt*	NorthEastern poland	22703041	30505	45236	*tt*
cab06o	F	*tt*	NorthEastern poland	28776800	33546	45535	*tt*
cab07o	F	*eet*	Polska Woda	40178876	174757	44243	*tt*
cab08o	F	*eet*	Polska Woda	4624141	55942	16893	*tt*
cab09o	F	*eet*	Polska Woda	25472396	164826	39787	*ee*
cab10o	F	*tt*	NorthEastern poland	30603626	39068	48032	*tt*
cab11o	F	*et*	Barycz	15764331	99653	33993	*tt*
cab13o	F	*een*	Polska Woda	47647538	203212	44892	*ee*
cab15o	F	*etn*	Barycz	29642299	139422	41208	*tt*
cab16o	F	*ett*	Barycz	19716928	119339	38996	*tt*
cab17o	F	*ett*	Barycz	24280629	124251	40271	*tt*
cab18o	F	*ee*	Budkowiczanka R, Odra basin	20188754	129733	39874	*ee*
cab19o	F	*ee*	Budkowiczanka R, Odra basin	21225612	131981	40271	*ee*
cab20o	F	*ee*	Budkowiczanka R, Odra basin	18846248	125399	38985	*ee*
cab21o	F	*ee*	Budkowiczanka R, Odra basin	20685617	147495	42029	*ee*
cab22o	F	*et*	Barycz	26201436	120212	36789	*tt*
cab23o	F	*et*	Swedrnia	19583650	93809	35014	*tt*
cab25o	F	*ett*	Barycz	27633355	117129	37552	*tt*

Tab. č. 3: Vzorky oocytů cDNA sekvenování nenormalizovaných dat platformou illumina – RNAseq oocytů

7.1 Příprava vzorků

Zdrojová tkáň pro izolaci DNA byla fixována v roztoku 99 % etanolu – ploutve a svalová tkáň, v ojedinělých případech byla zutilizována i čerstvá krev. Vzorky byly různého stáří, tedy i různého stádia degradace genomické DNA. Pro přípravu sekvenačních knihoven se ukázalo přínosné upotřebit vzorky s převládajícím obsahem nepoškozené, vysokomolekulární gDNA. Existuje totiž závislost mezi stabilitou DNA a její délkou, krátké *ready* prezentované v nadměrném množství již v iniciálním cyklu aplikace ultrazvuku (22 – 44 KHz) nedestruují tempem jako dlouhé sekvence – rozložení délek *readů* se vychyluje od normality. Vysokomolekulární gDNA je nezbytná především v případě, kdy je nutno se vyvarovat systematické chybě sonikace – jelikož preferenčně fragmentují konkrétní sekvenční motivy (Poptsova et al., 2014); tento *bias* může u nízkomolekulární DNA nabývat rozdílných intenzit. Na míru *degradace* má vliv rovněž složení roztoku, je nutné dbát na čistotu izolátu a volit vhodné rozpouštědlo.

4.1.1

4.1.2 Izolace RNA

RNA byla izolována metodou Trizol (guanidinium thiokyanátová-phenol-chloroformová extrakce).

1) Homogenát v roztoku trizolu rozmražen při RT.
2) Vortex vzorku 2 min – řádné rozpuštění žloutku oocytů.
3) Inkubace 5 min při RT.
4) Přidáno 200 µl chloroformu (200 µl chloroformu na 1 ml Trizolu).
5) Protřepáno v ruce po do15 s.
6) Následuje inkubace při RT trvající 2-3 min.
7) Vzorek je centrifugován při max. 12 000 g, 15 min, nutno chlazení rotoru na 2-8°C.
8) Horní vodná fáze s RNA je „přepipetována“ do nové čisté "eppendorfky", pozn.: vodná fáze tvoří cca 60% objemu Trizolu, interfázní vrstva s DNA a ani spodní fenolová, organická fáze nesmí kontaminovat RNA).

9) Na 1 ml Trizolu je přidáno 0,5 ml isopropanolu k vysrážení RNA (v případě izolace z minimálního množství vstupní tkáně je přidán navíc glycerol).
10) Následuje přiměřené promíchávání v ruce a vzorek je inkubován 10 min při RT.
11) Vysrážená RNA je centrifugována při max. 12 000 g, 10 min, 2-8°C.
12) Supernatant je odstraněn a k "peletce" je přidáno 1 ml 75% etanolu (1 ml etanolu na 1ml Trizolu).
13) Rozpuštění RNA "peletky" promícháním na *vortex* třepačce - 2x cca 5s.
14) Centrifugováno max. 7 500 g, 5 min, 2-8°C
15) Odebrán maximální objem supernatantu. Poté je vzorek ponechán k sušení 10-15 min na vzduchu (vakuově sušení není možné, jelikož zcela vyschlou "peletku" nelze snáze rozpustit)
16) RNA rozpuštěna v 10-20µl ddH2O (*RNase-free-water*) - dle velikosti "peletky".
17) Roztok několikrát „propipetován“ a inkubován 10 min při 57.5 °C na ledu.
18) Byly připraveny *alikvóty* pro NanoDrop, a qPCR, aby nedocházelo k rozmrazování RNA a tím k její degradaci.
19) Vzorky byly skladovány při -80°C.

4.1.3 Příprava cDNA a RNA sekvenační knihovny

cDNA byla připravena dle SMART přístupu (Zhu et al. 2001) s postupem doporučeným výrobcem (clonetech), nicméně s modifikovanými primery – CDS-T22 namísto primeru BD SMART CDS Primer II A. Přehled použitých oligonukleotidů:

SMART Oligo II oligo.	5′-AAGCAGTGGTATCAACGCAGAGTACGCrGrGrG-3′
CDS-T22 primer	5′-AAGCAGTGGTATCAACGCAGAGTTTTTGTTTTTTTCTTTTTTTTTTVN-3′
SMART PCR primer	5′-AAGCAGTGGTATCAACGCAGAGT-3′

Tab. č.: 1: Sekvence primerů užitých k přípravě cDNA reverzní transkripcí

1) Prvním krokem bylo provedení syntézy prvního vlákna cDNA ve směru 3’-5’ (primer hybridizuje s polyA sekvencí) a to s následujícími položkami kitu:
   - 0,3 µg RNA
   - 10 pmol SMART Oligo II oligonucleotide

- 10 pmol CDS-T22 primer

2) Reakční směs je zahřáta na teplotu 72 °C a prudce ochlazena na ledu po dobu 2 min.
3) Reakce syntézy cDNA prvního vlánka je iniciována přidáním hybridizačního primeru (primer-RNA) spolu s reverzní transkriptázou do celkového objemu 10 µl obsahujícího:
   - 1X First-StrAND Buffer (50 mM Tris-HCl (pH 8.3); 75 mM KCl; 6 mM MgCl2)
   - 2 mM DTT
   - 4 mM eqvimolárního roztoku dNTP
4) Syntéza prvního vlákna probíhá při 42 °C po dobu 2 h. Po ukončení reakce je směs ochlazena na ledu.
5) Pro přípravu dvouvláknové cDNA je první vlákno zředěno 5x v TE pufru, roztok je posléze inkubován 7 min při 70 °C k preparaci amplifikace dlouhých *readů* DNA. Reakční směr pro Long-Distance PCR (50 µl) obsahuje níže uvedené složky:
   - 1 µl cDNA
   - 1x Encyclo reaction buffer (Evrogen)
   - 200 uM dNTPs
   - 0.3 uM SMART PCR primer
   - 1 x Encyclo polymerase mix (Evrogen)

   Bylo provedeno 18 PCR cyklů tohoto nastavení:

Iniciální denaturace:	95 °C	2 min
Denaturace:	95 °C	10 s
Hybridizace:	65 °C	30 s
Extense	72 °C	3 min
Finální Extense:	72 °C	5 min

6) RNA sekvenační knihovna byla připravena a sekvenována v servisním oddělení EMBL Genomics Core Facilities. RNAseq knihovny byly sekvenovány na instrumentu HiSeq 2000, 50 pair-end (PE).

4.1.4 Příprava normalizované cDNA a normlizované RNAseq knihovny

V iniciálním kroku je cDNA hybridizována (míra abundance vzniknuvší dsDNA koreluje s rychlostí odbourávání dané dsDNA DNS (duplex specifickou nukleázou)) – tohoto faktu je v následném kroku využito k normalizaci (Zhulidov et al. 2005). Princip graficky znázorněn ve schématu č. 1. RNA integrita vztažena k 18S a 28S *peaku* byla testována Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA), RIN hodnoty byly u všech vzorků vyšší než 7,5.

1) Hybridizační směs obsahuje níže uvedené komponenty:
   - 3 µl (150 ng) purifikované cDNA
   - 1 ul 4x Hybridization Buffer (200 mM HEPES-HCl, pH 8.0; 2 M NaCl)
2) Reakční směs je překryta minerálním olejem a inkubována při 98 °C po dobu 3 min, po uplynutí denaturačního procesu je cDNA hybridizována při 68 °C 5 h.
3) Do hybridizační reakce je přidáno:
   - 5 µl 2x DNase Buffer (100 mM Tris-HCl, pH 8.0; 10 mM MgCl2, 2 mM DTT)
   - 1 ul DSN enzyme
4) Inkubace s DNasou probíhá 20 min při 67 °C. Pro inaktivaci enzymu je do reakce přidáno 10 µl 5mM EDTA
5) Po normalizaci je cDNA zředěna 20 µl ddH_2O a je připravena PCR reakční směs (50 µl):
   - 1µl zředěné cDNA
   - 1 x Encyclo reaction buffer (Evrogen)
   - 200 uM dNTPs
   - 0.3 uM SMART PCR primer
   - 1 x Encyclo polymerase mix (Evrogen)
6) Reakční směs podstupuje 18 PCR cyklů za těchto podmínek:

Denaturace	95 °C 7s
Hybridizace	65 °C 20s
Extenze	72 °C 3 min

7) 1 µg normalizované cDNA byl zpracován v servisním středisku Genomics Core Facilities EMBL k přípravě sekvenační knihovny dle protokolu Library Preparation Protocol (Roche).

8) Knihovny jedinců byly značeny MID adaptéry (Roche) a byl vytvořen tzv. „pool". Počet jedinců byl zvolen dle požadované sekvenační hloubky – 2 jedinci na „flowcell lane" sekvenační platformy GS FLX+ (454 Life Sciences, Roche).

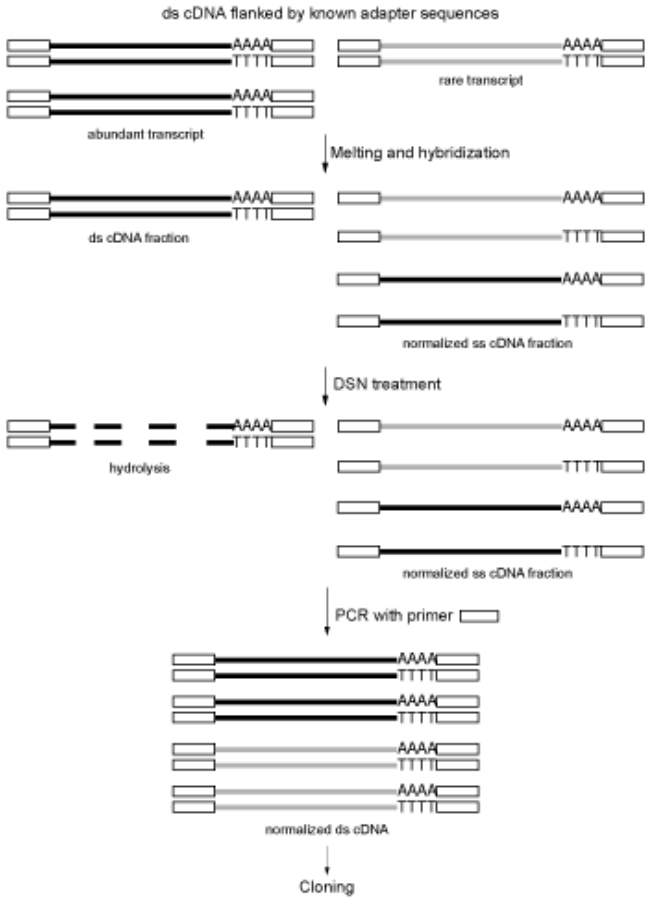


Schéma č. 1: Znázornění principu cDNA normalizace (zrdoj: http://www.evrogen.com/technologies/normalization.shtml)

8.1 Kompozice referenční sekvence – transkriptomu

Hlavním požadavkem na referenční sekvenci je obsáhnout maximální množství exprimovaných genů, jejich sestřihových variant a to nejvyšší možné délky. Z tohoto důvodu byly zvoleny tkáně, orgány ontogeneticky značně různorodé – oocyty v 6. vývojovém stádiu a játra – pouze samice. Cílem normalizace cDNA je navíc získání transkriptů byť i minimální kvantity v transkriptomu a to při zachování rozumné sekvenační hloubky.

Referenční sekvence transkriptomu byla sestavena pouze z jednoho druhu – *Cobitis taenia*, protože se ukázalo, že vnesením sekvenční variability smícháním druhů způsobuje problémy se správným složením cDNA, které následně neodpovídá realitě ani jednoho druhu. Frekventovaně tímto způsobem vznikají tzv. pseudoparalogní cDNA sekvence - homologní geny vykazující znaky paralogů – zdají se být výsledkem duplikace ancestrálního genu, nicméně v žádném stádiu vývoje nedošlo k jejich duplikaci (Koonin 2005). V tomto případě nehledejme příčinu v horizontálním genovém přenosu. Mapováním *readů* na referenci obsahující *in silico* vzniklé pseudoparalogy dochází k situaci, kdy jednotlivé druhy mapují své *ready* na rozličné pozice, tímto pseudoparalogní geny navíc zdánlivě zvyšují polymorfismus takových *loci*.

Ready získané sekvenací normalizovaných cDNA knihoven šesti jedinců *C. taenia* (viz tab. č. 4) byly zbaveny sekvencí a konců s nízkou kvalitou pomocí Trimmomatic softwaru (Bolger et al. 2014). Technické PCR multiplikáty byly odstraněny z datasetu *readů* utilizací cdhit-454softwaru (Li and Godzik 2006). Výsledných 1886536 *readů* – 648620753 párů bazí bylo selektováno k tvorbě cDNA assembly – referenční sekvence (provedeno: Mgr. Jan Pačeš, Ph.D.).

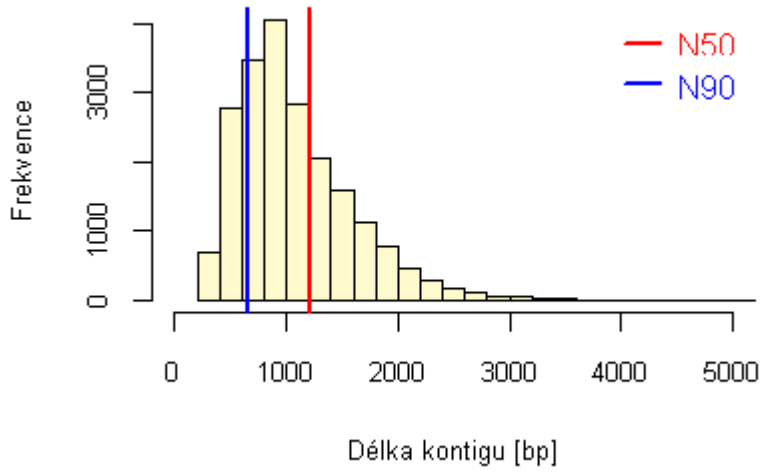
Jedinec	**Typ tkáně**	**Knihovna**	**Počet bazí**	**Počet *readů***
co05	oocyty	co05	97492888	364919
co06	oocyty	co06	66110724	243092
cab04	játra	cab04L	103716091	285342
cab05	oocyty	cab05o	150231825	374684
	játra	cab05L	91235506	208142
cab06	játra	cab06L	54725340	154445
cab10	játra	cab10L	96836134	255912

Tab. č.: 4: Tabulka jedinců tvořící referenční cDNA transkriptom se základní deskripcí

4.1.5 *Assembly* transkriptomu

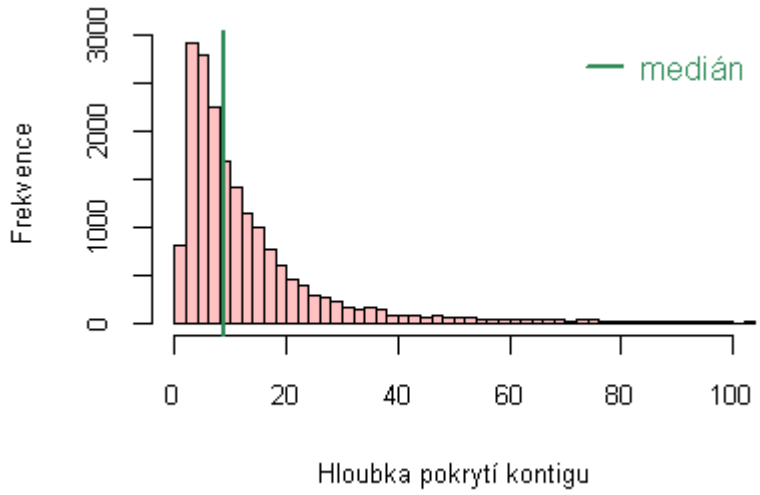
Vlastní *assembly* bylo provedeno Newbler softwarem (verze: 2.6 20110517_1502, Roche) modifikovaným pro analýzu 454, tak illumia datových setů, s parametry: 20 bp minimální délka *readu*, minimální překryv *readů* 40 bp, minimální identita překrývajících se *readů* 90 %, minimální délka všech sekvencí, minimální délka dlouhých sekvencí 300 bp, automatické odstranění vektorových sekvencí.

První verze *assembly* se skládá z 29333 cDNA sekvencí (včetně příslušných *in silico* generovaných alternativních sestřihových variant) s počtem bazí 32540805 a N50 [1] rovno 1291 viz tab. č. 4, přičemž 95.39 % sekvenovaných bazí má kvalitu vyjádřenou *phred score* (Q) větší než 40. Distribuce délek cDNA verze lom300tt_v5 je graficky vyjádřena v grafu č. 1. Sekvenační hloubka pokrytí cDNA sekvencí je znázorněna v graf č. 2 (*assembly* provedeno s pomocí: Mgr. Jan Pačes, Ph.D.).



Graf č. 1: Histogram distribuce délek cDNA ve finální verzi referenčních sekvencí

[1] [1] N50 značí délku nejkratšího kontigu definovaného jako sumu všech kontigů větší nebo rovno 50% celkové délky všech kontigů.



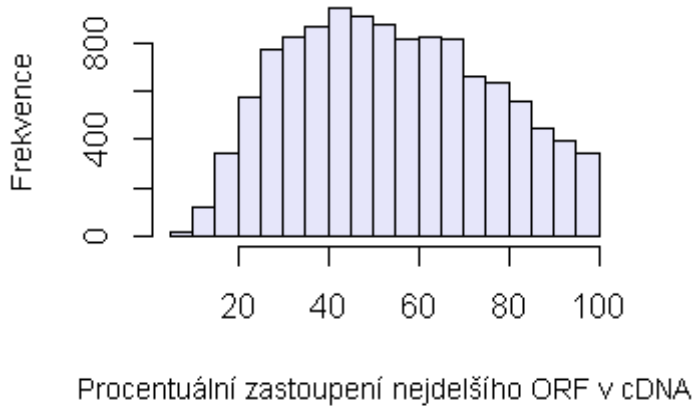
Graf č.: 2: Histogram hloubky pokrytí cDNA sekvencí; medián = 9 hloubka/kontig

4.1.6 Validace a dodatečné korekce transkriptomu

Reference musela být v několika postupných krocích vyčištěna, neboť byly artificiálně vygenerovány duplikáty cDNA (tzn. geny jsou polymorfní, identický gen vznikne několikrát). Pro navazující analýzy s cílem odstranit problematické cDNA bylo nutno zbavit se nebetyčných skupin sekvencí chovajících se jako pseudoparalogy, skupin sestřihových variant v jednotlivých generovaných sestřihových skupinách, rozdělit chimérické genů vzniklých spojením skrze sekvenčně parciálně homologické sekvence. Díky těmto krokům byly postupně získáno 5 verzí referenční transkripce. Basální popis změn v transkriptomu je uveden v tab. č. 5.

1) Z první verze transkriptomu byly odstraněny cDNA, u kterých docházelo k „nasedání“ *readů* na více *loci* totožné cDNA i jiných cDNA sekvencí – i po několikanásobném mapování stejného setu *readů* byly generovány nové polymorfní pozice vzhledem k referenční sekvenci.
2) Z druhé verze byly odstraněny sekvence, které při blastn analýze (*Basic Local Alignment Search Tool*) vůči své sekvenci byly obsaženy s téměř 100 % identitou ve více než jednom kontigu (cDNA) – detekce duplicitních cDNA (netýká se splice variant).

3) V transkriptomu byly ponechány pouze nejdelší sestřihové varianty z jednotlivých sestřihových skupin.
4) Zbylé sekvence kratší nežli 300 bazí, které prošly během procesu *assembly,* byly odstraněny.
5) Při anotaci cDNA sekvencí se ukázalo, že v procesu *assembly* došlo u 1432 cDNA sekvencí ke spojení dvou a více různých genů, jelikož tato část sekvencí získala k vícero separovaných blastx *hits* - *alignmentů* s nejlepším skóre přesahující limit e-value 0,001 - a to různých GI (jedinečných identifikátorů sekvencí v databázi). Nicméně část těchto cDNA nemusí být chimérických, identifikátory mohou odpovídat sekvenčně blízce příbuzným homologním sekvencím. Referenční cDNA s podezřením na vznik chimérického genu byly rozděleny v polovině mezi rozdílnými geny - na základě úzu pozičního rozsahu GI. Pokud nedošlo k fúzi např. přes UTR regiony cDNA, ale přes homologní konzervované regiony genových rodin nebylo možno takového sekvence rozdělit. Z reference bylo následně také 180 *kontigů* odstraněno, jelikož po separaci nedosáhly limitu 300 bp, rovněž bylo odstraněno 698 původních chimérických *kontigů*.
6) Finální verze prošla navíc změnou orientace *antisence* cDNA sekvencí: 3‘-5‘ do 5‘-3‘ *sence* orientace programem revseq (emboss, ver: 6.6.0.0). *Antisence* cDNA sekvence byly identifikovány podle orientace nejdelšího ORF (*open reading frame*) mezi stop kodóny tvořícího min. 20 % celkové délky cDNA (ORF detekovány programem getorf (emboss, ver: 6.6.0.0)) a orientace blastx na základě pozice alignmentů viz subkapitola 4.2.3. U 1293 cDNA se orientace bezi blastx a ORF neshoduje, u této cDNA byla orientace determinována pouze na základě blastx, neboť jistá část cDNA neobsahuje dostatečně dlouhý ORF, viz distribuce procentuálního obsahu ORF v cDNA graf č. 3, tudíž 20% limit nemusí být dostatečný. Takovýto ORF může být čistě sporadicky nejdelší v opačné orientaci, pokud dojde během *assembly* k posunutí ORF. V orientaci 3‘-5‘ bylo identifikováno **5583**, v 5‘-3‘ **6016** a bez determinované orientace zůstalo **9003** cDNA sekvencí.
7) Posledním způsobem validace bylo vygenerování řady funkčních primerových párů a to jak pro amplifikaci sekvencí z cDNA, tak i gDNA (housekeeping geny a diferenciálně exprimované geny pro validaci RNAseq experimentu; oliga obsahující morfolino k inhibici translace ad.)

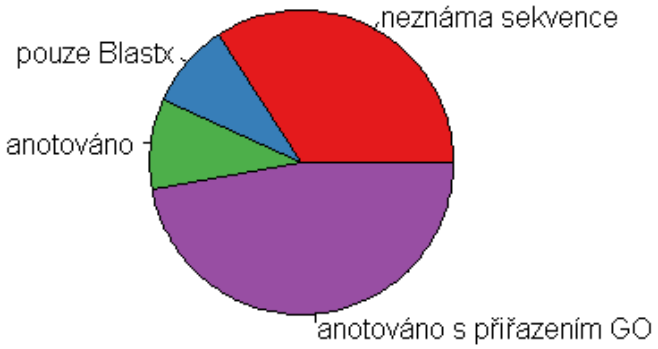


Graf č.: 3: Histogram distribuce procentuálního zastoupení nejdelšího ORF v cDNA sekvenci s mediánem 52,96.

verze	počet sekvencí	průměrná délka	celková délka	N50	N90
lom300tt_v1	29333	1109.4	32540805	1291	648
lom300tt_v2	22176	1094.5	24271233	1249	657
lom300tt_v3	20385	1096.5	22355325	1246	661
lom300tt_v4	20047	1111.6	22283778	1249	668
lom300tt_v5	20601	1079.9	22246219	1218	651

Tab. č. 5: Deskriptivní statistika jednotlivých setů referenčních sekvencí

4.1.7 Anotace cDNA transkriptomu



Graf č. 4.: Četnost jednotlivých kategorií, do nichž byly přiřazeny cDNA – rozlišené na základě existence signifikantního alignmentu, cDNA se signifikantními hity, leč bez anotace a anotovaných cDNA a to s a bez „váhově“ signifikantních GO term.

Finální verze referenčního transkriptomu byla anotována podle předpokládaných nalezených homologních proteinů lokálním *alignmentem* (blastx) proti neredundantní

proteinové databázi (nr – 12. 1. 2015) s prahem e-value 0,001. Na základě maximálního počtu 20 přiřazených *alignmentů* (menší než prahová hodnota e-value) blastx byla provedena programem Blast2GO (Conesa et al. 2005) vlastní anotace včetně přiřazení GO (*gene ontology*) a KEGG (KEGG: *Kyoto Encyclopedia of Genes AND Genomes*).

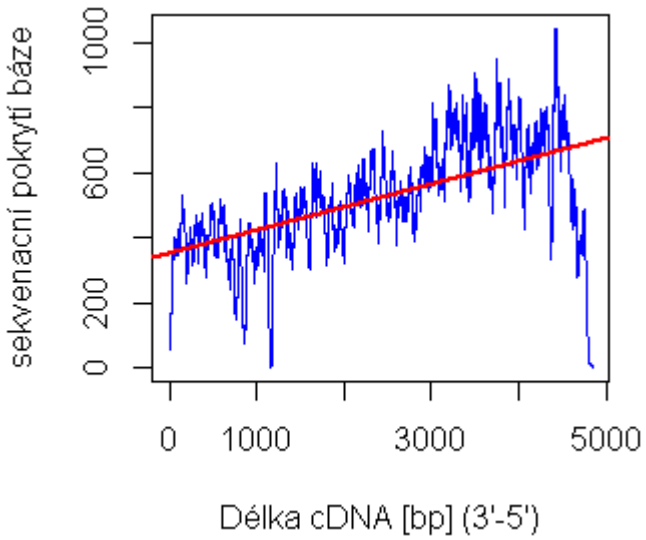
Přibližně třetina cDNA sekvencí nebyla anotována, viz graf č. 4. Primárním důvodem tak vysokého počtu genů bez anotace je to, že fylogeneticky nejbližší kompletně anotovaný genom náleží druhu *Danio rerio* patřícího rovněž do řádu *Cyprinoformes*, nicméně jejich společný předek je starý přibližně 120 mil. let. (Nakatani et al. 2011).

9.1 Mapování RNAseq sekvencí

Při mapování SE 50 bp *readů* RNAseq dat (HighSeq 2000) a PE *readů* (2x75 – NextSeq & 2x100 - HighSeq 2000) bylo přistoupeno k několika rozdílným algoritmickým přístupům: BWA (verze 0.7.12-r1039) (*Burrow wheel alignment*) (Li AND Durbin, 2009), gsnap (verze 2012-07-20) (Wu and Nacu 2010), novoalign (V3.01.02) (Ruffalo et al. 2012) a Mosaik (verze 2014-03-26) (Lee et al. 2014b). Mosaik se jeví jako vhodný nástroj k detekci SNP vzhledem k preciznosti mapování. Nicméně Mosaik systematicky opomíjí frakci *readů* mapujících se pouze parciálně. Veškeré programy byly spuštěny pouze se základními (*default*) parametry.

„Single-end“ ready o délce 50 bp byly mapovány programem bowtie2 (základní nastavení) pro získání počtu *readů* na cDNA, pro detekci SNP bylo zvoleno mapování programem mosaik. *Ready* byly zbaveny již adaptérových sekvencí (EMBL, genecore servisní centrum).

Nenamapované *ready* zřejmě pochází z UTR sekvencí, či se ve většině případů jedná o *ready*, již postrádají pozice referenční cDNA, které nebyly *assemblovány* díky nízkému pokrytí; tento problém se pochopitelně projevuje také díky nedokonalosti reverzní transkripce, především pak v genech, kterou jsou málo transkripčně aktivní (absence pozic směrem k 3’ konci cDNA viz kapitola 4.4.2, graf. č. 5).

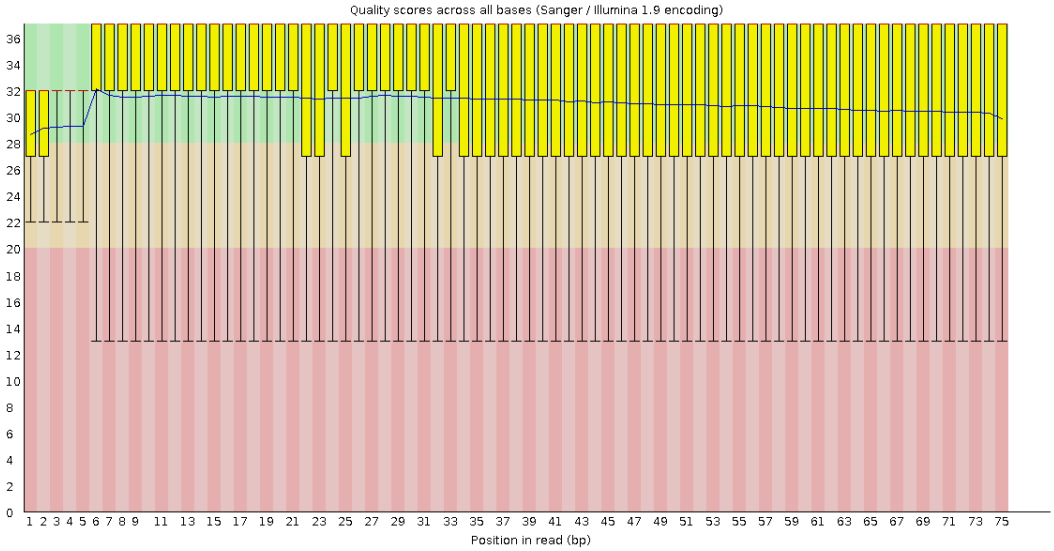


Graf č. 5: Sekvenační hloubka jednotlivých bází napříč cDNA kódující gen komplementu C4-2 s délkou 4843 bp; (vzorek nemusí být reprezentativní pro celý dataset; podobný trend lze ale pozorovat i u mnoha jiných náhodně vybraných cDNA). Červená přimka je znázorněním „nafitovaného“ lineárního modelu.

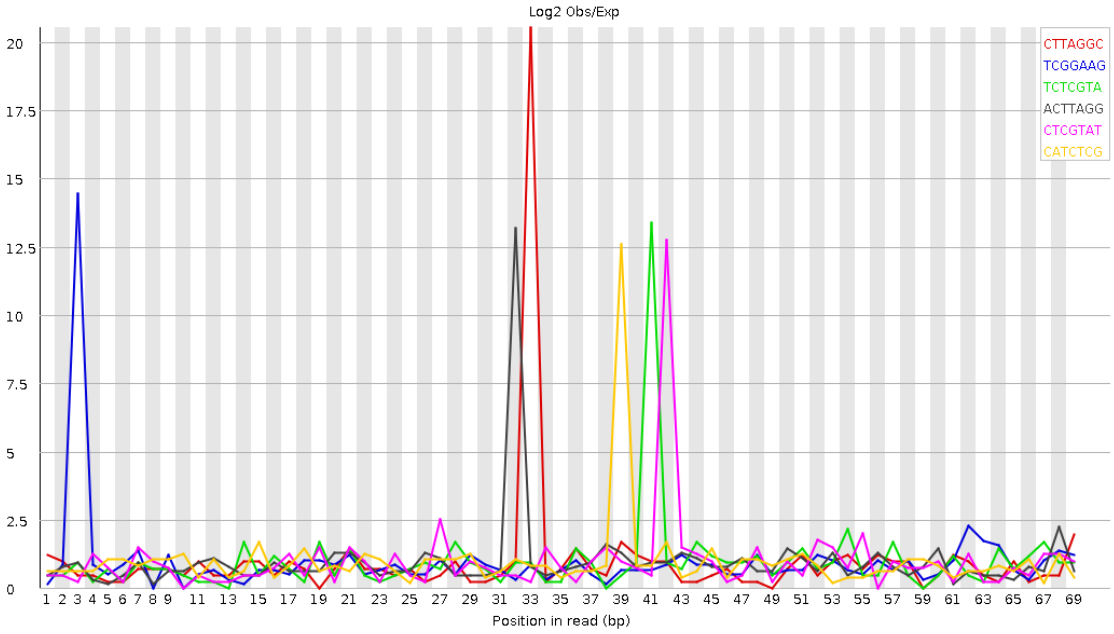
10.1 Kontrola kvality sekvenačních dat

Ke kontrole kvality *readů* RNAseq sekvenančích dat byl aplikován software fastqQC (Trivedi et al. 2014). V případě vygenerovaných sekvenačních dat obecně není problém s poklesem kvality (phred skóre) směrem ke 3’ koncům *readů*, na základě těchto zjištění nebyly konce *readů* s nízkou kvalitou odstraněny (v případě značného poklesu pred skóre by se snížila úspěšnost mapování, sic *ready* s nadlimitním počtem *missmatch* pozic jsou eliminovány), viz graf. č. 6. Jediné problémy sekvenovaných fragmentů činí duplikované sekvence a výskyt nabohacených Kmer v rámci *readů* (graf č. 7). Bohužel v obou případech těchto zjištění nelze jednoznačně určit primární příčinu prezence v sekvenci – neexistuje způsob, jak odfiltrovat PCR artefakty od biologických duplikátů (náhodná selekce identických duplikátů různých sekvenčních kopií). Artificiální duplikáty v případě RNAseq jsou v této kvantitě běžné, jedná se o nadexprimované geny, viz graf č. 8. V RNAseq *readech* se vyskytl další problém a to s *disbalancovaným* zastoupením frekvence bazí při začátcích syntézy *readů*. Naštěstí takovýto *bias* neafektuje výsledek expresí; příčinou *biasu* poziční kompozice bazí může být nenáhodná fragmentace (způsobenátepelným šokem v případě našich dat), selekce *randomizačních* primerů (nebyly aplikovány) nebo chybný, příliš horlivý zásah bioinformatika při odstranění sekvencí

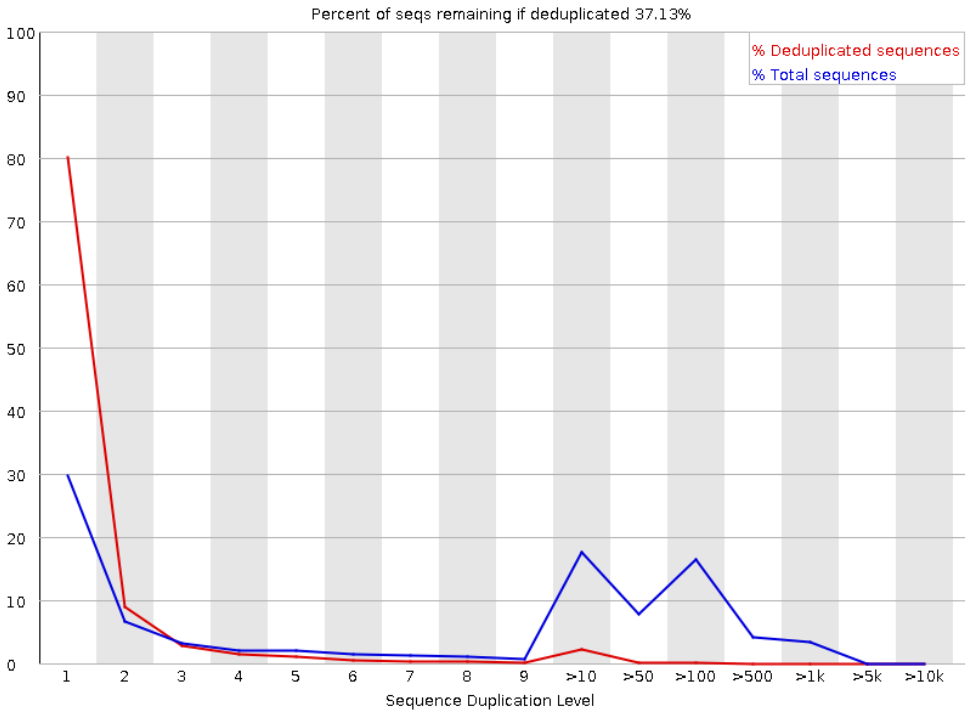
adaptéru fragmentu.



Graf č. 6: Boxplot zobrazení distribuce Q pro každou bázi v souborů PE *readů* (linka zobrazuje průměrnou Q).



Graf č. 7: Exspektance nabohacených Kmer sekvencí v rámci souboru *readů* (detekce hypergeom. testem)



Graf č. 8: Znázornění míry duplikovaných sekvencí v rámci souborů *readů* RNAseq vzorku jater (počet neunikátních sekvencí by neměl překročit 20 %)

11.1 Analýza polymorfismů 454 normalizovaných RNAseq dat

Východiskem k analýze druhově specifických SNP byla data pocházející ze 454 sekvenování normalizované cDNA. Z výstupních dat mapovaných na prvotní referenci pomocí GS Reference Mapper 3.0 (20140129_1709) byly z *high confidency* tabulek (HCD) extrahovány SNP pozice. Cely komentovaný postup je znázorněn v schématu č. 3. Použité skripty užité k získání SNP, jejich úpravu je možné si prohlédnout příloze. Níže uvedený postup schématu č. 2 popisuje získání a uložení informací do relační databáze jedno nukleotidových polymorfismů z HCD tabulek.

```
#!/bin/bash
# Získání SNP (cDNA, pozice, varianta, hloubka) z HCD výstupní tabulky
# newbler mapperu
cat cab04_map1/mapping/454HCDiffs.txt | grep ">" | sed -e "s/>//g" | sed
-e "s/%//g" > cab04.tsv
# Vytvoření tabulek v relační databázi MySQL s následným vložením dat
for i in $(cat 454_tables.txt)
   do
      mysql -e "CREATE TABLE ${i}_hcd10i (ref_acc char(16), pos_start
      int, pos_end int, ref_seq varchar(255), var_seq varchar(255), dep
      smallint, freq tinyint, KEY (ref_acc,pos_start));
      mysql -e "load data local infile
      "/mnt/raid1/Projects/Cobitis/ws_jenda/Mapping_vs_lom300tt/step3/${i
      }.tsv" INTO TABLE '$i'_hcd10i ignore 2 lines;"
# Příprava finální tabulky do níž budou vloženy finální
```

```
        mysql -e "CREATE TABLE $i_hcd10f (ref_acc char(16), pos_start
        int, pos_end int, ref_seq char(1), var_seq char(1), dep smallint,
        freq tinyint, KEY(ref_acc, pos_start));"
done
```

Schéma č. 2: Popis získání SNP z výstupních tabulek HCD a vložení do relační databáze MySQL

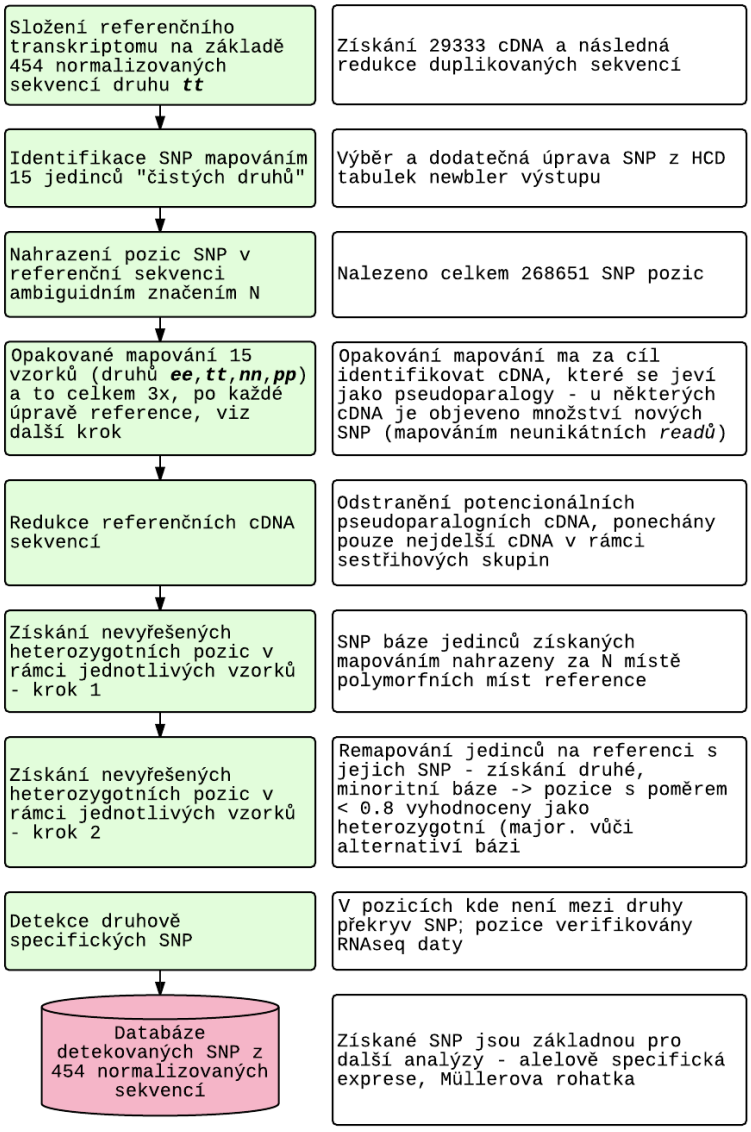
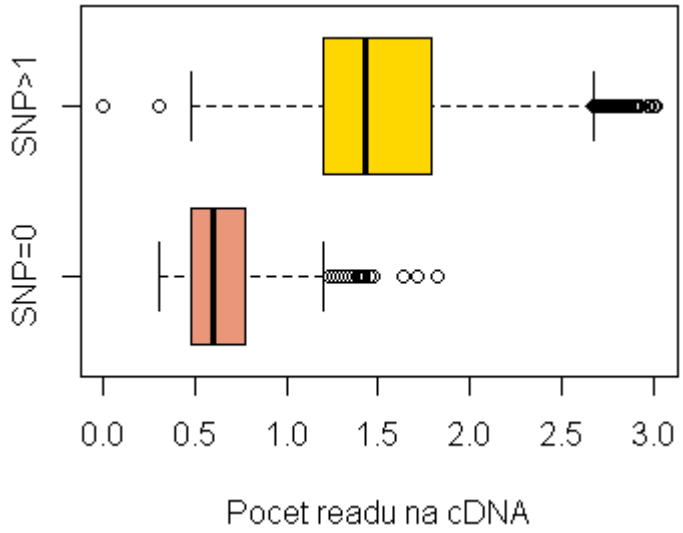


Schéma č. 3: Postup získání SNP z datasetu 454 normalizovaných dat

Do finální taulky jsou nahrány pouze rozdělené, rozřešené polymorfismy, vyjma *vyfiltrových* variant, které byly řešeny individuálními skripty. Hlavní překážkou bylo získání heterozygotních pozic, jelikož newbler této verze detekuje pouze majoritní pozici, či „balancovaného heterozygota“ blízkého poměru alternativních bazí 1:1. Newbler také neumí řešit problémy s vícenásobnými heterozygoty, více-násobnásobnými záznamy

polymorfismů stejných koordinát (rozuměj cDNA a pozice SNP); subsidiární programy (autor: Mgr. Jan Pačes, Ph.D.) Řešení těchto komplikací jsou uvedena v příloze.

Vážným nedostatkem 454 dat je sekvenační hloubka, která má za následek razantní snížení počtu nalezených polymorfismů, viz graf. č. 9. Z grafu č. 11 je řetelné, že hloubka sekvenování 454 normalizované RNA nebyla dostatečná pro identifikaci velké části prezentovaných polymorfismů daných jedinců (limitem pro detekci SNP je nutná přítomnost alespoň 5 namapovaných bazí s Q >= 20). Nízká sekvenační houbka je také spojena se zkrácením 3' konců rererenční sekvence viz kapitola č. I přes tyto problémy bylo identifikováno 268651 SNP.



Graf č. 9: Srovnání distribuce počtu *readů* na cDNA mezi datasety: počet *readů* na kontig bez nalezených SNP a počet *readů* na kontig obsahující více než 1 SNP. Data jsou transformovány log10+1.

4.1.8 Označení detekovaných polymorfismů v transkriptomu

Finální tabulky generovaných pozic polymorfismů a jejich frekvencí z HCD (*high confidency table*) 454 dat výše zmíněných druhů Sekavcovitých ryb byly spojeny do jedné tabulky na základě názvu cDNA a pozice SNP, kdy jedinci, potažmo vzorky zaujímají pozici ve sloupci. MySQL *query* (union všech řádků pozic cDNA a následný LEFT JOIN pro získání všech pozic napříč vzorky; autor: Mgr. Jan Pačes, Ph.D.) – schéma č. 6

```
CREATE TABLE var15_hcd18f_N
SELECT ref_acc, pos FROM (
   SELECT ref_acc, pos FROM cab04_hcd18f
   union SELECT ref_acc, pos FROM cab05L_hcd18f
   union SELECT ref_acc, pos FROM cab05o_hcd18f
   union SELECT ref_acc, pos FROM cab06_hcd18f
   union SELECT ref_acc, pos FROM cab10_hcd18f
   union SELECT ref_acc, pos FROM co01_hcd18f
   union SELECT ref_acc, pos FROM co02_hcd18f
   union SELECT ref_acc, pos FROM co03_hcd18f
```

```
   union SELECT ref_acc, pos FROM co04_hcd18f
   union SELECT ref_acc, pos FROM co05_hcd18f
   union SELECT ref_acc, pos FROM co06_hcd18f
   union SELECT ref_acc, pos FROM co07_hcd18f
   union SELECT ref_acc, pos FROM co08_hcd18f
   union SELECT ref_acc, pos FROM co09_hcd18f
   union SELECT ref_acc, pos FROM co10_hcd18f
   union SELECT ref_acc, pos FROM var15_hcd17x_N
  ) as a
GROUP BY ref_acc, pos;

-- Doplnění NULL v pozicích bez informace u ostatních vzorků
CREATE TABLE var15_hcd18f_pos
SELECT ref_acc, pos as pos,
cab04.var_seq AS cab04,
cab05L.var_seq AS cab05L,
cab05o.var_seq AS cab05o,
cab06.var_seq AS cab06,
cab10.var_seq AS cab10,
co01.var_seq AS co01,
co02.var_seq AS co02,
co03.var_seq AS co03,
co04.var_seq AS co04,
co05.var_seq AS co05,
co06.var_seq AS co06,
co07.var_seq AS co07,
co08.var_seq AS co08,
co09.var_seq AS co09,
co10.var_seq AS co10
FROM var15_hcd18f_N
LEFT JOIN cab04_hcd18f AS cab04 USING (ref_acc, pos)
LEFT JOIN cab05L_hcd18f AS cab05L USING (ref_acc, pos)
LEFT JOIN cab05o_hcd18f AS cab05o USING (ref_acc, pos)
LEFT JOIN cab06_hcd18f AS cab06 USING (ref_acc, pos)
LEFT JOIN cab10_hcd18f AS cab10 USING (ref_acc, pos)
LEFT JOIN co01_hcd18f AS co01 USING (ref_acc, pos)
LEFT JOIN co02_hcd18f AS co02 USING (ref_acc, pos)
LEFT JOIN co03_hcd18f AS co03 USING (ref_acc, pos)
LEFT JOIN co04_hcd18f AS co04 USING (ref_acc, pos)
LEFT JOIN co05_hcd18f AS co05 USING (ref_acc, pos)
LEFT JOIN co06_hcd18f AS co06 USING (ref_acc, pos)
LEFT JOIN co07_hcd18f AS co07 USING (ref_acc, pos)
LEFT JOIN co08_hcd18f AS co08 USING (ref_acc, pos)
LEFT JOIN co09_hcd18f AS co09 USING (ref_acc, pos)
LEFT JOIN co10_hcd18f AS co10 USING (ref_acc, pos)
;
```

Schéma č. 6: *Outer join* všech SNP ze 454 sekvenované, normalizované cDNA v jednu databázovou tabulku

Finální tabulka polymorfismů tohoto formátu – SNP vzorků uvedeny ve sloupcích s primárním klíčem ve formě názvu cDNA a pozice SNP (proces získání polymorfismů je vícestupňový, viz schéma č. 3; popis je uveden abstraktně a to z důvodu rozsáhlé úpravy reference, což se projevuje postupným úbytkem SNP, především pak redukcí problematických pseudoparalogních cDNA) byla převedena do binárního formátu a to následujícím způsobem – cílem je spojit tabulky variantních SNP pozic s vytvořenou tabulkou obsahující pouze identické informace o pozici SNP. Báze je v binární tabulce

nyní zapsána logicky v matematickém slova smyslu (ano, ne). Každý jedinec přispívá do souboru možných polymorfismů na dané pozici, tím, že je umožněn jejich přepis a to ve směru není přítomen >> je přítomen (0 >> 1), ale nikdy ne v opačném směru, viz schéma č. 7.

```
#!/bin/bash
# Transformace IUPAC bazí do binárního formátu SNP všech jedinců
# 454 normalizovaných dat
while read i
   do
mysql -e "var15_species (ref_acc, pos) SELECT ref_acc, pos FROM
var15_454_pos;"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=0,tt_c=0,tt_g=0,tt_t=0 WHERE $i IS NULL;"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=1,tt_c=0,tt_g=0,tt_t=0 WHERE $i ='A' AND $i !='C' AND $i !='G' AND
$i !='T';"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=0,tt_c=1,tt_g=0,tt_t=0 WHERE $i !='A' AND $i ='C' AND $i !='G' AND
$i !='T';"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=0,tt_c=0,tt_g=1,tt_t=0 WHERE $i !='A' AND $i !='C' AND $i ='G' AND
$i !='T';"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=0,tt_c=0,tt_g=0,tt_t=1 WHERE $i !='A' AND $i !='C' AND $i !='G' AND
$i ='T';"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=1,tt_c=0,tt_g=1,tt_t=0 WHERE $i !='A' AND $i !='C' AND $i !='G' AND
$i !='T' AND $i = 'R';"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=0,tt_c=1,tt_g=0,tt_t=1 WHERE $i !='A' AND $i !='C' AND $i !='G' AND
$i !='T' AND $i = 'Y';"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=0,tt_c=1,tt_g=1,tt_t=0 WHERE $i !='A' AND $i !='C' AND $i !='G' AND
$i !='T' AND $i = 'S';"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=1,tt_c=0,tt_g=0,tt_t=1 WHERE $i !='A' AND $i !='C' AND $i !='G' AND
$i !='T' AND $i = 'W';"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=0,tt_c=0,tt_g=1,tt_t=1 WHERE $i !='A' AND $i !='C' AND $i !='G' AND
$i !='T' AND $i = 'K';"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=1,tt_c=1,tt_g=0,tt_t=0 WHERE $i !='A' AND $i !='C' AND $i !='G' AND
$i !='T' AND $i = 'M';"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=0,tt_c=1,tt_g=1,tt_t=1 WHERE $i !='A' AND $i !='C' AND $i !='G' AND
$i !='T' AND $i = 'B';"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=1,tt_c=0,tt_g=1,tt_t=1 WHERE $i !='A' AND $i !='C' AND $i !='G' AND
$i !='T' AND $i = 'D';"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=1,tt_c=1,tt_g=0,tt_t=1 WHERE $i !='A' AND $i !='C' AND $i !='G' AND
$i !='T' AND $i = 'H';"
"UPDATE var15_species JOIN var15_454_pos USING (ref_acc) SET
tt_a=1,tt_c=1,tt_g=1,tt_t=0 WHERE $i !='A' AND $i !='C' AND $i !='G' AND
$i !='T' AND $i = 'V';"
done < 454_animals.txt
```

Schéma č. 7: Postup převodu SNP do binárního formátu.

4.1.9 Postup získání heterozygotních SNP

V případě 454 sekvenačních dat bylo získání heterozygotních pozic problematické a to z důvodu, že HCD tabulky obsahují záznam pouze majoritní alely, nebo již vyhodnocené heterozygotní pozice. Z toho důvody byl navržen následující způsob řešení (autor: Mgr. Jan Pačes, Ph.D.). Po iniciálním namapování na N referenci byla získána informace o majoritní alele (v místech *ambiguidnch* N pozic), která byla zapsána do referenční sekvence. Tedy ze SNP jednotlivých jedinců byly vytvořeny referenční sekvence obsahující pouze jejich konkrétní set jednonukleotidových polymorfismů. V dalším kroku byli jedinci mapováni na své reference se zapsanou majoritní bazí heterozygotní alely. Díky tomu je možné extrahovat druhou bázi a na základě arbitrárního poměru 0.2 - 0,8 (majoritní / suma počtu všech ostatních variant) stanovit heterozygota v dané cDNA a pozici. Skript určený pro zápis první namapované alely namísto N pozice v referenčním transkriptomu je uveden v příloze (autor: Mgr. Jan Pačes, Ph.D).

12.1 Analýza SNP RNAseq v pozicích identifikovaných 454 daty

Detekce SNP ze sekvencí RNAseq namapovaných programem mosaik byla provedena „manuálně“ (mapování GSmapper je příliš výpočetně náročné), neboť v případě, kdy je do reference záměrně vnesena *ambiguidní* báze N, nelze obvyklými software pro detekci SNP (samtools/bcftools/vcftools; samtools/GATK; samtools/freebayes) identifikovat variantní báze – pozice proti N referenci - lze je sice „zavolat“, ale následně nelze identifikovat použitím „zabudovaných“ modelů maximum likelyhood heterozygotní pozice. Z toho důvodu byly vypočteny frekvence všech bazí na základě tabulky cDNA a pozic, které byly dříve označeny jako polymorfní – N, viz kapitola č. Níže uvedený podstatný výňatek v prvním korku popisuje získání všech bazí *readů* podporující danou pozici získanou analýzou 454 dat, poté je uveden výňatek perl skriptu s deskripcí transliterace nalezené báze za počet, který je poté vložen do MySQL databáze, viz schéma č. 8.

```
# Podle výpisu polymorfních míst extrakce bazí mapovaných na daný loci a
# pozici pomocí samools (pominutí indels a bazí s Q <= 20)
$: for f in $(cat animals.txt); do for i in $(cat positions.txt); do
samtools mpileup -Q 20 -r echo $i ${f}_s.bam >> ${f}_var.txt; done; done
# Vybrané báze jsou nyní spočítány užitím skriptovacího jazyku perl a
# nahrány do sql databáze (výňatek skriptu č. uvedeného v příloze níže):
    $A = $variants =~ tr/Aa/../;
```

```
    $T = $variants =~ tr/Tt/../;
    $C = $variants =~ tr/Cc/../;
    $G = $variants =~ tr/Gg/../;
$sql="UPDATE $table SET A_cnt = $A, T_cnt = $T, G_cnt = $G, C_cnt = $C
WHERE ref_acc = '$ref_acc' AND pos = $pos";
```

Schéma č. 8: Výňatek postupu získání SNP z nenormalizovaných RNAseq dat

Databázové tabulky frekvencí bazí pro uvedené pozice byly přeneseny do prostředí jazyka R; je nezbytné odfiltrovat sekvenační chyby od pravé heterozygotnosti. Bylo navrženo testovaní chi-kvadrát testem vůči očekávaným poměrům alel (Mgr. Karel Janko, Ph.D.)

Tento program psaný v jazyce R (schéma č. 9) vytváří objekty pro testování nulové hypotézy – diploidní jednici by měli být homozygotní v poměru bazí 0:0:0:počet dané báze, či heterozygotní v poměru 0:0:(počet dané báze / 2). V případě polyploidů je situace komplikovanější, neboť hybrid může obsahovat tři nebo dvě alely, v případě dvou alel bude očekávaná frekvence majoritní alely 2/3 vůči minoritní. První řádek kontingenční tabulky tedy obsahuje očekávané hodnoty alel a druhý řádek pozorované hodnoty očekávaných alel. Objekty označeny jako "allel" testují pouze poměry majoritních a alternativních alel, tedy bez sekvenačních chyb – tento test je posléze užit pro detekci imprinting – "nevybalancovanosti" exprimované alely v (transkripční konflikt mezi genomy nebo typu pohlaví).

Vzhledem k faktu, že nelze dělit nulou, musely být alternativní hypotézy Chi-kvadrát testu úpraveny – k nulovým očekávaným hodnotám počtu bazí/readů podporujících SNP byla přičtena jednička, totéž je nutné posléze učinit v případě heterozygotů. Tato úprava afektuje pouze SNP s nižším pokrytím, ty ale byly řešeny banálnějším způsobem, viz schéma č. 10.

```
#!/usr/bin/env Rscript
chisq.test.for.alleles<-function(x=snp) {
  phomo<-NULL;pheter<-NULL;pheter3<-NULL;pheter2<-NULL;pallel<-
NULL;pallel2<-NULL;pallel3<-NULL
  len<-c(1:length(x$pos))
  for (i in len) {
# Seřazení bazí dle počtu readů podporujících jejich pozici
      mysort<-sort(x[i,(4:7)])
# Vtvoření očekávaných hodnot poměru bazí v případě homozygota
      ehomo<-c(1,1,1,sum(x[i,(4:7)])+1)
# Vtvoření očekávaných hodnot poměru bazí v případě heterozygota
      eheter<-c(1,1,(sum(x[i,(4:7)])+2)/2,(sum(x[i,(4:7)])+2)/2)
# Tvorba očekávaného poměru v případě heterozygotní pozice polyploidního
# jedince, kdy očekáváme tři různé alely
      eheter3<-c(1,(sum(x[i,(4:7)])+3)/3,(sum(x[i,(4:7)])+3)/3,(sum(x[i,(
      4:7)])+3)/3)
# Tvorba očekávaného poměru v případě heterozygotní pozice polyploidního
# jedince
```

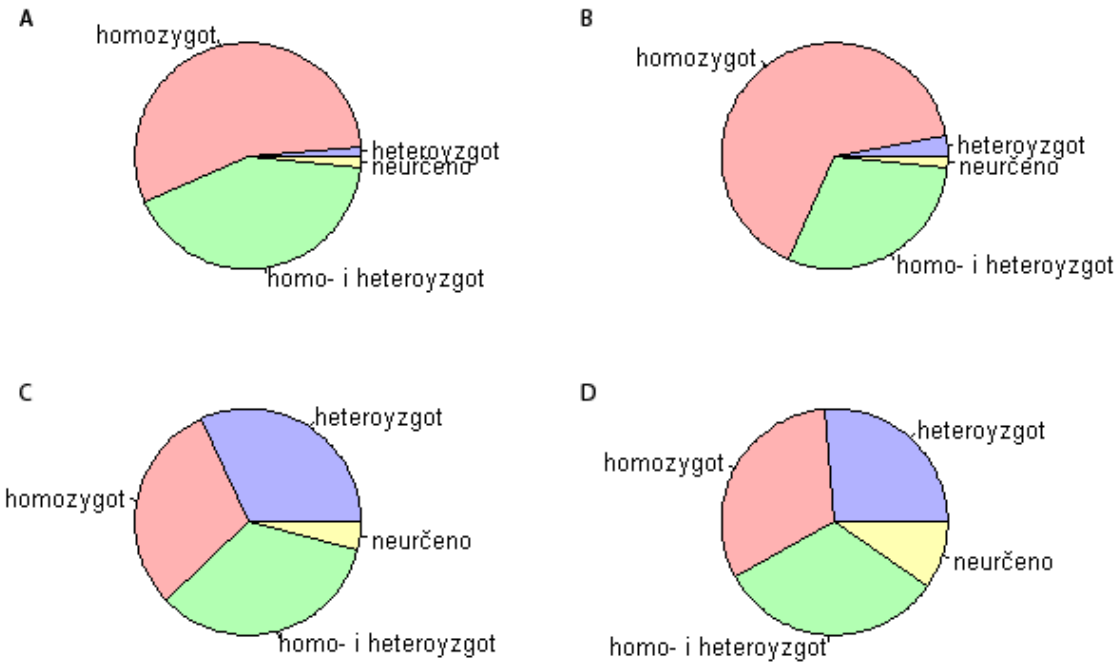
```
      eheter2<-c(1,1,((sum(x[i,(4:7)])+2)/3),(sum(x[i,(4:7)])+2)/3*2)
# Poměr dvou hlavních alel - testování disbalance (alel-specifická
# exprese)
      eallel<-c((sum(mysort[3:4])/2)+1,(sum(mysort[3:4])/2)+1)
# Poměr dvou nejabundantnějších alel v případě polyploida
      eallel2<-c((sum(mysort[3:4])/3)+1,(sum(mysort[3:4])/3*2)+1)
# expektance poměru u triploidních jedinců, kdy očekáváme 3 různé alelové
# varianty
      eallel3<-c((sum(mysort[2:4])/3)+1,(sum(mysort[2:4])/3)+1,(sum(mysor
      t[2:4])/3)+1)
# Výpočet chi-kvadrát testu
      phomo<-c(phomo,chisq.test(mysort+1,p=ehomo,rescale.p=T)$p.value)
      pheter<-c(pheter,chisq.test(mysort+1,p=eheter,rescale.p=T)$p.value)
      pheter3<-c(pheter3,chisq.test(mysort+1,p=eheter3,rescale.p=T)$p.val
      ue)
      pheter2<-c(pheter2,chisq.test(mysort+1,p=eheter2,rescale.p=T)$p.val
      ue)
      pallel<-c(pallel,chisq.test(mysort[3:4]+1,p=eallel,rescale.p=T)$p.v
      alue)
      pallel2<-c(pallel2,chisq.test(mysort[3:4]+1,p=eallel2,rescale.p=T)$
      p.value)
      pallel3<-c(pallel3,chisq.test(mysort[2:4]+1,p=eallel3,rescale.p=T)$
      p.value)
  }
# Korekce P hodnoty false rate discovery metodou kvůli mnohočetnému
# testování
      phomo.bonf<-p.adjust(phomo,method= "fdr")
      pheter.bonf<-p.adjust(pheter,method= "fdr")
      pheter3.bonf<-p.adjust(pheter3,method= "fdr")
      pheter2.bonf<-p.adjust(pheter2,method= "fdr")
      pallel.bonf<-p.adjust(pallel,method= "fdr")
      pallel2.bonf<-p.adjust(pallel2,method= "fdr")
      pallel3.bonf<-p.adjust(pallel3,method= "fdr")

p<-data.frame(ref_acc=x$ref_acc,pos=x$pos,Phomo=c(phomo),Pheter=c(pheter)
,Pheter3=c(pheter3),Pheter2=c(pheter2),Pallel=c(pallel),Pallel2=c(pallel2
),Pallele3=c(pallel3),Phomo_FDR=c(phomo.bonf),Pheter_FDR=c(pheter.bonf),P
heter3_FDR=c(phet
  return(p)
}
#Zápis tabulek s požadovanými informacemi
files <- list.files(path =
"/mnt/raid1/Projects/Cobitis/ws_jenda/RNAseq_ws/SNP_species_RNASeq/SNP_sp
ecies_verifikace/chitest/", pattern="*_var.txt")
      for (i in files) {
      y <- read.delim(i)
      yy<-chisq.test.for.alleles(y)
      write.table(x=yy,file=paste(i, '.chi.txt', sep = ""),row.names = F,
sep = '\t', quote = F)
}
```

Schéma č. 9: R program pro statistickou detekci alel.

V koláčových grafech č. jsou znázorněny počty SNP kategorií, ve kterých byly přijmuty, nebo zamítnuty nulového hypotézy o možném poměru mezi alelami v případě, homo a heterozygotů. Je zřejmé, že většina polymorfismů čistého druhu *tt* a *ee* (graf č. 10)

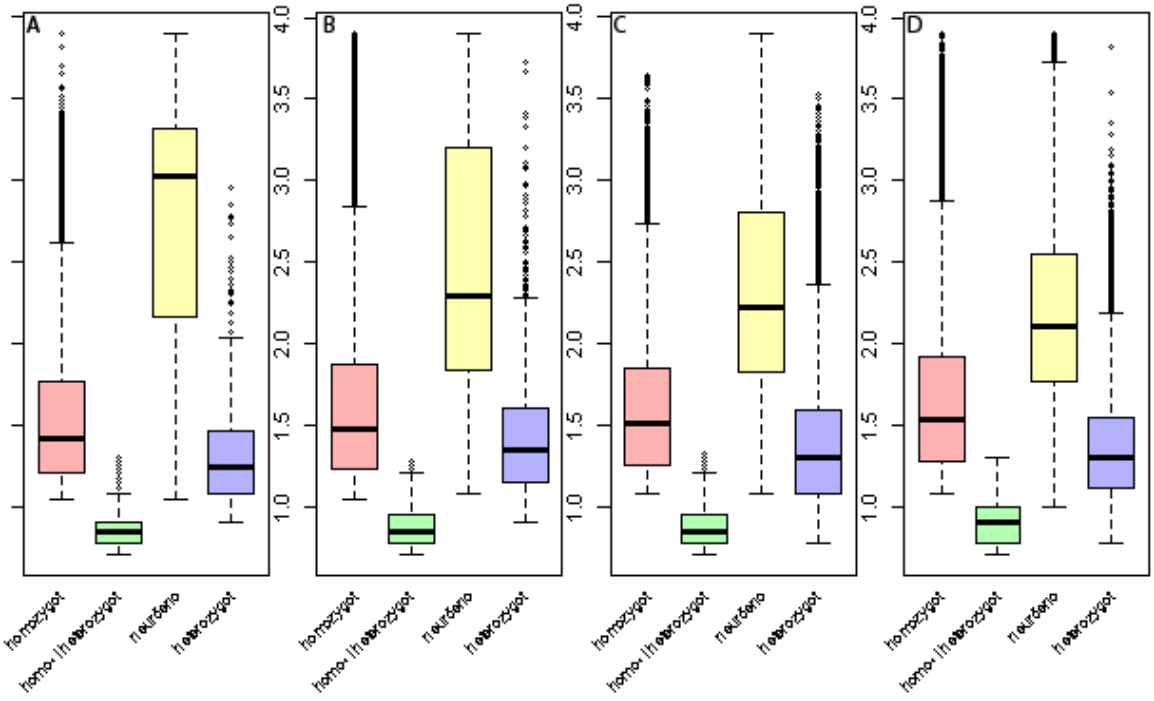
spadá do kategorie homozygot a minoritně pak heterozygot, bohužel u velké části polymorfismů nelze zamítnout ani jedna z hypotéz – kategorie homo- i heterozygot. V případě diploidního hybrida *et* a triploidního hybrida *ett* je heterozygotnost mnohem vyšší nežli u „čistých“ druhů, což podporuje hypotézu vzniku hybridizací.



Graf č. 10: Čtyři kategorie identifikovaných SNP vyhodnocených jako heterozygotní, homozygotní, nebo naopak nehomozygotní, neheterofzygotní jedince druhu *tt* označeného **A**, jedince druhu *ee* označeného **B**, hybridního jedince typu *et* označeného **C**, hybridního jedince typu *ett* označeného **D** na základě chi-kvadrát testování s $\alpha = 0{,}05$ (FDR korekce)

Báze s nízkou sekvenační hloubkou nelze chi-kvadrát testem správně identifikovat, protože je nulová hypotéza pro stav homozygotnosti přijata, ale alternativní hypotéza o heterozygotnosti taktéž (příčinou je samotné testování chi-kvadrát testem, bohužel mimo Yaytsovu korekci P hodnoty, či jiná úprava a forma testování uvedených hypotéz není možná (binominální test, Fisherův exaktní test), neboť aplikujeme kontingenční tabulky větší než 2x2). Proto jsou paradoxně tyto pozice s nižší sekvenační hloubkou zařazeny do kategorie potencionálních kandidátů *disbalancovaných*, v případě původu z rodičovských genomů taktéž alelově specificky exprimované (ASE). V případě SNP pozic naopak s vysokou sekvenační hloubkou, kde se vyskytuje pouze jedna nízce exprimovaná alternativní alela a je v nepoměrně nízkém poměru vůči majoritní alele, nelze zamítnout ani jednu z hypotéz a tyto pozice poté spadají do kategorie neurčených pozic –

pravděpodobně se jedná o sekvenační chyby. Deskripce impaktu sekvenační hloubky na detekci genotypu je znázorněn v grafu č. 11.



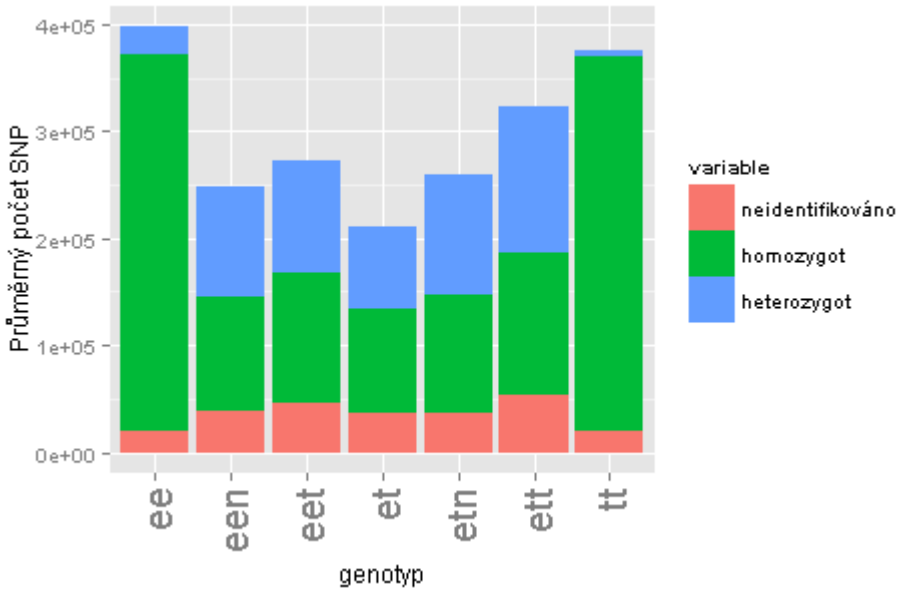
Graf č. 11: Znázornění distribuce sekvenanční hloubky formou boxplotu. Popisy a barevné označení jednitlivých skupin korespoduje s grafem č.

Část zmíněných problémů byla v případě homozygotů řešena tak, že pro polymorfismy s nízkým pokrytím byl nastaven homozygot jako poměr mezi druhou a první alternativní bazí roven nule. Heterozygotní ani homozygotní pozice ač s nízkým nebo vysokým pokrytím **nebyly** určovány na základě arbitrárních poměrů viz schéma č. 10 (perl pro stanovení výpočtu a poměru alel je uveden v příloze na CD).

```
#!/bin/bash
for f in $(cat RNAseq_tables.txt)
do
# Odstranění vších pozic, kdy SNP nebyl definován alespoň 5 ready
   mysql -e "delete FROM $f WHERE dep < 5;"
# Stanovení SNP v místech, kde byla detekována pouze jedna báze
   mysql -e "UPDATE $f set base = 1st_allele WHERE ratio = 0;
done
```

Schéma č.: 10 Stanovení homozygotních pozic v případě alel s jedou detekovanou bazí – "čistý homozygot"

V grafu č. 12 je znázorněn finální počet všech vyřešených SNP v rámci sekvenovaných genotypů. Hodnoty odpovídají průměru mezi všecmi identifikovanými jedinci daného genotypu. Velká procentuální část kategorie, kdy byly potvrzeny obě hypotézy, v případě nízké sekvenační hloubky byly vyřešeny; tedy kategorie nevyřešených pozic se znatelně nezvedla.



Graf č. 12: Barplot znázorňující průměrný počet nalezených homozygotů, heterozygotů a neidentifikovaných SNP mezi jednotlivými typy genotypů na polymorfních pozicích identifikovaných 454.

Na základě poměru majoritní a druhé "nejabundantnější" báze nelze správně detekovat heterozygotnost a to stanovením arbitrární hodnoty poměru druhé alternativní vůči majoritní bázi (v případě polymorfismů 454 dat je tento přístup parciálně v pořádku, ASE alely a somatické mutace jsou opomíjeny, především se jedná o normalizovaná data). Jedním se sdělení vyplývajícího z centrálního limitního teorému je, že s rostoucí velikostí výběru z celkového N též klesá střední chyba průměru – proto nelze srovnat poměr pouze poměr majoritní a alternativní báze. V úvahu při detekci frekvenci alely a genotypu je nezbytné brát také v potaz pravděpodobnost chyby a to jak báze, tak *namapování readu*. Pro stanovení genotypu je nejčastěji užíváno lilekyhood ratio testu a expektačního-maximalizačního algoritmu (Li 2011), do kterých je možno implementovat znalost apriori známého očekávaného poměru alel z již analyzovaných haplotypů. V našem případě se ale spokojíme pouze s chi-kvadrát testem díky faktu, že neexistuje program pro vyhodnocení SNP vůči N pozicemi v referenci. Tímto přístupem bohužel nebereme v potaz veškeré možné faktory ovlivňující poměr mezi bázemi.

4.1.10 Identifikace a validace druhově specifických SNP

Z MySQL transformace souhrnné tabulky SNP jedinců do binárního formátu je patrné (schéma č.), že v tento okamžik se již nejedná o polymorfismy jednotlivých jedinců, nýbrž o polymorfismy v rámci souboru jedinců náležících k druhu. Druhově specifické

SNP jsou definovány jako množiny SNP prezentovaných pouze u jednoho druhu (vyloučení prvků – SNP obsažených v sjednocení dvou a více množin – druhů) – viz MySQL detekce popsaná ve schématu č. 11:

```
# Označení druhově specifických SNP (zkratky druhů uvedeny v tab. č.)
UPDATE var15_species SET tt_ee = 1 WHERE (tt_a+ee_a) <= 1 AND (tt_t+ee_t)
<= 1 AND (tt_c+ee_c) <= 1 AND (tt_g+ee_g) <= 1 AND (tt_a+tt_c+tt_g+tt_t)
>= 1 AND (ee_a+ee_c+ee_g+ee_t) >= 1;
UPDATE var15_species SET tt_nn = 1 WHERE (tt_a+nn_a) <= 1 AND (tt_t+nn_t)
<= 1 AND (tt_c+nn_c) <= 1 AND (tt_g+nn_g) <= 1 AND (tt_a+tt_c+tt_g+tt_t)
>= 1 AND (nn_a+nn_c+nn_g+nn_t) >= 1;
UPDATE var15_species SET tt_ss = 1 WHERE (tt_a+ss_a) <= 1 AND (tt_t+ss_t)
<= 1 AND (tt_c+ss_c) <= 1 AND (tt_g+ss_g) <= 1 AND (tt_a+tt_c+tt_g+tt_t)
>= 1 AND (ss_a+ss_c+ss_g+ss_t) >= 1;
UPDATE var15_species set ss_nn = 1 WHERE (ss_a+nn_a) <= 1 AND (ss_t+nn_t)
<= 1 AND (ss_c+nn_c) <= 1 AND (ss_g+nn_g) <= 1 AND (ss_a+ss_c+ss_g+ss_t)
>= 1 AND (nn_a+nn_c+nn_g+nn_t) >= 1;
UPDATE var15_species set ee_ss = 1 WHERE (ee_a+ss_a) <= 1 AND (ee_t+ss_t)
<= 1 AND (ee_c+ss_c) <= 1 AND (ee_g+ss_g) <= 1 AND (ee_a+ee_c+ee_g+ee_t)
>= 1 AND (ss_a+ss_c+ss_g+ss_t) >= 1;
UPDATE var15_species set ee_nn = 1 WHERE (ee_a+nn_a) <= 1 AND (ee_t+nn_t)
<= 1 AND (ee_c+nn_c) <= 1 AND (ee_g+nn_g) <= 1 AND (ee_a+ee_c+ee_g+ee_t)
>= 1 AND (nn_a+nn_c+nn_g+nn_t) >= 1;
```

Schéma č. 11: Transformace SNP informací jedinců do binární tabulky definující varianty přítomné v daném druhu

Z celkového počtu druhově determinačních SNP mezi druhy *ee* a *tt* bylo nalezeno 32538 takových SNP. Nicméně získané druhově specifické SNP je nutno validovat, neboť geneze těchto dat je podepřena pouze čtyřmi vzorky pro druh *ee* a 8 vzorky pro druh *tt*. Díky *stringenci* detekce SNP aplikací chi-kvadrát testu a faktu, že RNAseq sekvence sekvenové na platformě HighSeq 2000 nejsou normalizovány, bylo získáno celkem 64955 pozic SNP z celkového počtu detekovaných 268651 SNP.

13.1 Detekce SNP z RNAseq dat vůči transkriptomu bez ambiguidních pozic

Pro získání SNP k výpočtu poměru nesynonymních / synonymních záměn v transkriptomu byly RNAseq sekvence mapovány vůči transkriptomu bez polymorfních pozic značených jako N. V tomto případě je možné užití běžných programů pro detekci polymorfismů. V scématu č. je uveden postup získání SNP. V prvním kroku je ze seřazených binárních výstupů mapování programem Mosaik získány veškeré polymorfismy. Následně je programem bcftools (ver. 1.2) stanoven likelyhood genotypů, počet genotypů, či alelická frekvence, pokud je analyzováno simultálně vícero jedinců. V

případě diploidů je užit při detekci SNP parametr --consensus-caller nepředpokládající minoritní varianty. V případě polyploidních vzorků je užit parametr --multiallelic-caller. Samotné SNP jsou filtrovány programem vcfutils.pl (ver. 0.1.19), přičemž je jako v případě 454 normalizovaných dat filtrována sekvenační hloubka SNP, ale také kvalita bazí a mapování, pozice SNP vzhledem k indel polymorfismům ad. V posledním kroku byly z VCF verze 4.1 (*variant call format*) (Danecek et al. 2011) vyextrahovány potřebné informace (viz schéma č. 12): název cDNA, pozice SNP, sekvenační hloubka na pozici, alternativní alely oproti referenci a kvalita Q.

```
$: for f in *_s.bam; do samtools mpileup -g -f lom300tf.fas $i | bcftools
call --consensus-caller (--multiallelic-caller) --variants-only --skip-
variants indels - | vcfutils.pl varFilter -d 5 -a 1 - > ${i}.vcf; done

$: for i in *.vcf; do cat $i | grep -v "^#" | sed -e "s/;/\t/g" | awk
'{print $1,$2,$4,$5,$6,$8}' | sed -e "s/DP=//g" > ${i}.txt;done

$: for i in $(cat tables.tab); do mysql -e "CREATE TABLE ${i}_var2
(ref_acc char(15),pos MEDIUMINT, ref char(1), var char(1), qual double,
dep int(11), PRIMARY KEY (ref_acc,pos))"; done

# Přepis možných kominací pořadí alel do IUPAC znaků
for i in $(cat tables.tab)
do
cat ${i}.txt | sed -e 's/A,G/R/g' | sed -e 's/G,A/R/g' | sed -e
's/C,T/Y/g' | sed -e 's/T,C/Y/g' | sed -e 's/G,C/S/g' | sed -e
's/C,G/S/g' | sed -e 's/A,T/W/g' | sed -e 's/T,A/W/g' | sed -e
's/G,T/K/g' | sed -e 's/T,G/K/g' | sed -e 's/A,C/M/g' | sed -e
's/C,A/M/g'  | sed -e 's/C,G,T/B/g'  | sed -e 's/C,T,G/B/g'  | sed -e
's/G,C,T/B/g'  | sed -e 's/C,T,G/B/g'  | sed -e 's/T,C,G/B/g'  | sed -e
's/T,G,C/B/g'  | sed -e 's/A,G,T/D/g'  | sed -e 's/A,T,G/D/g'  | sed -e
's/G,T,A/D/g'  | sed -e 's/G,A,T/D/g'  | sed -e 's/T,A,G/D/g'  | sed -e
's/T,G,A/D/g'  | sed -e 's/A,C,T/H/g'  | sed -e 's/A,T,C/H/g'  | sed -e
's/C,A,T/H/g'  | sed -e 's/C,T,A/H/g'  | sed -e 's/T,C,A/H/g'  | sed -e
's/T,A,C/H/g'  | sed -e 's/A,C,G/V/g'  | sed -e 's/A,G,C/V/g'  | sed -e
's/C,A,G/V/g'  | sed -e 's/C,G,A/V/g'  | sed -e 's/G,C,A/V/g'  | sed -e
's/G,A,C/V/g' > ${i}_F.txt
done
```

Schéma č. 12: Deskripce postupu získání SNP vůči referenci bez N ambiguidních pozic v místech polymorfismů detekovaných na základě 454 dat

14.1 Detekce SNP z SeqCap dat vůči transkriptomu bez ambiguidních pozic

SNP z gDNA mapované na referenční transkriptom programem BWA byly extrahovány zcela identickým způsobem (opět s ohledem na ploidii jedince) až na výjimku týkající se finálního filtrování SNP pozic programem vcfutil.pl VarFilter, kdy byly odstraněny všechny báze -5 a +5 v okolí *gap alignmentu* – cílem je zbavit se artificiálních

chyb způsobených přítomností extra genomových a intronových sekvencí přítomných v *readech mapujích* se na cDNA.

15.1 Analýza RNAseq exprese genů

K odhalení DE genů byly zvolen DESeq balík programu R (3.1.14) (Anders and Huber 2010) a balík edgeR (Zhou et al. 2014). V schématu č. je uveden postup získání DE genů balíkem DESeq a konsekventně i nabohacených GO termínů a KEGG drah těchto DE genů.

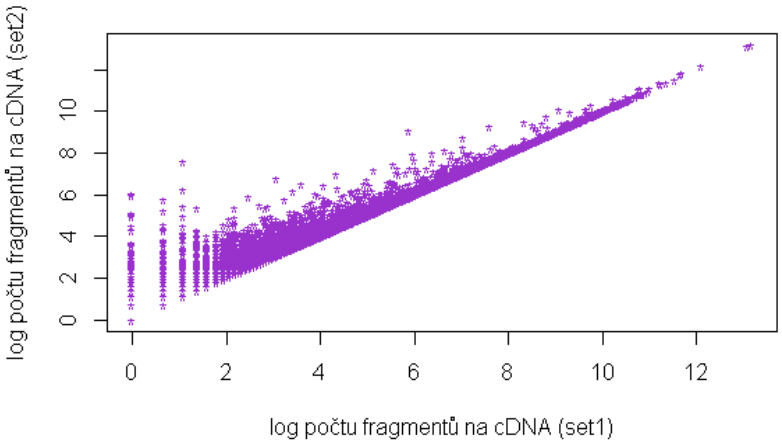
4.1.11 Extrakce počtu *namapovaných readů* na cDNA RNAseq vzorků

Informace o *namapovaných readech* je uložena v binárním formátu – aplikací samtools (1.2) je informace převedena do formátu SAM (Li et al. 2009). Ze SAM formátu byla informace o počtu *readů* na cDNA získána dvěma způsoby. V prvním případě byly zjišťovány sumy *readů* na gen pouze unikátně *namapovaných readů* na referenci značených v hlavičce formátu SAM jako @SQ. V druhém případě byly z binárního formátu získány *namapované ready*, které byly dále selektovány dle MAPQ skóre s limitem 10 (šance, že *read* se může mapovat na jinou pozici v referenci, činí 1:10) programem Count_reads_denovo.pl (Jennifer Shelton, sdíleno GitHub, nepublikováno;). Komparace mezi datasety má za cíl zjistit, zda hraje roli fakt, že reference může stále obsahovat špatně složené cDNA, které neodpovídají sekvenci genomu, ne následkem editace RNA, ale díky chybě při složení. Níže uvedené schéma příkazů č. 13, v prostředí shell popisují získání samotných expresních informací (počet *readů* namapovaných na cDNA).

```
# Selekce unikátně mapovaných readů – set1
$: cat animal.sam | grep -v ^@SQ | grep lom300tt | awk '{ print $3}' |
sort | uniq -c | awk '{ print $2, $1 }' > animal.txt
# Sumarizace počtu readů na cDNA v MySQL
> SELECT ref_acc, SUM(cnt) AS count FROM animal GROUP BY (ref_acc);

# Selekce namapovaných readů s limitní hodnotou MAPQ – set2
$: samtools view -h -F 4 -b animal.bam > mapped_animal.bam -@ 16
$: samtools sort mapped_animal.bam animal_sort_m.bam -@ 16
$: samtools view -h -o cab01L_sort_m.sam cab01L_sort_m.bam.bam -@ 16
$: perl Count_reads_denovo.pl -s animal_sort_m.sam -o animal_sort_m.txt
```

Schéma č. 13: Získání počtu *namapovaných readů* na cDNA – dvěma způsoby (selekce pouze unikátních, či i neunikátně *namapovaných readů*)



Graf č. 13: Srovnání počtu *readů* na cDNA mezi dvěma výše generovanými datasety jednoho jedince (provedena transformace dat log10 +1).

Spearmanův korelační koeficient mezi získanými datasety činí 0.994. Významný rozdíl je znatelný pouze u genů s velmi nízkou expresí, kde bylo v setu 2 ponecháno větší množství *readů*. Pro analýzu DE byl zvolen set č. 2; sice není ztracena část informace přílišnou *stringencí* výběru, viz graf č. 13; neunikátní, zároveň možné chybně *namapované ready* lze očividně během normalizace a detekce DE odfiltrovat – při analýze nebyl nalezen signifikantní rozdíl, byly získány téměř identické DE geny porovnáním níže zmíněných skupin. Naopak *stringence* mezi nástroji pro detekci DE genů se liší značně. Po filtrování DE genů na stejné hodnotě α byla nelezeny stejné geny, ale rozdílné, výrazně posunuté P hodnoty.

4.1.12 Identifikace diferenciálně exprimovaných genů

V prvním kroku jsou získané počty *radů* na cudná normalizovány, neboť mezi jedinci není sekvenační hloubka identická viz graf č., nastává rovněž problém s délkou genů, i variance dat je nechtěně obohacena o technickou variabilitu, v poslední řadě je nutno vzít v úvahu preference delších *radů* následkem přípravy sekvenační knihovny a naopak preferenci těch kratších při klastrování *radů* sek věnováním na iluminát platformě (Dillies et al. 2013). V našem případě je volba vhodné metody normalizace zcela zásadní, protože *bias* délek genů hraje v našem datasetu významnou roli, neboť homologní sekvence mezi druhy se mohou v ojedinělých případech lišit i co do délky. EdgeR a DESeq jsou adekvátní rovněž z toho důvodu, že vycházejí z předpokladu, že mezi porovnávanými skupinami z biologického hlediska není mnoho DE genů.

V následujícím kroku je stanovena variance, disperze a jejich průměr v rámci deklarovaných skupin určených ke komparaci včetně variance, disperze mezi těmito skupinami. Disperzi lze popsat jako kvadratické umocnění koeficientu biologické variance. Samotná variance je sumou dvou faktorů: úrovně variance mezi skupinami v rámci replikátu a „nejistotou“ vypočítanou na základě koncentrace *readů*; tento faktor predikuje také náhodnou biologickou/technickou variabilitu z poissonovy distribuce, což má svůj význam zejména u genů s velmi nízkou hladinou exprese. Čím je vyšší disperze, biologická variabilita mezi vzorky v rámci skupiny, tím více je třeba replikátů, nebo o to větší musí být rozdíl v expresi, aby DE geny mohly přejít práh signifikance. V grafu č. 16 je znázorněn vztah mezi průměrem normalizovaných počtů *readů* a disperze v log škálách. DE geny jsou určeny pomocí negativně binominálního modelu vycházejícího aproximací z poissonova modelu. Výsledná P hodnota DE je korigována FDR (*false rate discovery*), ježto práh P hodnoty je při mnohonásobném testování zvyšován - tedy i chyba II. řádu. Postup analýzy DE užitím DESeq potažmo i edgeR softwaru je znázorněn ve schématu č. 16. Fyzický příklad postupu získání DE genů mezi skupinami asexuálně a sexuálně se rozmnožujícími jedinci je uveden s popisy ve schématu č. 14. Ve schématu č. 15 je uveden postup, jakým bylo provedeno znázornění heatplot a shlukovací analýza genů uvedených ve výsledcích sekce diferenciální genové exprese.

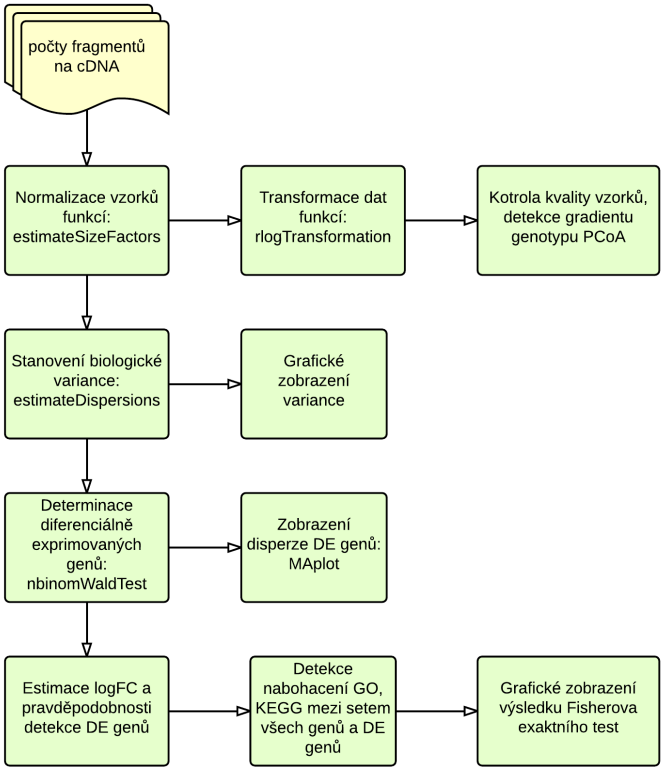


Schéma č. 16: Znázornění postupu získání DE genů utilizací balíku DESeq (přeneseně platí i pro edgeR využívající velmi podobný postup)

```
library(sqldf)
library(edgeR)
# Odstranění genů se sumou readů menší než 20
cnt_o <- dat_o[rowSums(dat_o)>20,]
# Odstranění jedinců, kteří nejsou přítomni v obou datasetech, či nemají
# dostatečnou sekvenační hloubku
in_o_sex <- sqldf("select * from animals_o where name != 'cab07o' and
name != 'cab23o' and name != 'cab15o' and name != 'cab14o'")
dat_o_sa<- cnt_o[, names(cnt_o) %in% in_o_sex$name]
cds_o_sa <- DGEList(dat_o_sa, group = in_o_sex$sex)
# normalizace dat metodou TMM (trimmed mean median)
cds_o_sa <- calcNormFactors(cds_o_sa)
# Stanovení disperse v rámci skupiny
cds_o_sa <- estimateCommonDisp(cds_o_sa, verbose=TRUE)
# Stanovení disperse mezi skupinami
cds_o_sa <- estimateTagwiseDisp(cds_o_sa)
# Výběr skupin ke komparaci
et_o_sa <- exactTest(cds_o_sa, pair=c("asex","sex"))
top_o_sa <- topTags(et_o_sa, n=nrow(cds_o_sa$counts))$table
# Selekce genů s P hodnotou < 0.05
de_edgeR_o_sa <- top_o_sa[top_o_sa$FDR<0.05]
de_edgeR_o_sa$ref_acc <- rownames(de_edgeR_o_sa)
# Spojení výsledků s náležitou anotací
de_edgeR_o_sa_ann <- sqldf("select ref_acc, logFC, logCPM, PValue, FDR,
annotation from de_edgeR_o_sa join annotation using (ref_acc) where
annotation != 'NA'")
# Výběr genů s největšími rozdíly exprese mezi skupinami seřazených dle
# absolutní hodnoty (logFC) určeným
```

```
real_top_o_sa <- sqldf("select ref_acc, logFC, logCPM, PValue, FDR,
annotation from de_edgeR_o_sa join annotation using (ref_acc) where
annotation != 'NA' order by abs(logFC) desc limit 50")
```

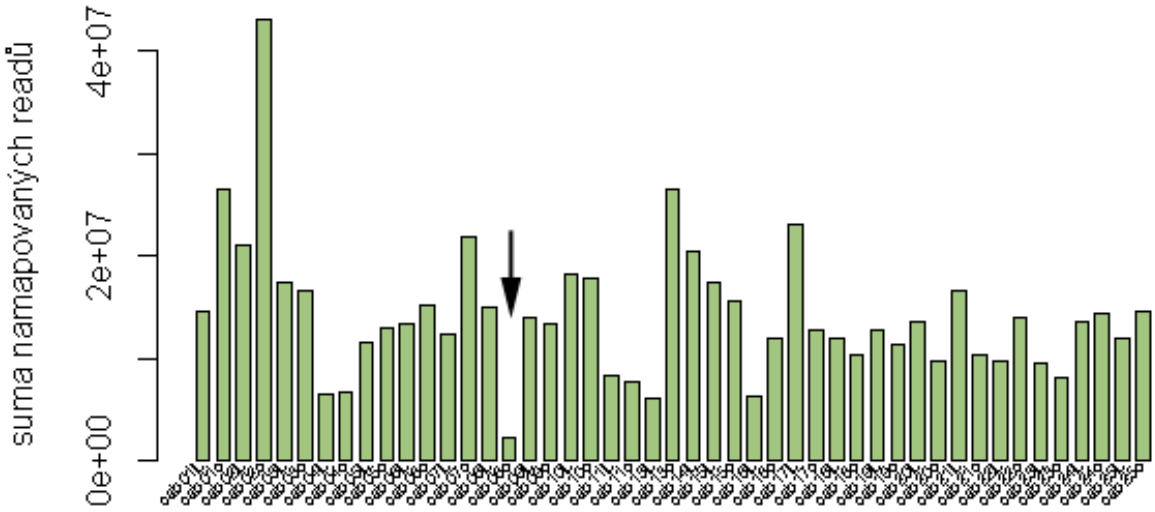
Schéma č. 14: Deskripce fyzického postupu při získání DE genů metodou edgeR

```
# Získání dat pro heatplot2 a shlukovací analýzu spojedním dle klíče DE
# genů s TMM normalizovanými daty napříč použitými vzroky
cluster<-merge(cds_o_sa$pseudo.count,de_edgeR_o_sa,by="row.names")
# Vygenerování heatplot grafu
library(RColorBrewer)
cols <- colorRampPalette(brewer.pal(8,"Blues"))(100)
mydatascale <- t(scale(t(data.matrix(cluster[,2:21]))))
# Shluková analýza řádků na základě korelační parametrické matice
hr <- hclust(as.dist(1-cor(t(mydatascale), method="pearson")),
method="complete")
# Shluková analýza sloupů na základě korelační neparametrické matice
hc <- hclust(as.dist(1-cor(mydatascale, method="spearman")),
method="complete")
# barevné zvýraznění shluků genů
mycl <- cutree(hr, h=max(hr$height)/1.5)
mycolhc <- sample(rainbow(256))
mycolhc <- mycolhc[as.vector(mycl)]
# Tvorba heatmap diagramu z dat normalizovaných vzroků použitých k
# analýze DE genů – příklad komparace skupiny "sex" vs "asex" (20 vzorků
# oocytů)
heatmap.2(data.matrix(cluster[,2:21]), Rowv = as.dendrogram(hr), Colv =
as.dendrogram(hc), scale="row", RowSideColors=mycolhc, labRow = '',
cexCol = 1.5,col= cols,labCol = in_o_sex$biotype,trace = "none",keysize =
2,key.title = "")
# Shlukovací analýza genů
library(parallel)
cl <- makeCluster(5, type = "PSOCK")
cluster1 <- merge(cds_o_sa$pseudo.count,real_top_o_sa,by="row.names")
pv <- parPvclust(cl,t(data.matrix(cluster1[,2:21])),
method.dist="correlation", method.hclust="median", nboot=1000)
# Explorace shlukovací analýzy formou dendrogramu se zvýrazněným větvením
# podle signifikance P hodnoty
library(dendextend)
dend<- as.dendrogram(pv)
labels(dend) <- as.character(cluster$annotation)
dend %>%
  pvclust_show_signif_gradient(pv) %>%
  pvclust_show_signif(pv) %>%
  plot(main = "",xlab = "")
pv %>% text
pv %>% pvrect(alpha=0.95)
```

Schéma č. 15: Deskripce aplikovaných statistických metod programu R vyobrazených ve výsledcích týkajícín se diferenciální genové exprese.

Z analýzy DE byli vyloučeni ti jedinci, kteří se vyznačovali neúměrně nízkou sekvenační hloubkou vzhledem k sumám počtu *readů* na cDNA ostatních jedinců viz graf č. 16, protože se projevil i *bias* v datech, kdy na PCA znázornění. Jedinec této distance na ose PC1 (*principal component* 1) deformuje samotné znázornění převedení mnohorozměrného prostoru do 2D projekce. Byly ponechány vzorky, které byly získány

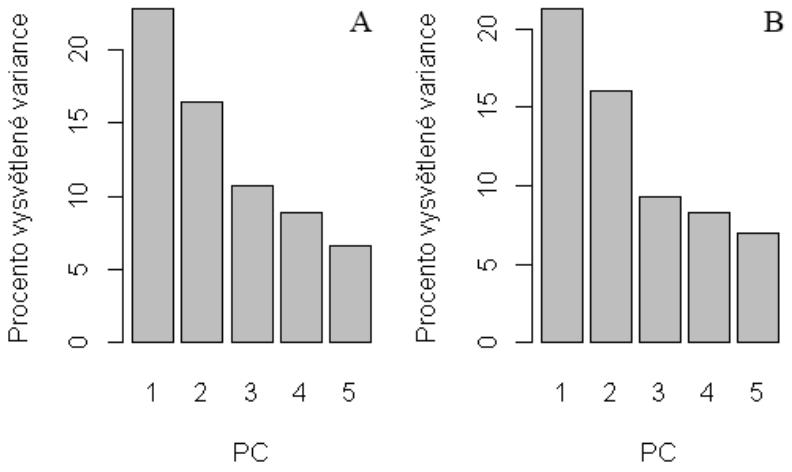
jak z jater, tak oocytů. Celkem bylo analyzováno 40 jedinců: 20 vzorků jater a 20 vzorků oocytů. Pro kontrolu tedy byly analyzovány čtyři sety dat – geny anotované, anotované včetně neanotovaných a to jak unikátně, tak neunikátně *namapovaných*, mezi kterými však není rozdíl - nebudou dále zmiňovány. Diference mezi výsledky získaných ze setu pouze anotovaných a neanotovaných dat je uvedena ve výsledcích. Jedinci vykazují distanci především na základě svého genotypu, viz graf PCA č. 17, 18. Nejedná se o PCA pouze několika set genů s největší variabilitou mezi předem definovanými skupinami; jak bývá obecně zvykem, byl použit set všech genů přítomných v analýze (anotované i neanotované). Vzorky stejných genotypů obou skupin vykazují mírně rozdílnou standartní chybu – vyšší u jater, což může být dáno environmentálními faktory, které na oocyty nemají takový vliv. Pro znázornění PCA distancí byly vybrány dvě majoritní komponenty, vysvětlující nejvíce variability mezi vzorky, graf č. 15.



Graf. č. 16: Sloupcový diagram znázorující sekvenační hloubky vzorků oocytů a jater; vzorek označený šipkou byl z datasetu odstraněn pro nedostatečnou sekvenační hloubku

Druh	**haplotyp**
Cobitis elongatoides	***e***
Cobitis taenia	***t***
Cobitis tanaitica	***n***

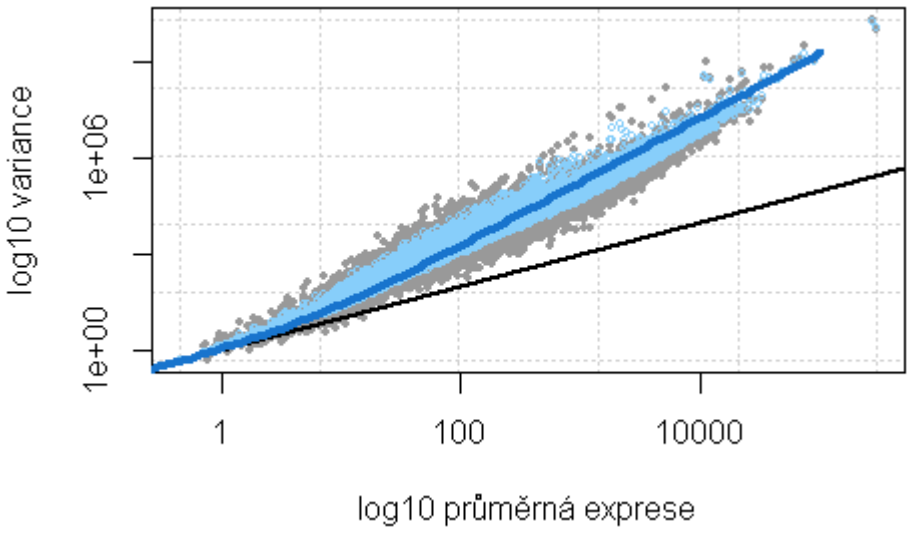
Tab. č. 6: Pracovní označení haplotypů analyzovaných genomů druhů.



Graf č. 15: 1 Variance pěti komponent PCA s nejvyšším příspěvkem; A = vzorky oocytů; B = vzorky jater

Graf č. 17: PCA plot znázorňující vzájemnou podobnost mezi vzorky **oocytů** při vybrání dvou hlavních komponent. Elipsy znázorňují standartní chybu centroidu, samotný centroid dané skupiny je vyznačen křížkem

Graf č. 18: PCA plot znázorňující vzájemnou *similaritu* mezi vzorky **jater** při vybrání dvou hlavních komponent. Elipsy znázorňují standartní chybu centroidu; centroid dané skupiny je vyznačen křížkem



Graf č. 19: Znázornění vztahu mezi průměrnou expresí a variancí vzorků; šedé body reprezentují netransformované variance počty *readů* na cDNA; světle modré body reprezentují varianci mezi vzorky; tmavě modrá linka znázorňuje společnou disperzi, trend dat, zatímco černá přímka je zobrazením poissonovské variance

Analýzu RNAseq velmi silně ovlivňuje především sestavení referenční sekvence, zvláště pokud je sestavena z několika druhů, či polymorfních populací. Dochází totiž k formování výše popsaných pseudoparalogních sekvencí. Majoritní část problematických sekvencí byla odstraněna tím způsobem, že na referenci byly sekvence *remapovány*; cDNA, ve kterých byly detekovány nové SNPs, došlo k jejich odstranění.. Vzhledem k metodě získání cDNA, jsou častěji „prosekvenovány“ oblasti 5’ konce cDNA, neboť

reverzní transkripce byla provedena aplikací primeru polyT, přepisující sekvenci od polyA konce mRNA. Ačkoliv je užitá reverzní transkriptáza schopna přepisu až 20 kb, často od templátu disociuje a to zřejmě lineárně, jak je patrno z grafu č. 5. Tímto způsobem je do výsledků vzhledem k nadměrně dlouhým sekvencí vnášen další *bias*, neboť tyto dlouhé cDNA sekvence se mohou jevit jako podexprimované. Naštěstí datasety RNAseq byly získány aplikací identického protokolu včetně téže reverzní transkriptázy, a proto tato chyba nehraje v komparaci dat s cílem identifikace diferenciálně exprimovaných genu žádnou roli (*procesivita* reverzní transkriptázy je sekvenčně *independentní*).

Pro kontrolu specifity testování získaných výsledků DE genů a verifikace gradientu (testování, jakým způsobem reflektuje genotyp jedince jeho *similaritu*, polohu vůči rodičovským druhům) mezi druhy *tt* a *ee* byl Mgr. Ladislavem Pekárikem, Ph.D. vypočtena míra příspěvku jednotlivých genů na vytvořeném gradientu mezi těmito druhy. Podstatný výpočet výňatku programu je uveden ve schématu č. 16 (autor: Mgr. Ladislav Pekárik, Ph.D).

```
ordscores$naxis1<-(ordscores$axis1*cos(ang)+ordscores$axis2*sin(ang))
ordscores$naxis2<-(-ordscores$axis1*sin(ang)+ordscores$axis2*cos(ang))
```

Schéma č. 16: trigonometrický výpočet příspěvku genů vzhledem k PCoA "nafitovanému" gradientu mezi genotypy *tt* a *ee* (Mgr. Ladislav Pekárik, Ph.D.)

V předešlých odstavcích této kapitoly věnované deskripci transkripčních dat a vybraných problémů spojených s analýzou transkripčních dat RNAseq, byl znázorněn simplifikovaný pohled na kontolní body postupu získíni DE genů a samotná vstupní data. Posledním kontrolním bodem je vlastní explorace distibuce výsledků DE genů. Graf č. 20 nám říká, jaký je vztah mezi expresí (fold-change) diferenciálně exprimovaných genů a normalizovaným počtem readů na cDNA (*counts per milion* – CPM) z porovnání skupin sexuálně a asexuálně se rozmnožujících jedinců, jejich jater a oocytů.

Graf č. 20: Závislot míry exprese (CPM) na násobku změny mezi srovnávanými skupinami (sexuálě a asexuálně se reprodukující jedinci, A = vzorky oocytů, B = vzorky jater); červené body značí DE geny stanoveneé na hladině α 0.05 FDR korigované P hodnoty.

4.1.13 Identifikace nabohacených GO termínů a KEGG metabolických drah

K anotaci *kontigů* referenčního transkriptomu náleží též GO (gene ontology) identifikátory a KEGG identifikátory metabolických drah. Užitím zmíněného softwaru Blast2GO byla provedena analýza nabohacení *subsetu* GO získaných analýzou diferenciálně exprimovaných genů programem edgeR a to z důvodu nižší *stringence* testování. Analýza nabohacení GO byla provedeno Fisherovým exaktním testem (komparace pozorovaných a očekávaných počtu GO mezi všemi anotovanými geny a zvolenou podmnožinou identifikovaných DE genů) na hladině α = 0,05 korigované P hodnoty FDR metodou.

4.1.14 Validace RNAseq srovnání výsledků RT-qPCR vybraných DE genů

Ačkoliv jsou RNAseq data genové exprese rutinně interpretovány a jejich výpovědní hodnota je podpořena stovkami článků, je stále nezbytné získané výsledky validovat, neboť může dojít k nezanedbatelnému množství chyb, ať již při preparaci vzorků, sekvenování, analýze dat, nebo již nevhodným experimentálním designem. Je ale známo, že korelace mezi relativní expresí genů qPCR a RNAseq je velmi těsná, jak naznačuje tato publikace: (Gavery and Roberts 2012). RT-qPCR

První otázkou validace RNAseq je: Jsou detekované transkripty diferenciálně exprimované mezi vzorky (hovoříme o technické reprodukcibilitě)? Zadruhé: Jsou detekované transkripty diferenciálně exprimované mezi skupinami (biologická variance). Zatřetí: Mají tyto rozdíly biologickou signifikanci – např. fenotypová kauzalita (qPCR v tomto ohledu není nápomocná).

cDNA pro stanovení relativní exprese užitím qPCR byla získána reverzní transkripcí popsanou v metodice. Pro qPCR byly navrženy 3 *house-keeping* geny (HS) a to pro: rpl13a, non-POU, hprt1 získaných na základě identifikace HS nejnižší variance mezi vzorky jater užitím softwaru normfinder (version 5, 2015-01-05) (Andersen et al. 2004). Software je určen primárně pro 2^{-Ct} (*cycle trashold*) hodnoty qPCR a *microarray* normalizovaných, ale nelogaritmovaných výstupních dat, nicméně dle mínění autora jej lze aplikovat na RNAseq nenormalizovaná data (software ale neumí normalizovat vzhledem k délce sekvence, využívá pouze geometrického průměru napříč vzorky).

non-POU[cobitis] FWD	5′-CAGGTGGAGCGTAACATCAA-3‘
non-POU[cobitis] REV	5‘-CGCAGGAGATCTTGTCTCATC-3‘
rpL13a[cobitis] FWD	5‘-GCCACATTGAGGAGGTCAAA-3‘
rpL13a[cobitis] REV	5‘-CAGCCTGGCGTCAATAAGAA-3‘
hprt1[cobitis] FWD	5‘-ACGGACTACCATAACCCATTTC-3‘
hprt1[cobitis] REV	5‘-GGTCATAGCCTTGCTCTTCAT-3‘

Tab č. 7: Sekvence primerů HS genů „posazených“ blízko 3’ konci sekvence užitých pro RT-qPCR validaci RNAseq relativní exprese

Výsledky korelace relativní exprese mezi qPCR a RNAseq jaterní tkáně několika náhodně vybraných genů vykazující diferenciální expresi bohužel nebyly dokončeny. V tab. č. 8 uvedeny primery pro vybrané diferenciálně exprimované geny mezi skupinou sex asex jaterní tkáně. Všechny uvedené primery byly testovány.

fatt_elong FWD	TGG TGG TTT GTC TTG AAC TGG
fatt_elong REV	5-CAG CAG ACA GCC CAT AAT ACG
GSTs FWD	GCT GGA GCT GAG TTT GAG G
GSTs REV	CTG CAT TCC ATC CAT TTC AAC C
CH25H FWD	CCA GAA CAG AGA AGA TGT CTG G
CH25H REV	GAA GAG CAC TGG GAA GAA GG
CPA2 FWD	ATG TGG CTC TAT CTG CAA GC
CPA2 REV	ATG CCA CGC TGG TAA GC
triqt FWD	TCT CTC AGG TGT AGA AGG ATG G
triqt REV	CGA TCT GTT TAC GGT ACT GAT CC

Tab č. 8: Sekvence primerů DE genů „posazených“ blízko 3’ konci sekvence užitých v RT-qPCR validaci RNAseq relativní exprese

16.1 Analýza alelově specifické exprese (ASE) hybridních jedinců

Dalším důležitým bodem mé práce je test hypotézy, zda u hybridů dochází k atenuaci exprese alel jednoho rodičovského genomu na úkor druhého, nebo dokonce k systematickému imprintingu jednoho rodičovského genomu.

Premisa detekce disbalance exprese RNA v závislosti na původu z rodičovského druhu je následující. V první řádě je nezbytné adekvátním statistickým způsobem testovat, zda alela je, či není disbalancovaná. Nulová hypotéza je uvedena v schématu č. 9 – parametr eallel, jakožto poměr 1:1 dvou hlavních bazí na dané pozici SNP. Za druhé test je prováděn pouze na výše získaných druhově specifických SNP pozicích, protože cílem je diferencovat ready pocházející z konkrétního druhu. Dále jsou vybrány všechny SNP, u kterých lze jejich původ jednoznačně připsat rodičovským druhům – to znamená, že testujeme pouze ancestrální polymorfismy přítomné v době vzniku hybridního jedince. Pokud je během fylogeneze klonální linie alela ztracena, nebo dojde z záměně báze na dané pozici, nebylo by možné určit původ alely, a proto takovéto pozice v našem testu nezohledňujeme.

4.1.15 Determinace původu alely na základě druhově specifických SNP

Ve schématu č. 16 je uveden postup identifikace alel, u kterých můžeme předpokládat původ z jednoho, nebo obou rodičovských druhů. Pokud je jedna z alel transkripčně umlčena jedním z genomů, jeví se v hybridním jedinci jako homozygotní, v případě, že umlčení není úplné, můžeme stále nalézt přítomnost heterozygotní alely a v tomto případě se musí alela shodovat s oběma rodiči. Pozn.: heterozygotní pozice byly vyhodnoceny na základě P hodnoty hypotézy nazvané jako ehomo (*expected homozygot*), kde jsme testovali royložení bazí na základě chi kvadrát testu oproti expektanci v případě homozygota a ehetero, kde jsme dělali totéž, ale oproti očekávanému rozložení oproti heterozygotu. Očekávané rozložení readů v tomto případě bralo v potaz fakt, že pozice mohla být i postižena sekvenační chybou – testovali jsme tedy všechny 4 možné stavy dané báze. Je dobré si zde uvědomit, že daná pozice nemusí zamítnout ani jednu hypotesu, anebo naopak může vést k zamítnutí obou najednou, což může být způsobeno jak

sekvenační chybou, chybným assembly, tak i samotným procesem nevyvážené transkripce. Po určení homo-heterozygotnosti na všech testovaných pozicích jsme pak přistoupili k testování samotné disbalance alel a to na základě P hodnoty podle nulové hypotézy *eallel*, která se již specificky pomocí chi-kvadrát testu táže na to, zda jsou dvě majoritní báze na dané pozici v rovnováze; viz schéma (bash/MySQL) č. 17

```
#!/bin/bash
for i in $(cat ASE_all.tab)
do
# Stanovení původu homozygotních pozic hybrida
    mysql -e "UPDATE ${i}_ASE a JOIN var15_species_final b USING
              (ref_acc,pos) SET a.ok = 1 WHERE a.base = 'A' AND b.tt_a = 1
              AND b.ee_a = 1;"
    mysql -e "UPDATE ${i}_ASE a JOIN var15_species_final b USING
              (ref_acc,pos) SET a.ok = 1 WHERE a.base = 'T' AND b.tt_t = 1
              AND b.ee_t = 1;"
    mysql -e "UPDATE ${i}_ASE a JOIN var15_species_final b USING
              (ref_acc,pos) SET a.ok = 1 WHERE a.base = 'C' AND b.tt_c = 1
              AND b.ee_c = 1;"
    mysql -e "UPDATE ${i}_ASE a JOIN var15_species_final b USING
              (ref_acc,pos) SET a.ok = 1 WHERE a.base = 'G' AND b.tt_g = 1
              AND b.ee_g = 1;"
# pro heterozygotní báze – alespoň jedna rodičovská alela se musí
# shodovat bazí hybridního jedince
    mysql -e "UPDATE ${i}_ASE a JOIN var15_species_final b USING
              (ref_acc,pos) SET a.ok = 1 WHERE a.base = 'R' AND (b.tt_a =
              1 AND b.ee_g = 1) or (b.ee_a = 1 AND b.tt_g = 1);"
    mysql -e "UPDATE ${i}_ASE a JOIN var15_species_final b USING
              (ref_acc,pos) SET a.ok = 1 WHERE a.base = 'Y' AND (b.tt_c =
              1 AND b.ee_t = 1) or (b.tt_t = 1 AND b.ee_c = 1);"
    mysql -e "UPDATE ${i}_ASE a JOIN var15_species_final b USING
              (ref_acc,pos) SET a.ok = 1 WHERE a.base = 'S' AND (b.tt_c =
              1 AND b.ee_g = 1) or (b.tt_g = 1 AND b.ee_c = 1);"
    mysql -e "UPDATE ${i}_ASE a JOIN var15_species_final b USING
              (ref_acc,pos) SET a.ok = 1 WHERE a.base = 'W' AND (b.tt_a =
               1 AND b.ee_t = 1) or (b.tt_t = 1 AND b.ee_a = 1);"
    mysql -e "UPDATE ${i}_ASE a JOIN var15_species_final b USING
              (ref_acc,pos) SET a.ok = 1 WHERE a.base = 'K' AND (b.tt_g =
              1 AND b.ee_t = 1) or (b.tt_t = 1 AND b.ee_g = 1);"
    mysql -e "UPDATE ${i}_ASE a JOIN var15_species_final b USING
              (ref_acc,pos) SET a.ok = 1 WHERE a.base = 'M' AND (b.tt_a =
              1 AND b.ee_c = 1) or (b.tt_c = 1 AND b.ee_a = 1);"
    # Vsude kde nelze zjistit - neni ok tzn. 0
    mysql -e "UPDATE ${i}_ASE SET ok = 0 WHERE base IS NULL;"
done
```

Schéma č. 16: Stanovení logické hodnoty (1,0) v případě, že platí podmínka původu SNP u hybridních jedinců (MySQL, bash)

4.1.16 Stanovení disbalancovaných alel jedinců hybridního původu

V níže uvedeném výňatku bash skriptu je znázorněn postup označení disbalancovaných SNPs. Ptáme se, jaký genom převažuje – zda byla zamítnuta nulová

hypotéza o balancovaném stavu počtu bazí na SNP. Pokud je tato podmínka splněna, ptáme se, zda je tento SNP druhově určující a shoduje-li se báze s rodičovskými druhy? V posledním kroku je stanoveno, zda převažuje alela pocházející z genomu *ee* či *tt* (v schématu č. 14 je uvedeno pouze nahrazení, zda převažuje alela pocházející z genomu *ee* a to pouze vybraných bazí pro demonstraci postupu).

```
# Určení disbalance ve směru genomu ee pokud je má hybrid disbalancovanou
# alelu na P < 0.05, pozice je druhově určující, alela se shoduje s
# jedním z rodičovských genomů a báze hybrid je homozygotní a odpovídá
# rodičovskému polymorfismu, pro kontrolu je udána také podmínka počtu
# bazí (nejvyšší počet musí mít výsledná báze SNP). Všechny typy bází pro
# přehlednost neuvedeny
    mysql -e "UPDATE ${i}_ASE a JOIN ${i}_var b USING (ref_acc,pos) JOIN
              var15_species_final c USING (ref_acc,pos) set a.balance =
              'ee' WHERE a.Pallel_FDR < 0.05 AND a.tt_ee = 1 AND ok = 1
              AND a.base = 'A' AND c.ee_a =1 AND b.A_cnt > b.T_cnt AND
              b.A_cnt > b.C_cnt AND b.A_cnt > b.G_cnt;"
# Totéž jako výše, nicméně zde je hybrid v heterozygotním stavu
    mysql -e "UPDATE ${i}_ASE a JOIN ${i}_var b USING (ref_acc,pos) JOIN
              var15_species_final c USING (ref_acc,pos) set a.balance =
              'ee' WHERE a.Pallel_FDR < 0.05 AND a.tt_ee = 1 AND ok = 1
              AND c.ee_a = 1 AND a.base = 'R' AND b.A_cnt > b.T_cnt AND
              b.A_cnt > b.C_cnt AND b.A_cnt > b.G_cnt;"
    mysql -e "UPDATE ${i}_ASE a JOIN ${i}_var b USING (ref_acc,pos) JOIN
              var15_species_final c USING (ref_acc,pos) set a.balance =
              'ee' WHERE a.Pallel_FDR < 0.05 AND a.tt_ee = 1 AND ok = 1
              AND c.ee_t = 1 AND a.base = 'Y' AND b.T_cnt > b.A_cnt AND
              b.T_cnt > b.C_cnt AND b.T_cnt > b.G_cnt;"
# Zde je řešen případ, kdy je druhově specifická SNP pozice určena ve
# dvou bazích - ambiguidně, je tedy nezbytné dotazovat se adekvátním
# způsobem
   mysql -e "UPDATE ${i}_ASE a JOIN ${i}_var b USING (ref_acc,pos) JOIN
              var15_species_final c USING (ref_acc,pos) set a.balance =
              'ee' WHERE a.Pallel_FDR < 0.05 AND a.tt_ee = 1 AND ok = 1
              AND c.ee_a =1 AND c.ee_t =1 AND ((b.A_cnt > b.T_cnt AND
              b.A_cnt > b.C_cnt AND b.A_cnt > b.G_cnt) OR (b.T_cnt >
              b.A_cnt AND b.T_cnt > b.C_cnt AND b.T_cnt > b.G_cnt));"
```

Schéma č. 17: Výňatek bash scriptu popisující detekci disbalance ve směru ke genomu *ee* (MySQL, bash)

4.1.17 Statistická analýza ASE loci hybridních jedinců

Pro znázornění počtu disbalancovaných alel byl počítán počet všech utilizovatelných SNP a počet disbalancovaných, balancovaných SNP na gen (detekce počtu balancovaných SNP je obdobná, ptáme se zde, zda byla P hodnota menší než α 0.05 hypotézy eallel – tedy přijata nulová hypotéza o "vybalancovanosti" alel, ve schématech neuvedeno). Pro každý gen byl vypočten také medián, exponent průměru logaritmů P hodnot (P hodnoty v tomto případě nelze průměrovat) a to pro všechny SNP druhově

specifické s bazí pravděpodobného původu od rodičovských druhů. Postup je znázorněn ve schématu č. 18

```
#!/bin/bash
for i in $(cat ASE_diploid.tab)
do
# Tvorba tabulky obsahující počet všech použitelných SNP, balancovaných a
# disbalancovaných alel a to ve prospěch tt a ee genomu (počet
# balancovaných alel v genu pro přehlednost odstraněny)
    mysql -e "CREATE TABLE ${i}_ref_stat(
SELECT ref_acc,cnt_valid_SNP,cnt_ee,cnt_e,cnt_f,cnt_tt,cnt_t,cnt_u FROM
var15_species_final AS x
              LEFT JOIN (SELECT ref_acc, count(*) AS cnt_valid_SNP FROM
${i}_ASE WHERE tt_ee =1 AND ok =1 GROUP BY (ref_acc)) as a USING
(ref_acc)
              LEFT JOIN (SELECT ref_acc, count(*) AS cnt_ee FROM ${i}_ASE
WHERE tt_ee =1 AND ok =1 AND (balance = 'ee') GROUP BY ref_acc) AS b
USING (ref_acc)
              LEFT JOIN (SELECT ref_acc, count(*) AS cnt_tt FROM ${i}_ASE
WHERE tt_ee =1 AND ok =1 AND (balance = 'tt') GROUP BY ref_acc) AS e
USING (ref_acc)
              GROUP BY ref_acc);"
    mysql -e "ALTER TABLE ${i}_ref_stat add column median double, ADD
COLUMN PlogAVG double;"
    mysql -e "ALTER TABLE ${i}_ref_stat add PRIMARY KEY (ref_acc);"
done
# Výpočet exponentu průměru logaritmovaných P hodnot
for f in $(cat ASE_diploid.tab)
do
    mysql -e "UPDATE ${f}_ref_stat a JOIN (SELECT
              ref_acc,exp(avg(log(Pallel_FDR))) as PlogAVG FROM ${f}_ASE
              WHERE tt_ee = 1 AND ok =1 GROUP BY (ref_acc)) AS b USING
              (ref_acc) set a.PlogAVG = b.PlogAVG WHERE a.ref_acc =
              b.ref_acc;"

    for i in $(cat ref_acc.tab)
          do
# MySQL nemá funkci medián - nutno počítat obskurním způsobem
          mysql -e "UPDATE ${f}_ref_stat i JOIN (
          SELECT ref_acc,avg(t1.Pallel_FDR) as median_Pallel_FDR FROM (
          SELECT @rownum:=@rownum+1 as row_number, d.Pallel_FDR,
          d.ref_acc, d.ok, d.tt_ee
          FROM ${f}_ASE d, (SELECT @rownum:=0) r
          WHERE 1 AND ref_acc = '$i' AND tt_ee =1 AND ok =1
          ORDER BY d.Pallel2_FDR
          ) as t1,(
  SELECT count(*) as total_rows
  FROM ${f}_ASE d
  WHERE 1 AND ref_acc = '$i' AND tt_ee =1 AND ok =1
) AS t2
  WHERE 1 AND t1.row_number in (floor((total_rows+1)/2),
  floor((total_rows+2)/2) )) y USING (ref_acc) set
  i.median=y.median_Pallel2_FDR WHERE i.ref_acc = y.ref_acc;";
done
          done
```

Schéma č. 18: Postup získání dat pro statistickou analýzu – identifikaci ASE genů

```
SELECT a.med_Pallel et1, b.med_Pallel et2, c.med_Pallel et3, d.med_Pallel
eet1, e.med_Pallel eet2, f.med_Pallel eet3, g.med_Pallel ett1,
h.med_Pallel ett2, i.med_Pallel ett3, j.med_Pallel ett4 FROM
cab11L_ref_stat a
 left join cab22L_ref_stat b USING (ref_acc)
 left join cab23L_ref_stat c USING (ref_acc)
 left join cab07L_ref_stat d USING (ref_acc)
 left join cab08L_ref_stat e USING (ref_acc)
 left join cab09L_ref_stat f USING (ref_acc)
 left join cab16L_ref_stat g USING (ref_acc)
 left join cab17L_ref_stat h USING (ref_acc)
 left join cab24L_ref_stat i USING (ref_acc)
 left join cab25L_ref_stat j USING (ref_acc) WHERE a.med_Pallel < 0.05
AND a.cnt_valid_SNP >= 5 AND b.cnt_valid_SNP >= 5 AND c.cnt_valid_SNP >=
5 AND d.cnt_valid_SNP >= 5 AND e.cnt_valid_SNP >= 5 AND f.cnt_valid_SNP
>= 5 AND g.cnt_valid_SNP >= 5 AND h.cnt_valid_SNP >= 5 AND
i.cnt_valid_SNP >= 5 AND j.cnt_valid_SNP >= 5
    UNION
SELECT a.med_Pallel et1, b.med_Pallel et2, c.med_Pallel et3, d.med_Pallel
eet1, e.med_Pallel eet2, f.med_Pallel eet3, g.med_Pallel ett1,
h.med_Pallel ett2, i.med_Pallel ett3, j.med_Pallel ett4 FROM
cab11L_ref_stat a
 right join cab22L_ref_stat b USING (ref_acc)
 right join cab23L_ref_stat c USING (ref_acc)
 right join cab07L_ref_stat d USING (ref_acc)
 right join cab08L_ref_stat e USING (ref_acc)
 right join cab09L_ref_stat f USING (ref_acc)
 right join cab16L_ref_stat g USING (ref_acc)
 right join cab17L_ref_stat h USING (ref_acc)
 right join cab24L_ref_stat i USING (ref_acc)
 right join cab25L_ref_stat j USING (ref_acc) WHERE a.med_Pallel < 0.05
AND a.cnt_valid_SNP >= 5 AND b.cnt_valid_SNP >= 5 AND c.cnt_valid_SNP >=
5 AND d.cnt_valid_SNP >= 5 AND e.cnt_valid_SNP >= 5 AND f.cnt_valid_SNP
>= 5 AND g.cnt_valid_SNP >= 5 AND h.cnt_valid_SNP >= 5 AND
i.cnt_valid_SNP >= 5 AND j.cnt_valid_SNP >= 5;
```

Schéma č. 19: Popis získání dat pro korelační matici mezi jedinci všech genů, při čemž hybrid et má P medián

17.1 Analýza Müllerovy rohatky

4.1.18 Identifikace otevřených čtecích rámců v genech

Pro identifikaci ORF bylo aplikováno programu getorf (6.6.0.0) balíku embossy s parametry pro hledání ORF mezi stop kodóny, protože u dlouhých genů často chybí start kodón. Byl vybrán pouze nejdelší ORF – min daleky 87 bp, přičemž byly analyzovány pouze anotované, protein kódující sekvence. Sekvence ORF byly exportovány z databáze na základě pozic stop kodónu nejdelšího ORF užitím skriptu uvedeného v příloze. Celkem byly analyzováno 12432 cDNA sekvencí vybraných na základě výše zmíněných kritérií.

Z tabulek SNP RNAseq popsaných v kapitole č. 4.2.2 byly do sekvence ORF zapsány veškeré nalezené SNP pro daného jedince a to vzorků jater, tak oocytů. Tento zápis polymorfismu do sekvence byl proveden jazykem perl – funkce substring (modifikovaný perl skript pro vnesení majoritní báze v případě identifikace heterozygotních pozic ze 454 dat (Mgr. Jan Pačes, Ph.D.)

4.1.19 Výpočet dN/dS poměru z párového srovnání

Pro otestování hypotézy Müllerovy rohatky jsem provedl výpočty poměru nesynonymních a synonymních mutací (dN/dS) pro různé biotypy, včetně ET hybridů. Dále jsem stejný výpočet provedl pro uměle vytvořené ET hybridy, které jsem získal náhodným zkombinováním odpovídajících sekvencí rodičovských druhů. Tyto rodičovské sekvence jsem nejdříve zduplikoval, poté jsem každé variantní pozici (SNP) náhodně (ale unikátně) přiřadil bázi tak, že každý duplikát obsahoval jinou variantu. Takto vzniklé sekvence z rodičovských druhů jsem zkombinoval mezi druhy za vzniku umělých F1 hybridů, na kterých jsem opět měřil dN/dS poměr. Ostatní sekvence jsem upravil stejným způsobem a všechny takto upravené sekvence jsem následně srovnával v rámci biotypů. Cílem bylo získat představu o distribuci dN/dS poměru v rámci jednotlivých druhů a biotypů, včetně laboratorních F1 ET hybridů, stejně jako v rámci mnou vytvořených in silico ET hybridů.

Poměry dN/dS pro všechny páry sekvencí jsem počítal opět ve statistickém prostředí R. Import a alignment sekvencí jsem provedl s pomocí knihovny ape (Paradis et al. 2004), samotný výpočet pak s pomocí knihovny seqinr (Charif and Lobry 2007), která pro výpočet dN/dS poměru používá model LWL85 .(Li 1993)(Zhang and Yu 2006)

Sekvence, které neobsahují žádnou synonymní a/nebo nesynonymní mutaci, a mají tedy v čitateli a/nebo jmenovateli zlomku nulu, představují problém. Ten lze vyřešit například přičtením čísla 1 ke každé hodnotě dN a dS ještě před výpočtem jejich podílu (Paradis et al. 2004) (Bajgain et al. 2011) (Novaes et al. 2008), což ale přináší nová úskalí. Předně sekvencím bez mutací (a tedy bez informace) je chybně přiřazen dN/dS poměr roven 1, tedy neutrální. Navíc přičtení 1 k hodnotám dN a dS přinejmenším na mých datech způsobovalo nahloučení dN/dS hodnot kolem neutrální hodnoty 1 v rozsahu, který znemožňoval rozumnou vizualizaci dat.

Využil jsem tedy toho, jak software R řeší dělení nulou. V R je zlomku 0/0 přiřazeno jako výsledek "NaN" (Not a Number). Naproti tomu dělením kladného čísla

nulou získáme hodnotu "Inf" (infinity, tedy nekonečno). Zatímco první případ popisuje situaci, kdy nemám dost informací k vyvození závěru o selekčním tlaku na danou sekvenci, druhý případ nějakou informaci nese. První případ by měl být tedy z výpočtů vyřazen, kdežto druhý by měl zůstat zachován.

Vytvořil jsem tedy dvě matice: První matice zahrnovala úpravu hodnot dN a dS přičtením čísla 0.01, které se ukázalo být dobrým koeficientem při následné vizualizaci dat. Druhá matice tuto úpravu neobsahovala a sloužila k poskytnutí souřadnic neinformativních hodnot ("NaN"), což mi umožnilo jejich vyfiltrování z první matice.

4 Výsledky

18.1 Evaluace referenční sekvence

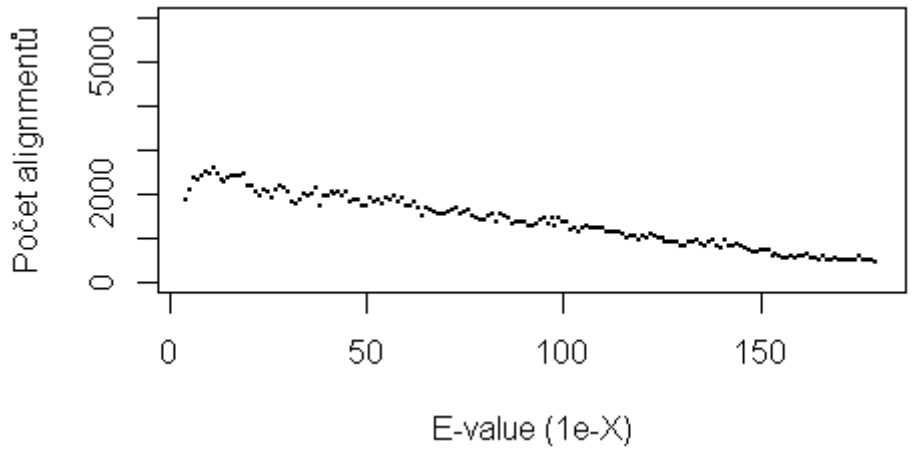
Jedním z prvních výsledků, na které navazují veškeré dalších analýzy této práce, je samotná příprava a vyhodnocení referenční sekvence.

Délka sekvence má na úspěšnost anotace velký význam, protože se snižující se délkou sekvence stoupá pravděpodobnost, že sekvence v databázi může být zcela randomní kompozice (e-value se snižuje exponenciálně s délkou sekvence). Jelikož limit délky cDNA byl nastaven pro sekvence delší než 300, je význam délky cDNA na anotaci již marginální, nicméně u krátkých sekvencí je úspěšnost anotace nižší, viz graf č. 7. V transktriptomu by ale mohly být zastoupeny lncRNA, je tedy zřejmé, že z porovnání obsahu parametrů GC (viz graf č. 22), ORF (viz graf č. 23) a délek sekvencí mezi anotovanými a neanotovanými *datasety* (viz graf č. 24) vyplývá, že část těchto sekvencí je zřejmě nekódujících. Detekce dlouhých nekódujích RNA je obzvláště složitá, neboť většina lncRNA je nestrukturovaná (není tomu tak v případě, kdy je prekurzorem malých RNA) a lze ji charakterizovat pouze na základě původu z intronových, či intergenových pozic obsahující v některých případech i regulační elementy transkripce. Homologie mezi lncRNA bývá často také nízká (Wang et al. 2013). Pro detekci lncRNA byly zvoleny dva přístupy detekce na základě homologie čistě sekvenční (blastn) a také kombinaci sekvenčního a strukturního s přístupu Rfam (databáze verze sekvencí 11.0) softwaru využitím modelů kovariance (Burge et al. 2012).

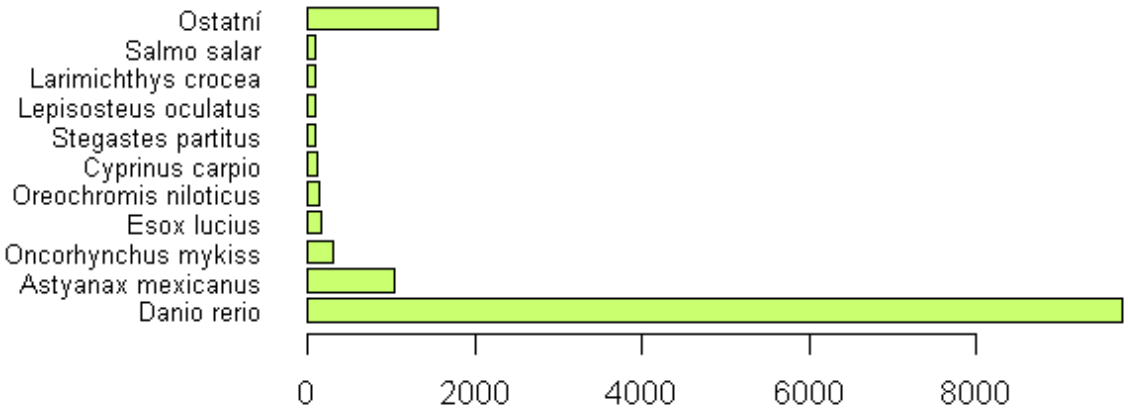
Nekódující RNA byla analyzována dvěma způsoby: blastn proti databázi všech nekódujících RNA *Danio rerio* (Ensemble, 17.02.15). Blastn bylo nalezeno 114 ncRNA pod prahovou hodnotou $1x10^{-6}$ čehož 108 sekvencí získalo *bitscore* větší než 80. Rfam

přístupem na základě definovaných, známých strukturních modelů kovariance bylo označeno 143 sekvencí jako ncRNA. Průnik těchto množin je mizivý - činí pouhých 8 cDNA sekvencí. Z frekvence označených lncRNA vyplývá, že množství nekódujících sekvencí v transkriptomu pravděpodobně nemá valný vliv na úspěšnost anotace, ani u ostatních modelových organismů kostnatých ryb nebyl nelezen exces lncRNA. Naopak zřejmě validní premisou může být přítomnost nebetyčného množství UTR sekvencí, především pak 3' URT, vycházíme-li z přípravy reverzní transkripcí od 3' polyA. Bohužel detekce je opět problematická, zaměřuje se především na přítomost polyA signálu a dalších regulačních motivů, pro jejichž detekci je nutné aplikování algoritmů strojového učení s učitelem.

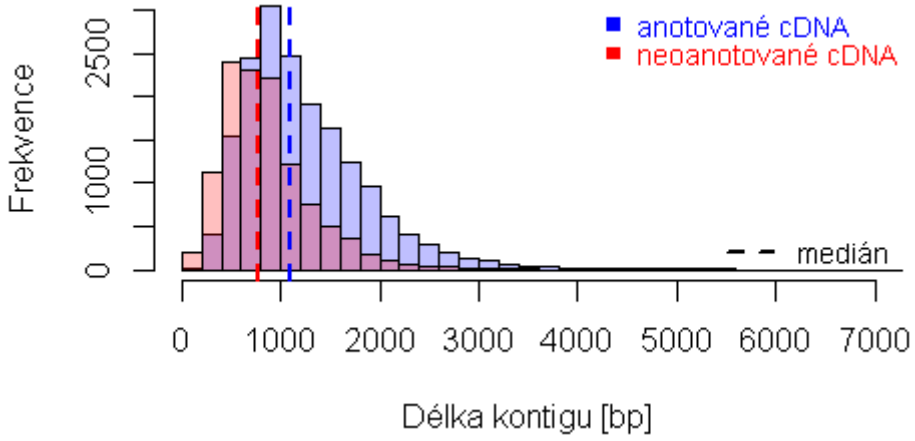
Při přípravě cDNA zřejmě nedošlo k „závažné“ kontaminaci, ať již parazitem, či následkem nesterilní přípravy. Naprostá většina sekvencí náleží kostnatým rybám. V datech se objevilo přibližně 100 sekvencí přiřazených k savčím a bezobratlých živočichům, nejvíce pak lidských (22) a myších (21), tyto sekvence vysazují velmi nízkou evalue - zřejmě se jedná o kontaminace, tudíž byly z následujících analýz vyloučeny. Četnost nejlepších *alignmentů* podle druhů je uvedena v grafu č. 20; distribuce *evalue* je uvedena v grafu č. 21.



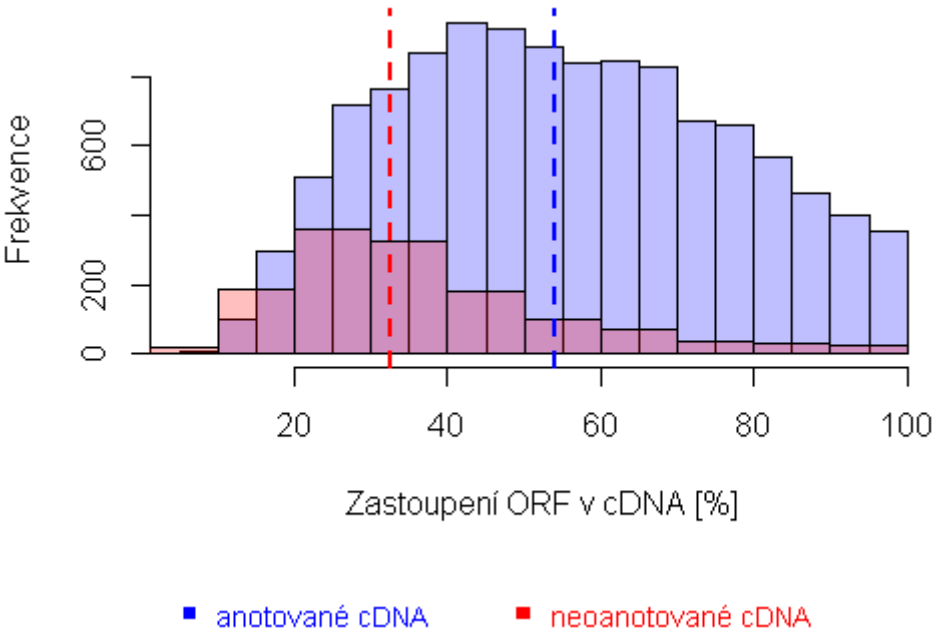
Graf č.: 21: Rozdělení eValue hodnot sekvencí vzhledem k počtu alignmentů (na ose x je uveden pouze exponent X: 10^{-X}).



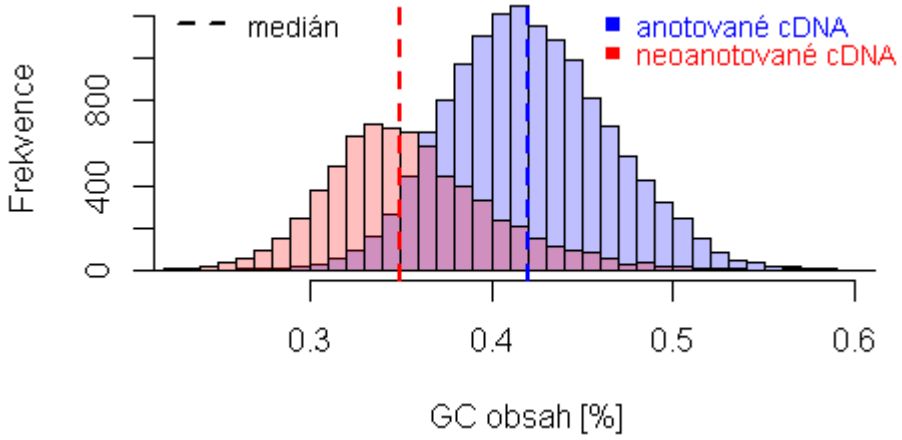
Graf č. 20: Sloupcový diagram znázorňující frekvenci deseti druhů s nejvyšším počtem alignentů a sumu všech ostatních druhů - v kategorii ostatní.



Graf č. 24: Histogram, komparace rozdělení délek kontigů anotovaných a bez anotace

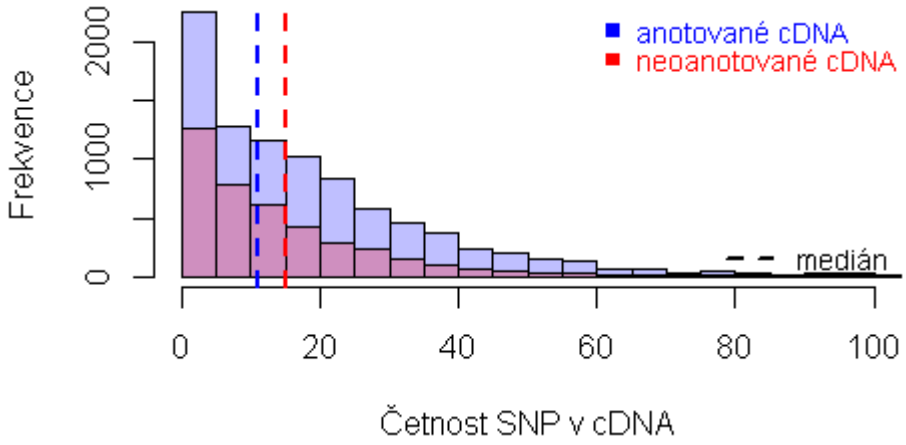


Graf č. 23.: Porovnání souborů rozdělení procentuálního zastoupení otevřeného čtecího rámce (ORF) v rámci cDNA mezi anotovanými a neanotovanými soubory sekvencí.



Graf č.: 22: Histogram porovnání procentuálního obsahu GC v setu anotovaných a neoanotovaných cDNA.

V neanotovaných sekvencích byl také identifikován vyšší počet polymorfních pozic, viz graf č. 25, což je dalším indikátorem přítomnosti excesu sekvencí, jež nejsou pod silným selekčním tlakem (Mann-Whitneův pořadový test – zamítáme nulovou hypotézu o shodě rozdělení veličin s P hodnotou < 2,2 x 10^{-16}).



Graf č.: 25: Histogram distribuce frekvence polymorfismů na cDNA (četnost SNP není normalizována na délku sekvence, neboť neanotovaný soubor sekvencí má nižší medián délky)

19.1 Evalace SNP polymorfismů

Pro získání SNP bylo využito několik algoritmických a statistických přístupů, navíc dat sekvenovaných různými platformami. Základní databázová sestava SNP vycházející ze

454 sekvenovaných normalizované cDNA tkáně oocytů a jater má za cíl identifikovat maximum vysoce důvěryhodných SNP přijatelného sekvenačního pokrytí, které nejenže umožnily mapování na ambiguidní referenci - pro vyváženost mapování vše analyzovaných druhů, ale též pro získání informací o původu alely – druhově determinující pozice. Tyto SNP byly získány velmi přesným *mapperem* Newbler a dodatečně upraveny pro maximalizaci výtěžku, řešením i problematicých pozic, které samotný program nebyl schopen analyzovat.

Na takto detekovaných pozicích byly řešeny SNP dat RNAseq, nenormalizované cDNA technologii illumina, z důvodu absence programového přístupu řešení dat, vzhledem k refrenci se ambiguidními pozicemi byl zvolen statistický přístup chi-kvadrát testování, který dnes není standartně používán na řešení homo- a heterozygotních pozic, z chi distribuce, ale stále vychází mnoho analýz aplikující likelyhood testování. Chi-kvadrát test je recentně aplikován pouze na testování Hardy-Weinbergovy rovnováhy alel v populaci.

V posledním případě byly získany SNP mapovaných na totožnou referenci, avšak bez ambiguidních pozic. Byly tedy detekovány i takové pozice, které nebyly definovány na základě 454 reference a to standartním postupem programu Bcftools aplikující likelyhood testování. Mezi získanými sety SNP je rozdíl pouze v průměrné heterozygotnosti analyzovaných jedinců výše uvedených přístupů, testování chi kvadrát určuje přibližně o čtvrtinu více heterozygotních pozic, které se ale v průniku pozic obou přístupů shodují.

V případě SNP získaných SeqCap bylo nutné odstranit problematické pozice v místech intron-exon hranice, protože byly mapovány sekvence genomu na cDNA. U těchto dat byla sekvenační hloubka naopak často až příliš velká, proto byla horní hranice sekvenační hloubky stanovena 20000; statistika v takovýchto případech neumí pracovat se sekvenačními chybami a může je chybně považovat za pravé substituce.

Veškerá získaná data jsou získaná z dat min kvality phre skóre vetší než 20.

20.1 Diferenciální genová exprese

V prvé řadě bych se rád zaměřil na výsledky získané z diferenciální exprese. Hlavní premisou v této analýze bylo detekovat rozdíly mezi skupinami sexuálně a asexuálně se rozmnožujícími jedinci, zejména pak na oocyty 6. stádia vývoje, jimž je také věnována největší pozornost. Druhořadně se výsledky zaměřují na prezentaci DE genů, které vznikly následkem polyploidizace a DE genů vycházejících z mezidruhových rozdílů.

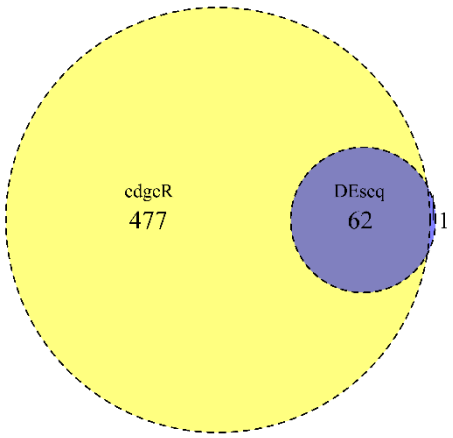
Sumarizace cílů detekce diferenciální exprese:

1) Analýza diferenciálně exprimovaných genů (DE) mezi klonálně se reprodukujícími hybridy a „čistými“ druhy reprodukujícími se pouze sexuální cestou.
2) Identifikace DE genů mezi polyploidními a diploidními jedinci.
3) Detekce DE genů mezi jednotlivými genotypy hybridů z pohledu impaktu na analýzu DE mezi diploidními a polyploidními jedinci
4) Identifikace DE genů mezi druhy *C. elogatoides* a *C. taenia* (a to dvěma přístupy detekce: testovaní skupin *fitovaných* na negativně binominální model a také přístupem zjištění nejvyšší divergence ve směru gradientu PCoA *C. taenia* a *C. elongatides*, vysvětleno v kapitole č.) vzhledem k nutnosti determinovat rozdíly exprese spjatou s divergencí druhů.
5) V poslední řadě se pozornost upíná k otázce, zda si hybridní jednici zachovávají na globální úrovni genomu původní míru regulace, či zda jeden z rodičovských genomů je zcela, nebo částečně imprintován, nebo dokonce zda parentální determinace transkripce není alterována jevem zvaným „genomic shock“, který se projevuje ve ztrátě alel - LOH, problémy s párováním chromosomů, deregulace metylace, aktivace retrotrnaspozónů. To vše může být příčinou transkripce generálně značně vychýlené od oboru parentálních druhů (Wang et al. 2015). Analýza imprintingu je řešena detekcí alel specifické exprese – kapitola 4.3. Cytogenetické aspekty gynogenetického rozmnožování jsou rozvedeny v úvodu. Analýzu zvýšené retrotranspozice nelze prozatím analyticky pojmout, protože reference se složena pouze z druhu *tt,* nejsme tedy schopni rekonstruovat RNA retrotranspozonů, která by v tomto druhu za normálních okolností neměla být nabohacena.

Je nezbytné si uvědomit, že oocyty uvedeného stádia jsou téměř transkripčně neaktivní a většina mRNA je pouze maternálním pozůstatkem, který může určovat vývoj embrya. Naše pozorování diferenciální exprese týkající se oocytů je pouze následek již proběhnuvší determinace na gynogeneticky se replikující embryo. Problematickým fenoménem vnášejícím do dat diferenciální exprese mezi oocyty hybridů a sexuálně se reprodukujícími druhy je polyploidie, ta může nebo nemusí, a to zcela nepredikovatelně, změnit transkripční profil organismu. Ve třetím případě se u hybridů, a to tkáně

jakéhokoliv původu, mohou projevit expresní mezidruhové rozdíly (změny v cis- trans regulaci exprese RNA), které jsou akumulovány mnohem rapidněji, nežli substituce v germinální linii. Naším primárním cílem diferenciální exprese oocytů skupin rozdělených dle formy rozmnožování je odhalit geny, které by mohly stát za funkční příčinou vzniku gynogenetického rozmnožování.

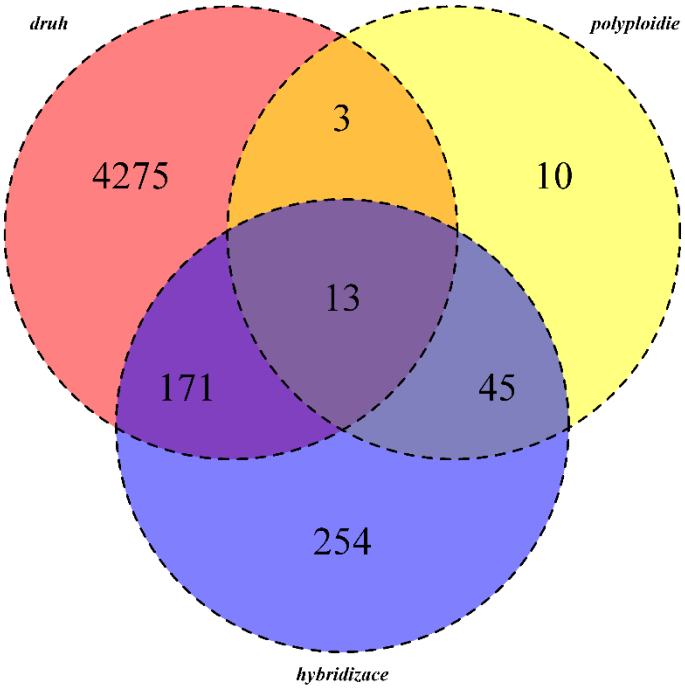
V grafu č. 25 jsou porovnány množiny detekovaných DE genů, mezi skupinou hybridů a rodičovských druhů pocházejících z ooytů, získanými dvěma přístupy detekce: edgeR a DEseq. DE geny získané programem DESeq jsou téměř podmnožinou DE genů získaných přístupem edgeR, ale výběrem na stejné hodnotě alfa byly získány velmi rozdílné počty DE genů. Program DEseq stanovuje obecně P hodnoty mnohem vyšší, a může tak zamítat validní výsledky, proto budou veškeré výsledky analyzovány pouze přístupem programu edgeR,



Graf č. 25: Vennův diagram množin DE genů nalezených dvěma přístupy selekcí na hladině α = 0.05 (edgeR a DESeq) srovnáním skupin asexuálních a sexuálně se reprodukujících skupin tkáně **oocytů**

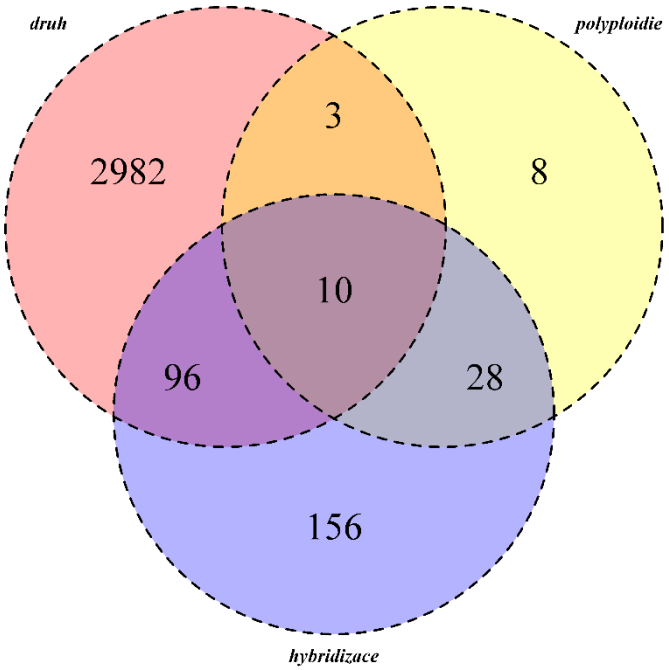
Z celkového počtu 483 DE genů oocytů výše uvedených skupin (127 podexprimováno u hybridů, 356 nadexprimováno z celkového počtu 15835 cDNA reference), všechny tyto geny ale nemůžeme označit jako DE geny s kýženým biologickým významem tj. geny, které jsou zodpovědné za gynogenetickou formu rozmnožování, potažmo klonální genezi embrya. Významným faktorem produkující rapidní změny exprese u hybridních forem rodičovských druhů je polyploidie, především pak u organismů lichého počtu chromosomových sad - anorthoploidie. Hlavním činitelem co do počtu DE genů je v našem případě diference mezi srovnávanými druhy ryb rodu *Cobitis* (*tt* vůči *ee*). V grafu Vennova diagramu č. 26 je znázorněn *intersekt* těchto tří množin: DE geny mezi druhy *ee* versus *tt*, DE geny srovnání diploid versus polyploid a DE geny z komparace

asex- forem vůči sexuálně se množícím rodičovský druhům. Ze srovnání byli odstraněni jedinci obsahující haplotyp *n*, protože k hybridním formám nemáme jejich rodičovské druhy.

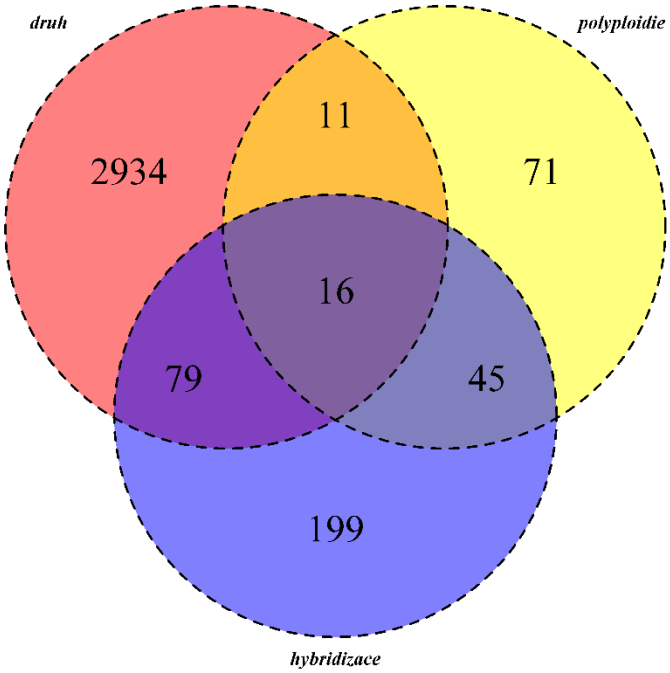


Graf č.: 26: Vennův diagram množiny párového srovnání nalezených DE genů tkáně **oocytů** mezi druhy *ee*, *tt* = kategorie **druh**; množiny DE genů získaných srovnáním diploidních a polyploidních jedinců = kategorie **polyploidie** a množiny DE párového srovnání gynogeneticky a sexuálně se reprodukujících jedinců = kategorie **hybridizace**

Pokud bylo totožné srovnání množin DE genů provedeno na datech, ze kterých byly vyřazeny neanotované geny, zůstaly poměry průniků množin velmi podobné, viz graf. č. 27. Poměrové zastoupení DE genů mezi uvedenými skupinami odpovídá grafu č. 26. V jaterní tkáň je celkový počet DE genů vše srovnávaných skupin je naopak nižší nežli v oocytech, došlo ale k znatelnému navýšení DE genů mezi skupinami di- a polyploidů, viz graf č. 28.



Graf č.: 27: Vennův diagram množiny párového srovnání nalezených DE genů tkáně **oocytů pouze anotovaných genů (selektovaných předem)** mezi druhy *ee*, *tt* = kategorie **druh**; množiny DE genů získaných srovnáním diploidních a polyploidních jedinců = kategorie **polyploidizace** a množiny DE párového srovnání gynogeneticky a sexuálně se reprodukujích jedinců = kategorie **hybridizace**

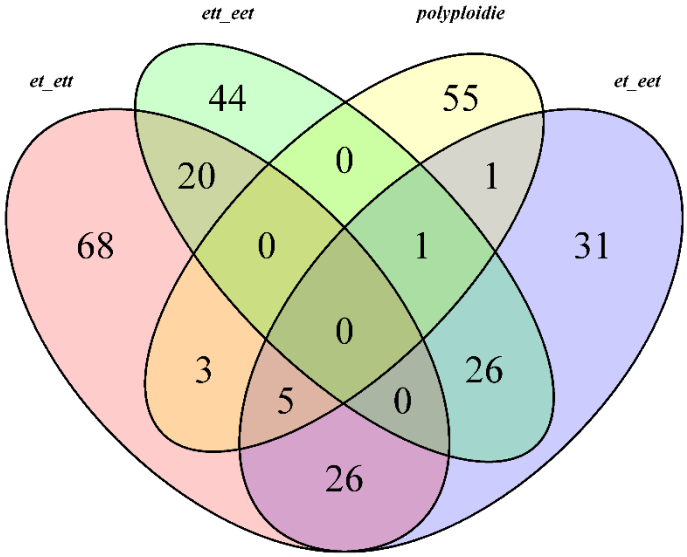


Graf č.: 28: Vennův diagram množiny párového srovnání nalezených DE genů tkáně **jater** mezi druhy *ee*, *tt* = kategorie **druh**; množiny DE genů získaných srovnáním diploidních a polyploidních jedinců = kategorie **polyploidnie** a množiny DE párového srovnání gynogeneticky a sexuálně se reprodukujích jedinců = kategorie **hybridizace**

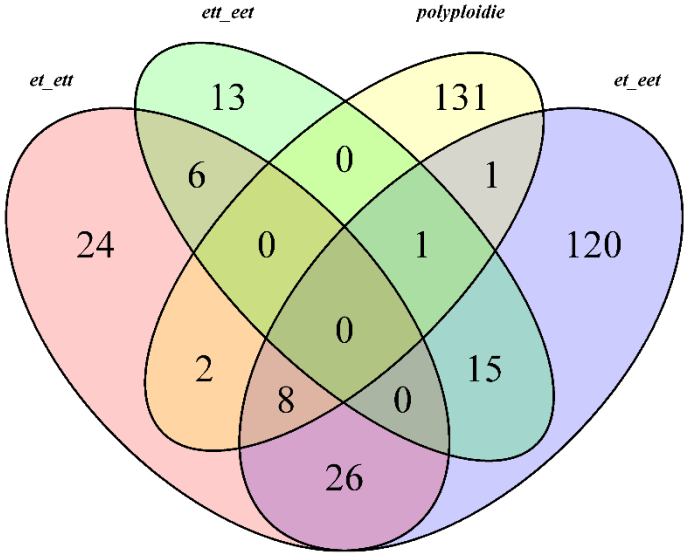
Ze zkoumání počtu DE genů mezi diploidními a triploidními jedinci se jeví, že polyploidizace nemá na hybridní jedince valný vliv, jelikož je počet nalezených DE genů

minimální. Nicméně srovnáváme-li několik různorodých skupin hybridů, přičemž každá kombinace haplotypů může generovat unikátní sadu DE genů způsobenou polyploidním šokem. Při globálním srovnání di – a polyploidů mohou být jednotlivé individuální rozdíly v rámci konkrétních specifik hybridů odfiltrovány, proto nelze jednoznačně tvrdit, že polyploidizace nemá vliv na diferenci v expresi. Graf č. 29 a 30 indikuje, že ani rozdíly na úrovni párové komparace jednotlivých typů hybridů mezi sebou nehrají významnou roli, neboť nebyl zaznamenán exces DE genů mezi skupinami hybridů. Nejvíce DE genů pak bylo nalezeno mezi skupinami triploidů *ett* vůči *eet*, mezi kterými je pozorovatelný nejznatelnější nárůst počtu DE genů, u nichž se může jednat o rozdíly na úrovni exprese mezidruhové; pokud dochází v hybridních jedincích k "zprůměrování" regulace cis exprese mezi rodičovskými haplotypy.

Na základě této premisy byli jednotliví jedinci rozdílných kombinací haplotypů rodičovských druhů srovnáni s oběma rodičovskými druhy tkáně jater a oocytů, viz graf č. 31, 32, 34, 33. Znázorněné průniky množin indikují vliv zastoupení haplotypu hybrida na množství DE genů. Jinými slovy srovnáme-li triploidního hybrida *ett* s druhem *ee*, narůstá počet genů, které vycházejí z rozdílu mezidruhového nikoliv vlivem ploidie. V grafu č. 35 je vyobrazen namísto polyploidie množina DE genů mezidruhových rozdílů mezi druhy *ee* a *tt*. Majoritní část DE genů množiny *tt* vůči *eet* náleží k mezidruhovým rozdílům – 87 % DE genů, mezi skupinami *tt* a *ett* náleží 70 %, což je neočekávaně vysoké číslo a mezi *tt* – *et* rozdíly mezi druhy *tt* a *ee* činí 42 %.



Graf č. 29: Vennův diagram tří skupin hybridů porovnaných mezi sebou (množiny DE genů komparace *et* versus *ett*, *et* - *eet*, *ett* - *eet*) tkáně **oocytů**



Graf č. 30: Vennův diagram tří skupin hybridů porovnaných mezi sebou (množiny DE genů komparace *et* versus *ett*, *et* - *eet*, *ett* - *ee*t) tkáně **jater**

Graf č. 31: Vennův diagram množin DE genů z porovnání tří skupin **hybridů** s rodičovským druhem ***ee*** tkáně **jater**

Graf č. 32: Vennův diagram množin DE genů z porovnání tří skupin **hybridů** s rodičovským druhem ***ee*** tkáně **oocytů**

Graf č. 33: Vennův diagram množin DE genů z porovnání tří skupin **hybridů** s rodičovským druhem ***tt*** tkáně **jater**

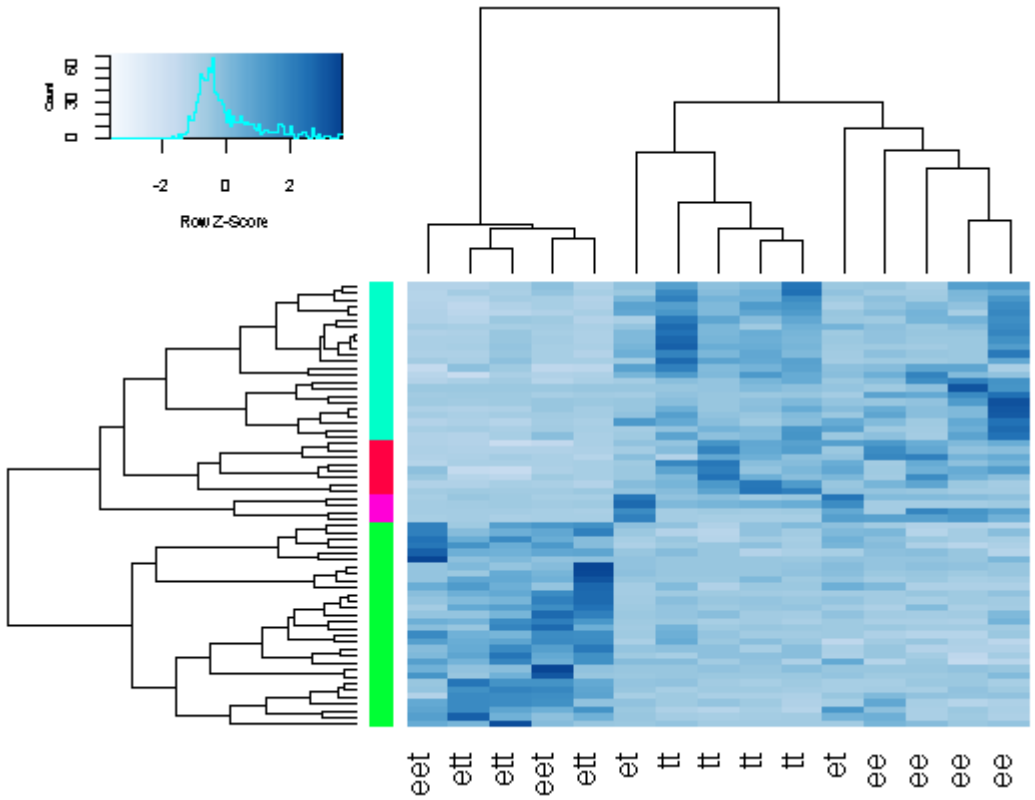
Graf č. 34: Vennův diagram množin DE genů z porovnání tří skupin **hybridů** s rodičovským druhem ***tt*** tkáně **oocytů**

Graf č. 35: Vennův diagram množin DE genů z porovnání tří skupin **hybridů** s rodičovským druhem ***tt*** tkáně **oocytů**

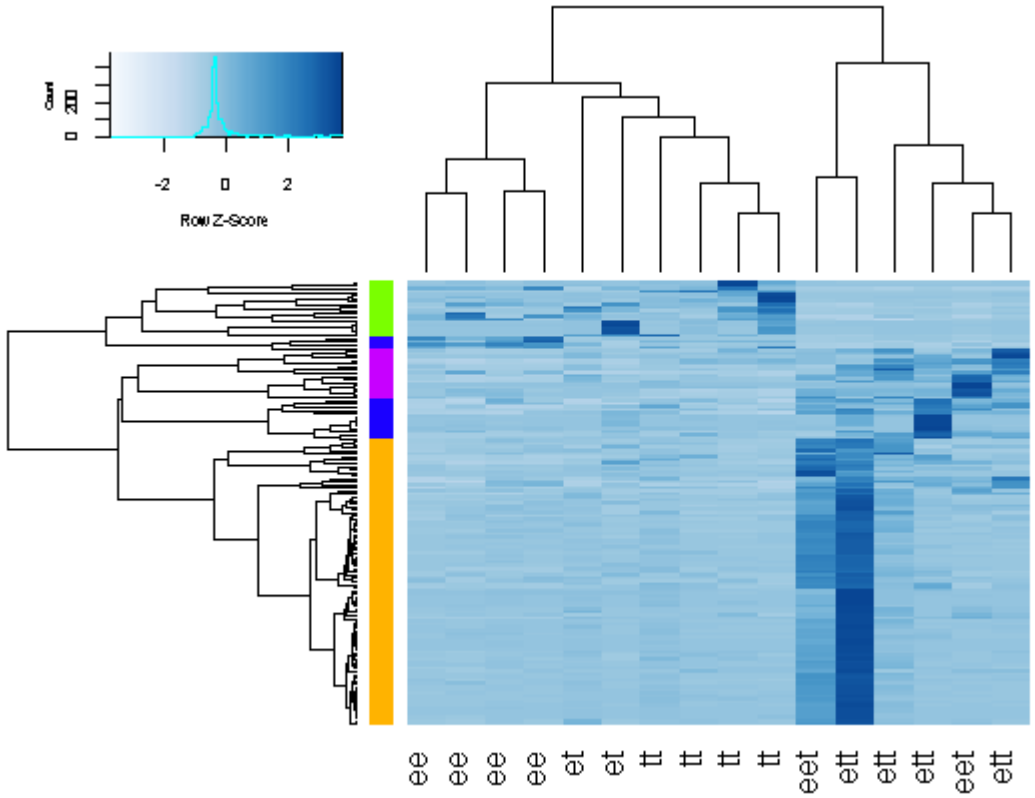
Z funkčního hlediska byly nalezeny nabohacené GO termy mezi DE geny získaných ze srovnání vzorků jater skupiny di- versus polyploidních jedinců. U oocytů nebyly nalaeny žádné signifikantní změny na FDR korigované α 0.05. Celkem bylo mezi jaterními vzorky nalezeno celkem 143 DE genů, z toho 22 podexprimovaných u diploidů a 121 nadexprimovaných u polyploidů, v případě oocytů je jedná o 35 a 30 stejného pořadí. V grafu č. 36 a 37 je znázorněn celkový pohled na kvantitativní rozdíly a podobnost vzorků. V řádcích *heatmap* jsou prezentovány TMM normalizované počty readů na cDNA, intenzita modré barvy indikduje míru exprese. Dendrogramy ve sloupcích jsou vyjádřením hierarchického klastrování – distance nejvzdálenějšího souseda (*complete method*) na základě hodnot spearmanovu korelace mezi vzorky (N vzorků je nízké, data nemají normální rozdělení), v řádcích je užita metoda totožného hierarchického klastrování, nicméně vychází z hodnot korelačního koeficientu dle Pearsona, protože můžeme pracovat daty normálního rozdělení. Popis tvorby heatplot je uveden v metodice Graf č. 36 a 37 tedy uveden zaprvé pro kontrolou experimentu – konkrétně vyjádření podobnosti analyzovaných skupin di – versus polyploidní jedinci. Dále toto zobrazení podává informací o divergenci genů mezi vzorky a genových skupinách sloučených na základě podobnosti v expresi. V grafu č. 37 je patrný problém se vzorkem, chovající se jakou odlehlý vůči ostatním vzorkům své kategorie; je též možné, že při DE analýze vznikla značná část genů dík tomuto vzroku. Nabohacené kategorie jsou znázorněny v grafu č. 38.

Jedná se o geny s funkcemi jako buněčný transport, či vazba heterocyklických sloučenin, DNA.

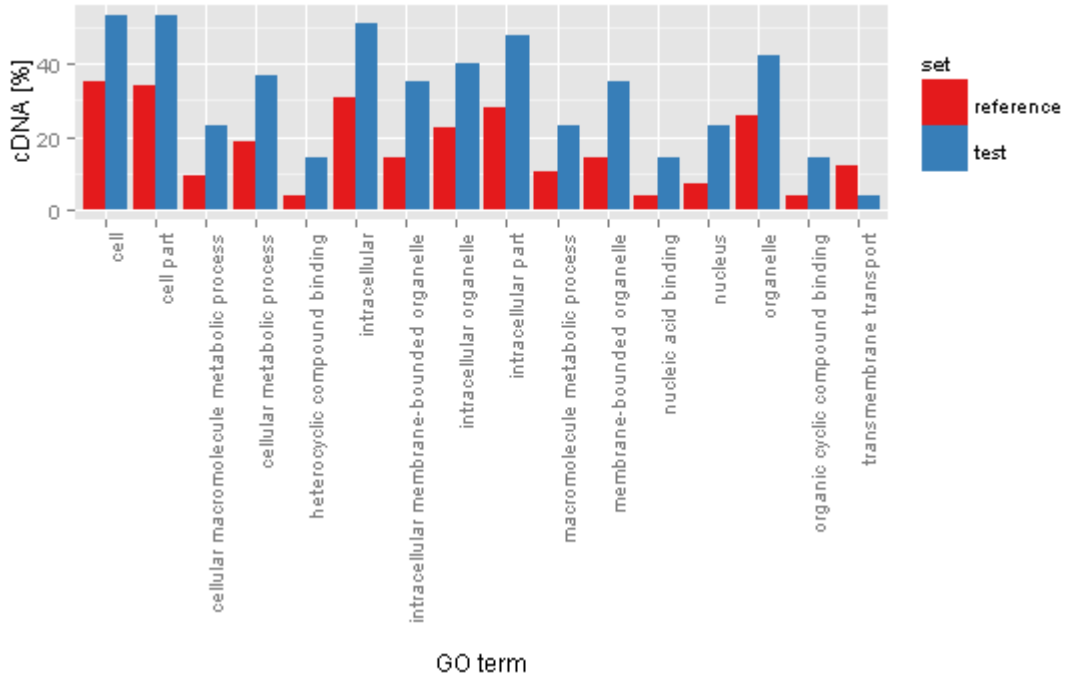
Bližší pohled na podobnosti DE genů v expresi mezi skupinami di- versus polyploidní jedinci oocytů a jater se naskýtá v grafech č. 39 a 40. Similarita exprese mezi geny může být následkem čistě stochastických jevů, nicméně častým jevem živých systémů je přítomnost koexprese, která může mít několik příčin. Geny mohou být například v silné vazbě a sdílet regulační oblast nebo mohou být přítomny ve stejné funkční, metabolické dráze, kdy je transkripci nutno oboustranně regulovat. Cílem je detekovat geny koexprimující, *koinherentní*, duplikované (předpoklad totožné regulace), geny pod koordinovanou epigenetickou kontrolou, či geny postižené dávnou genovou konverzí. Pro přehlednost byly vybrány DE geny s největší absolutní změnou exprese mezi vzorky (*fold change*). Byla zvolena metoda shlukování na základě nejpodobnějšího mediánu mezi klastry, abychom se vyvarovali efektu *outlier* vzorků, viz graf č. 40,41. K těmto DE geny byly přiřazeny anotace pro bližší prozkoumání jejich biologické souvislosti. Seznam anotovaných DE geny mezi skupinami uvedenými v grafu č. 26 je přiložen v apendixu práce.



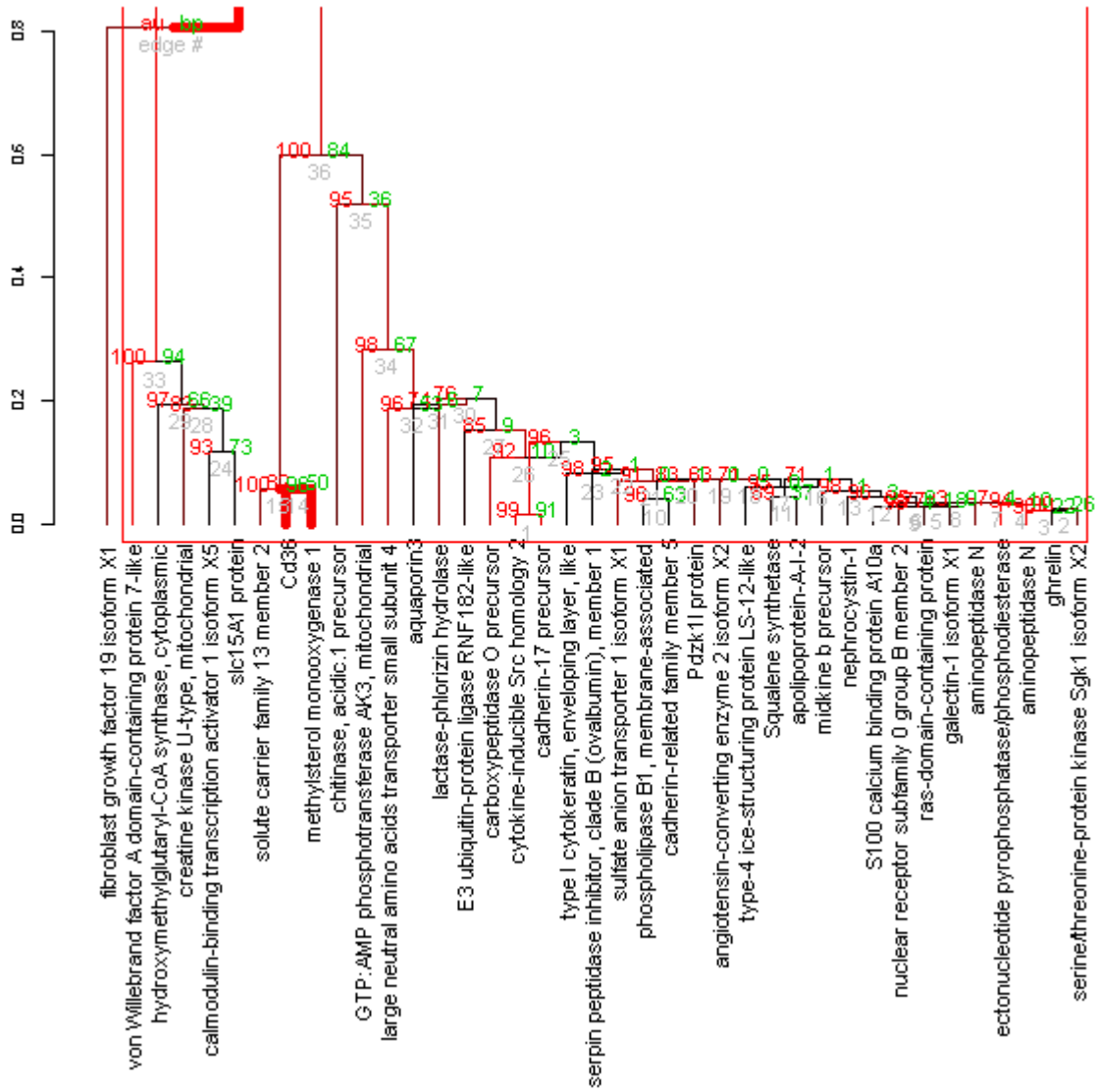
Graf č. 36: Heatplot – zobrazení míry exprese jednotlivých DE genů vzniknuvších komparací **diploidních** a **polyploidních** jedinců vzorků **oocytů** (**s intenzitou modré barvy stoupá míra exprese – normalizovaná data, bílá - modrá**). Dendogramy na horizontální ose vyjadřují výsledek shlukovací analýzy korelační koeficientů mezi geny, na vertikální ose vyjadřuje podobnost mezi vzorky.



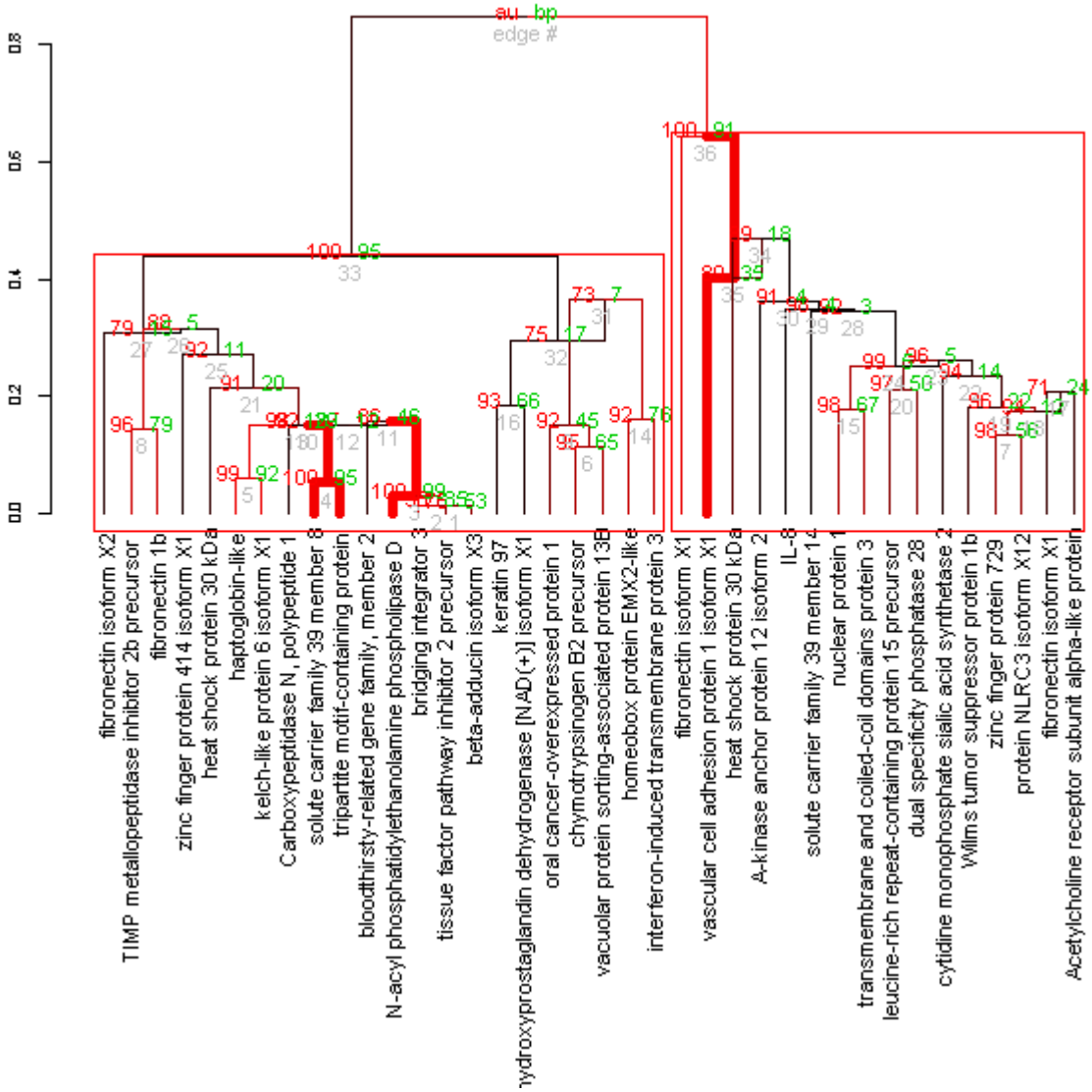
Graf č. 37: Heatplot – zobrazení míry exprese jednotlivých DE genů vzniknuvších komparací **diploidních** a **polyploidních** jedinců vzorků **jater** (**s intenzitou modré barvy stoupá míra exprese – normalizovaná data, bílá - modrá**). Dendogramy na horizontální ose vyjadřují výsledek shlukovací analýzy korelační koeficientů mezi geny, na vertikální ose vyjadřuje podobnost mezi vzorky.



Graf č. 38: Explorace nabohacených GO term identifikátorů v setu DE genů mezi **di** – a **polyploidními** jedinci vzorků **jaterní tkáně**. Procento výskytu v referenčním setu anotovaných genů je znázorněn červeně, zatímco procento GO term v testovaném setu je znázorněn modře.



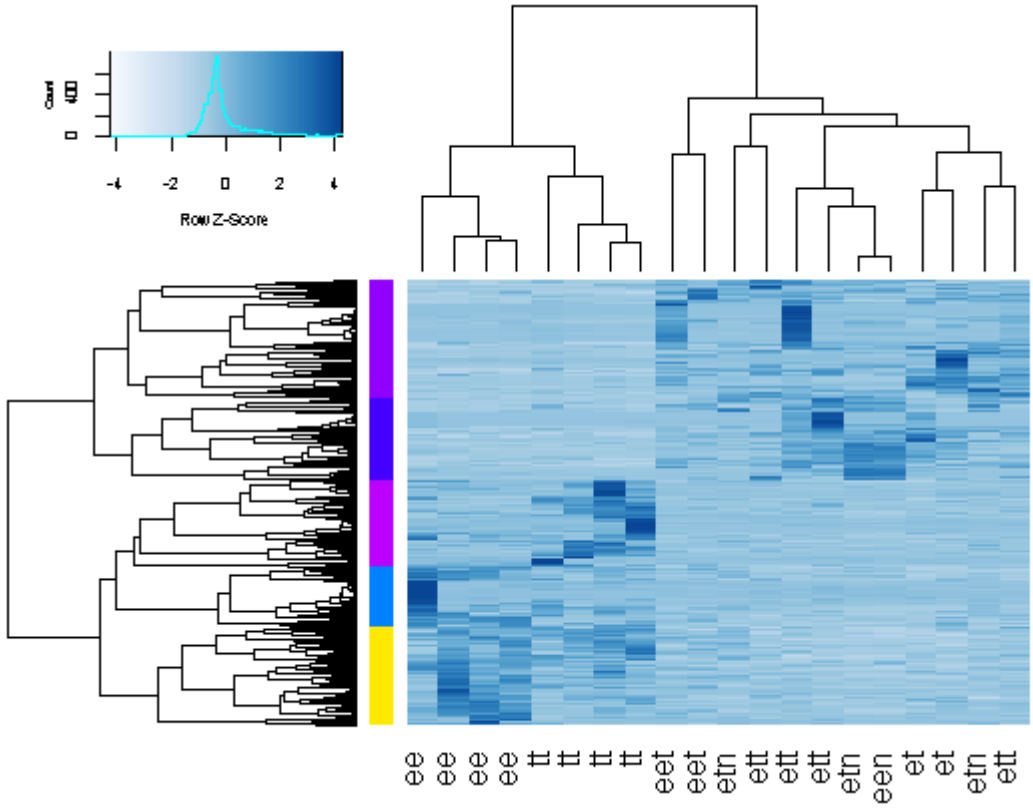
Graf č. 39: Hierarchické klastrování **40** DE geny s největší logFC **DE genů přítomných pouze mezi skupinami di- vs polyploidní jedinci tkáně jater**; *similarita* mezi geny je vyjádřena formou dendrogramu. Bootstrap P hodnoty jsou vyjádřeny červeně, AU – *aprroximately unbiased* hodnoty zeleně. Červené obdélníky vyznačují signifikantní větvení; tučné červené linky, či tučné červené linky značí míru signifikance – spolu s uvedenými AU a BP hodnotami podpory větvení.



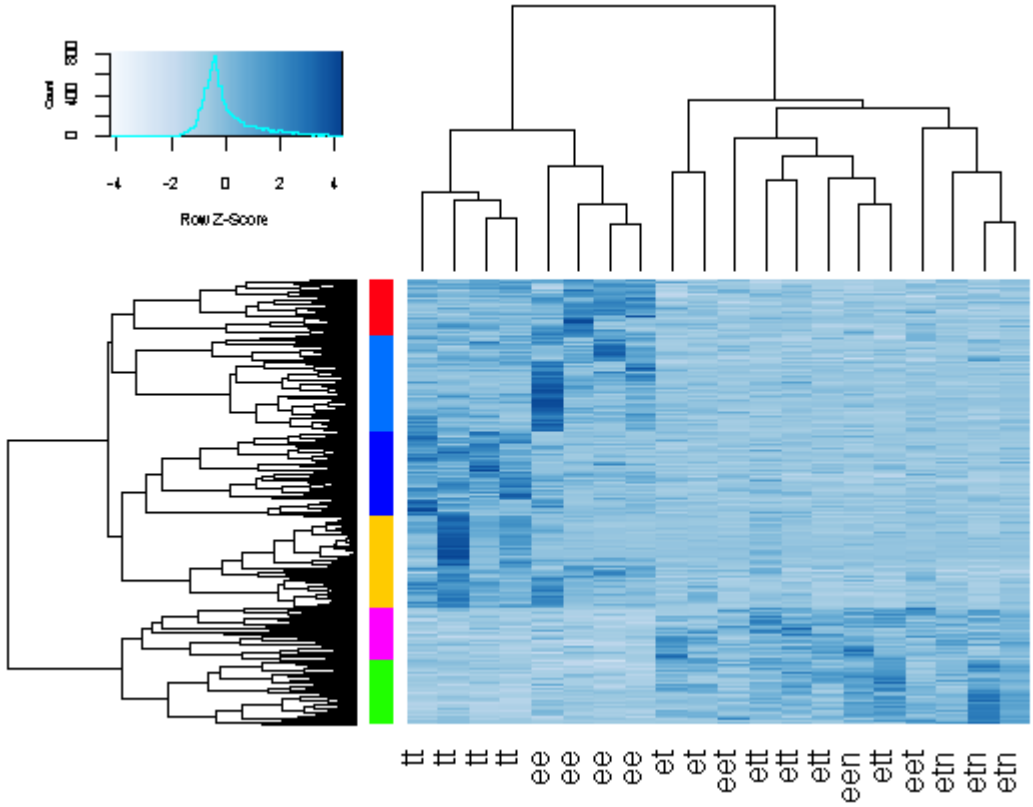
Graf č. 39IK: Hierarchické klastrování **40** DE geny s největší logFC **DE genů přítomných pouze mezi skupinami di- vs polyploidní jedinci tkáně oocytů**; *similarita* mezi geny je vyjádřena formou dendrogramu. Bootstrap P hodnoty jsou vyjádřeny červeně, AU – *aprroximately unbiased* hodnoty zeleně. Červené obdélníky vyznačují signifikantní větvení; tučné červené linky, či tučné červené linky značí míru signifikance – spolu s uvedenými AU a BP hodnotami podpory větvení.

V grafu č. 40 je znázorněna exprese všech signifikantní DE genů vzhledem k typu rozmnožování v oocytech. Přibližně polovina anotovaných genů je nadexprimována v případě rodičovských druhů a *vice versa*. Opět zde nacházíme jedince, kteří mírně vybočují ve své expresi a mohou deformovat pohled na množství signifikantní DE genů. Bohužel edgeR je náchylný vůči *outlier* hodnotám, protože transformuje data vzhledem k trendu disperze, zatímco DESeq příspěvky jednotlivých vzorků navíc váží (Zhou et al. 2014). V případě této analýzy byl ale zvolen edgeR, protože DESeq naopak příliš často zamítá; na základě několika desítek – jednotek anotovaných genů nelze implikovat zcela žádné

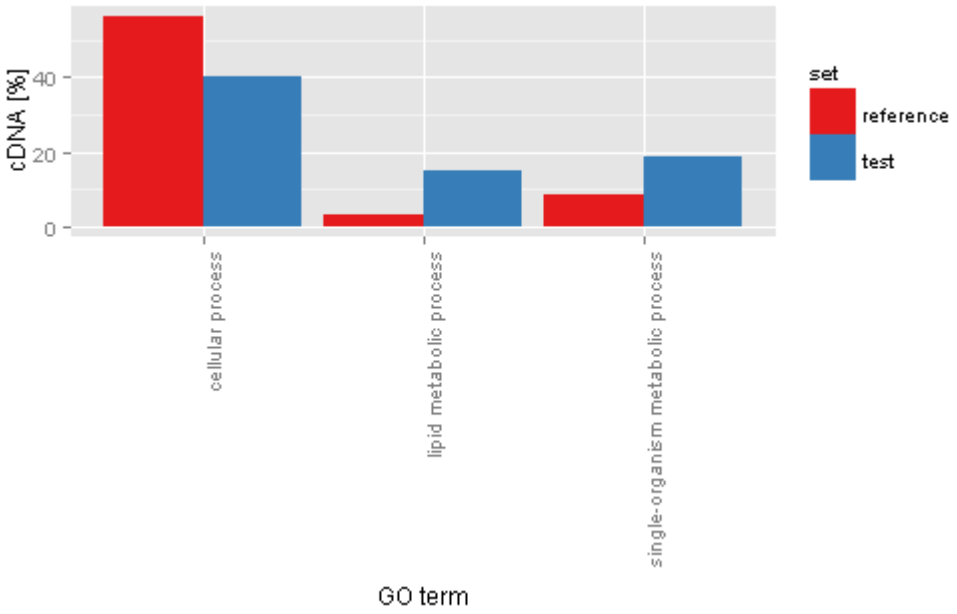
biologické konsekvence, musíme se ale smířit s opačným problémem. Čtyřicet genů s největšími změnami v expresi mezi skupinami (*fold change*), byly podrobeny shlukovací analýze hierarchického klastrování opět s distanční maticí vycházející z korelační matice a shlukovací metody využívající mediánu, viz graf č. 42, 43.



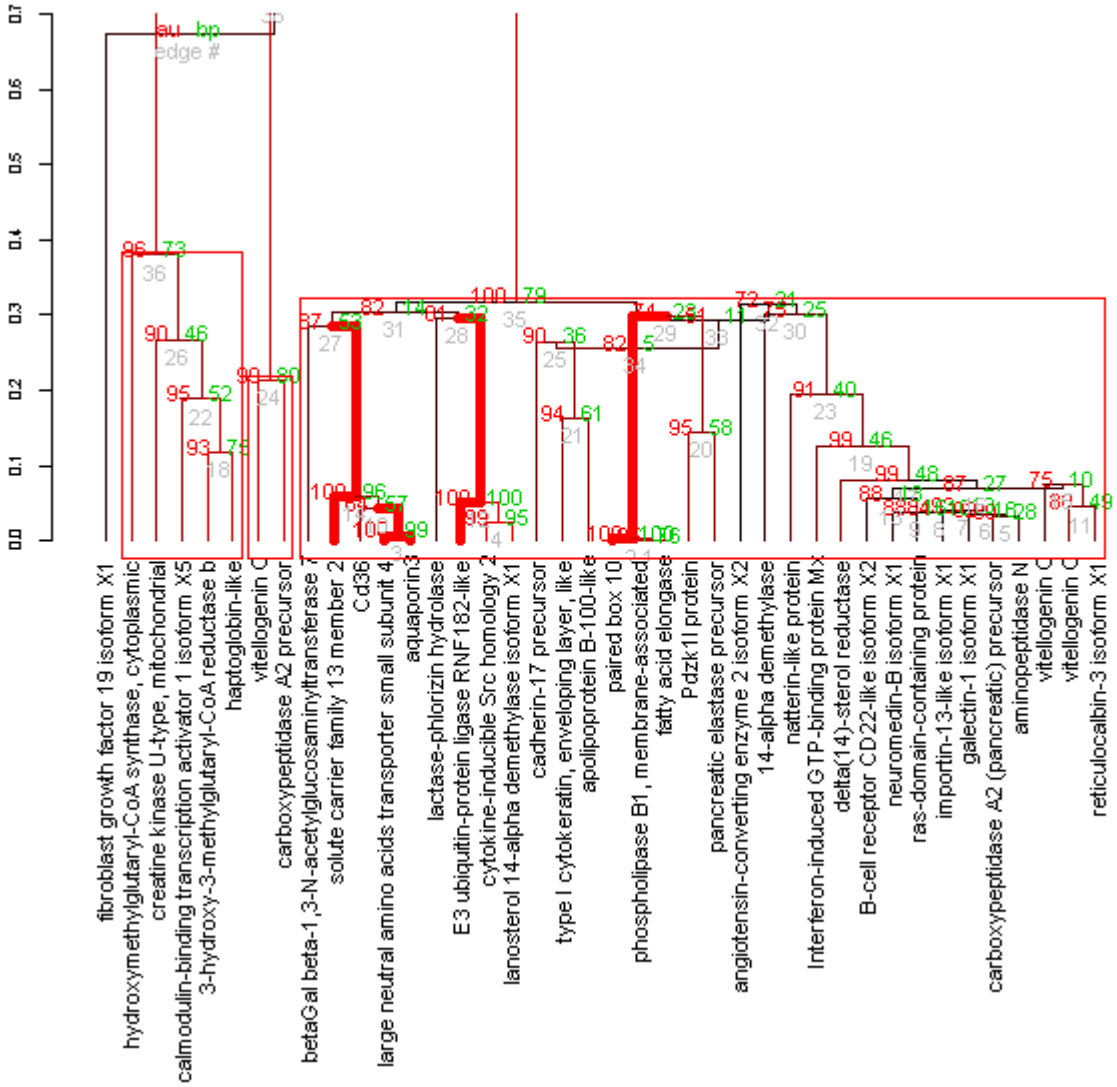
Graf č. 40: Heatplot – zobrazení míry exprese jednotlivých DE genů vzniknuvších komparací sexuálně a asexuálně se reprodukujících jedinců vzorků jater (s intenzitou modré barvy stoupá míra exprese – normalizovaná data, bílá - modrá). Dendogramy na horizontální ose vyjadřují výsledek shlukovací analýzy korelační koeficientů mezi geny, na vertikální ose vyjadřuje podobnost mezi vzorky.



Graf č. 37: Heatplot – zobrazení míry exprese jednotlivých DE genů vzniknuvších komparací **sexuálně** a **asexuálně** se reprodukujících jedinců vzorků **oocytů** (**s intenzitou modré barvy stoupá míra exprese – normalizovaná data, bílá - modrá**). Dendogramy na horizontální ose vyjadřují výsledek shlukovací analýzy korelační koeficientů mezi geny, na vertikální ose vyjadřuje podobnost mezi vzorky.



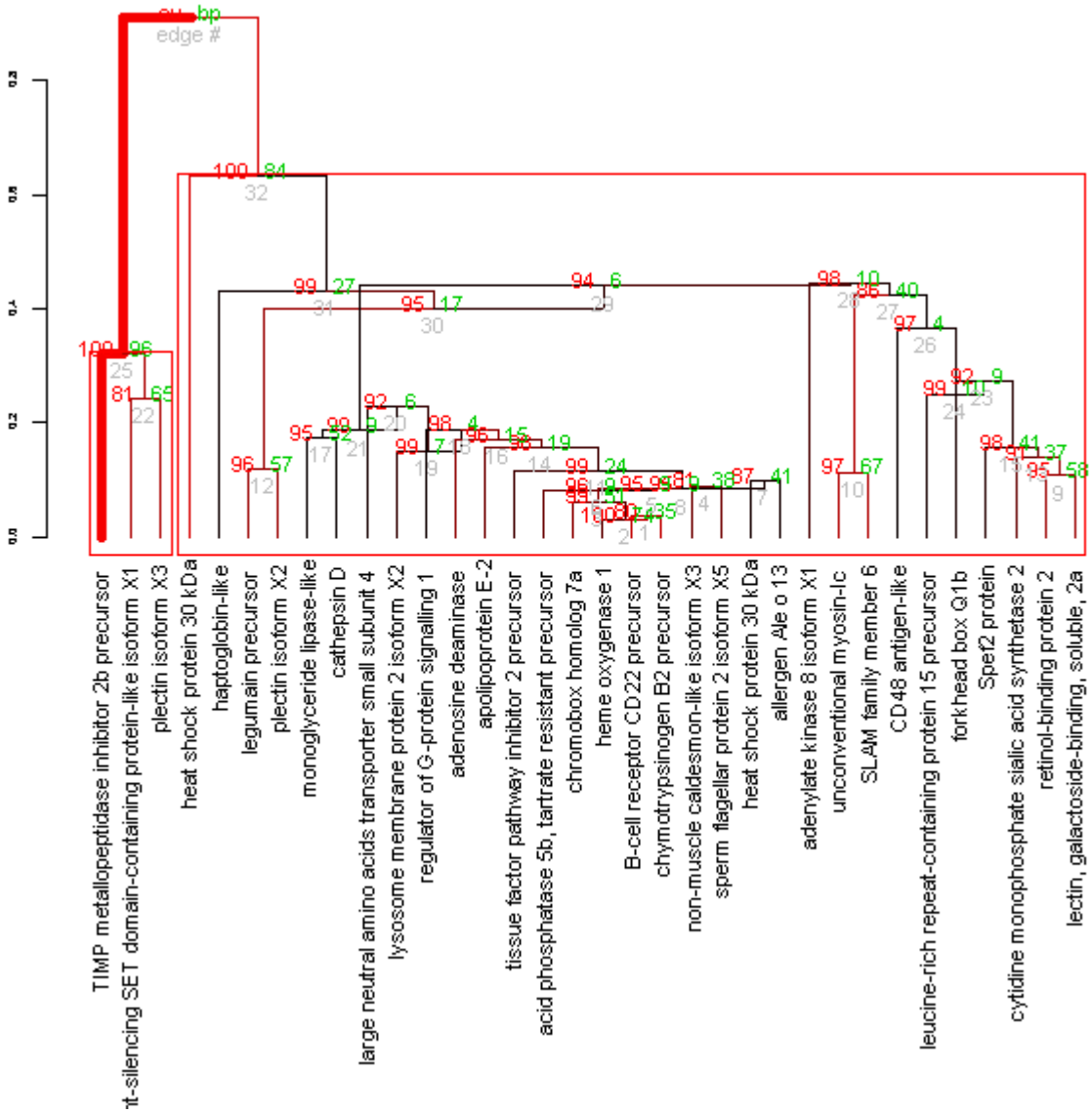
Graf č. 41: Explorace nabohacených GO term identifikátorů v setu DE genů vybraných z rozdílu množin grafu č. 27 – cílem je identifikovat pouze DE geny rozdílu **sexuálních a asexuálních jedinců** vzorků **jaterní tkáně**. Procento výskytu v referenčním setu anotovaných genů je znázorněn červeně, zatímco procento GO term v testovaném setu je znázorněno modře.



Graf č. 42: Hierarchické klastrování **40** DE geny s největší logFC ***subsetu* DE genů přítomných pouze mezi skupinami asexuálně – sexuálně reprodukujícími se jedinci tkáně jater**; *similarita* mezi geny je vyjádřena formou dendrogramu. Bootstrap P hodnoty jsou vyjádřeny červeně, AU – *aprroximately unbiased* hodnoty zeleně. Červené obdélníky vyznačují signifikantní větvení; tučné červené linky, či tučné červené linky značí míru signifikance – spolu s uvedenými AU a BP hodnotami podpory větvení.

U genů, kde bychom mohli navrhnout funkční koncept z hlediska spojitosti s gynogenetickým rozmnožováním, tj. rozdíl množin, odfiltrování interferujících kategorií polyploidie a druhově specifických rozdílů, zbývá 204 genů, ze kterých má přibližně 75 % anotaci, nelze implikovat na základě nabohacených GO funkci, buněčný kompartment ani metabolický proces. Naopak v tkáni jater (viz graf č. 41) lze pozorovat znatelný vliv na dráhu produkující vitelogenin, jež tvoří majoritní část žloutkového vaku (apolipoprotein – zejména transportní funkce, vazba na lipidy, N- doména signálál pro export). Bylo pozorováno, že hybridní jedinci produkují větší množství žloutku. Připomínám, že mezi analyzovanými jedinci byly pouze samice. V játrech byl také zaznamenán nárůst exprese

genů stojících za produkcí nukleotidů a replikačních kompent, což odpovídá představě, že hybridní jedinci jsou nuceni syntetizovat a reparovat o třetinu více DNA.

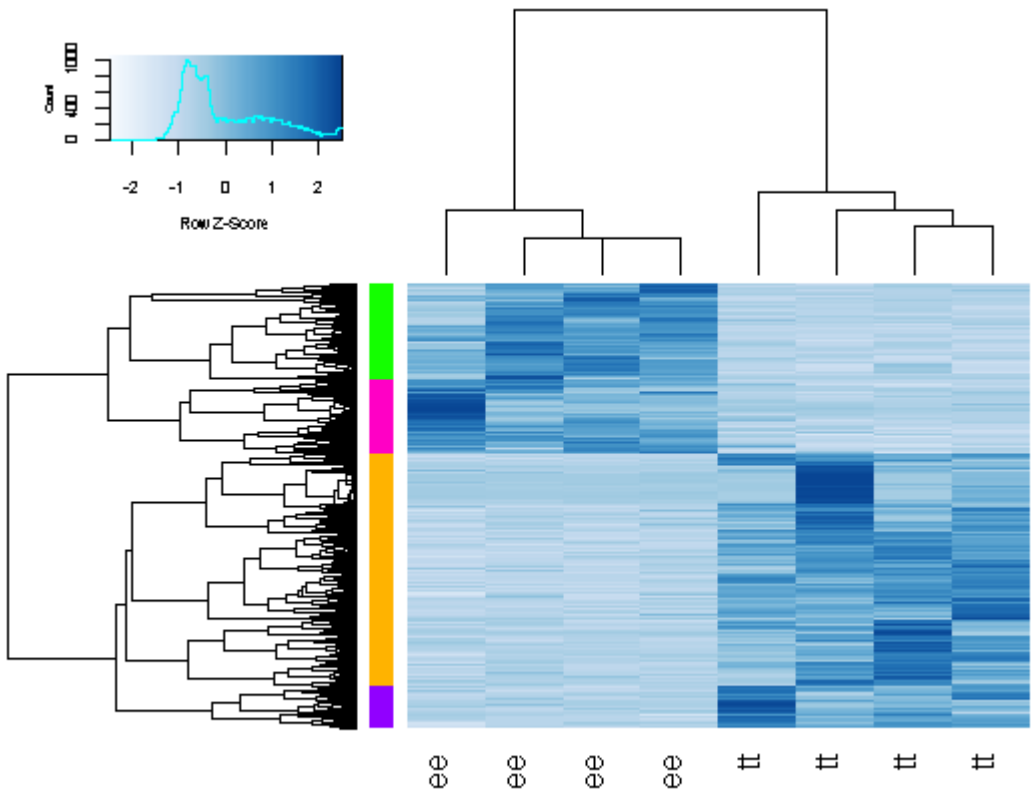


Graf č. 43: Hierarchické klastrování **40** DE geny s největší logFC ***subsetu* DE genů přítomných pouze mezi skupinami asexuálně – sexuálně reprodukujícími se jedinci tkáně oocytů**; *similarita* mezi geny je vyjádřena formou dendrogramu. Bootstrap P hodnoty jsou vyjádřeny červeně, AU – *aprroximately unbiased* hodnoty zeleně. Červené obdélníky vyznačují signifikantní větvení; tučné červené linky, či tučné červené linky značí míru signifikance – spolu s uvedenými AU a BP hodnotami podpory větvení.

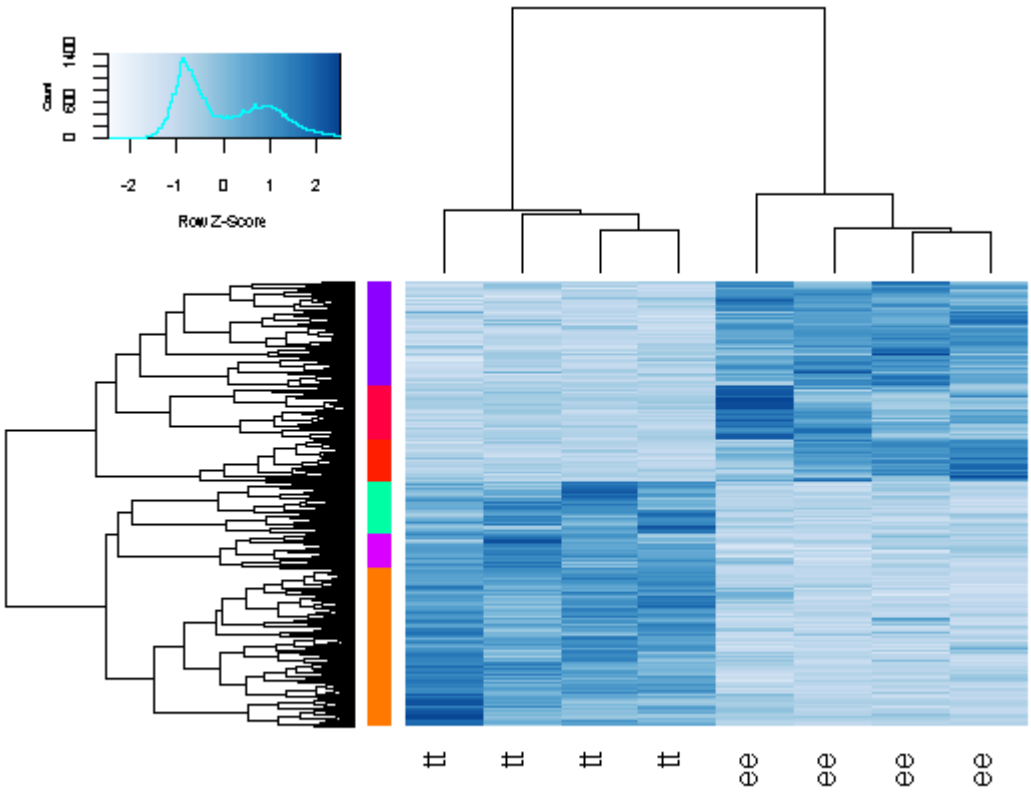
Poslední skupinou, z níž jsou prezentovány získané DE geny, je srovnání mezidruhové, konkrétně druhů *tt* a *ee*. Rozdíly genové exprese mezi druhy jsou nejmarkantnější, celkem bylo ve tkáni oocytů nalezeno 4462 DE geny, z čehož je 2452 *podexprimováno* u druhu *tt*. V tkáni jaterní je situace obdobná, bylo identifikováno 1855 *podexprimovaných* DE geny a 1141 *nadexprimovaných* u druhu opět ve srovnání skupin v pořadí *tt* vůči *ee*. *Heatmap* vyobrazení těchto skupin je uveden v grafech č. 44 a 45.

Klastrování úzké skupiny genů na základě své absolutní změny exprese v rámci celé množiny DE genů není uvedeno a to z důvodu nereprezentovatelnosti celkového počtu genů.

Z grafického vyjádření nabohacených funkčních skupin GO č. 46, 47 vyplývá, že tyto geny mohou být zapojeny především v metabolismu - metabolismus lipidů, oxidačně redukční mitochondriální procesy, translaci, transkripci (transkripční faktory), geny asociovány s jaderným prostorem. Tyto kategorie nacházíme jak ze srovnání oocytů, tak jater. Největší rozdíly byly nalezeny u genů pro cytochrom-oxidázy, ATPázy a vodíkových přenašečů mitochondrionálních krist. Dále je jedná o geny imunitní odpovědi genů jak viperin, tool like receptor a komplementy. Významné změny exprese jsou zaznamenány také u několika typů lektinů a v poslední řadě cyklinů A2, E1.

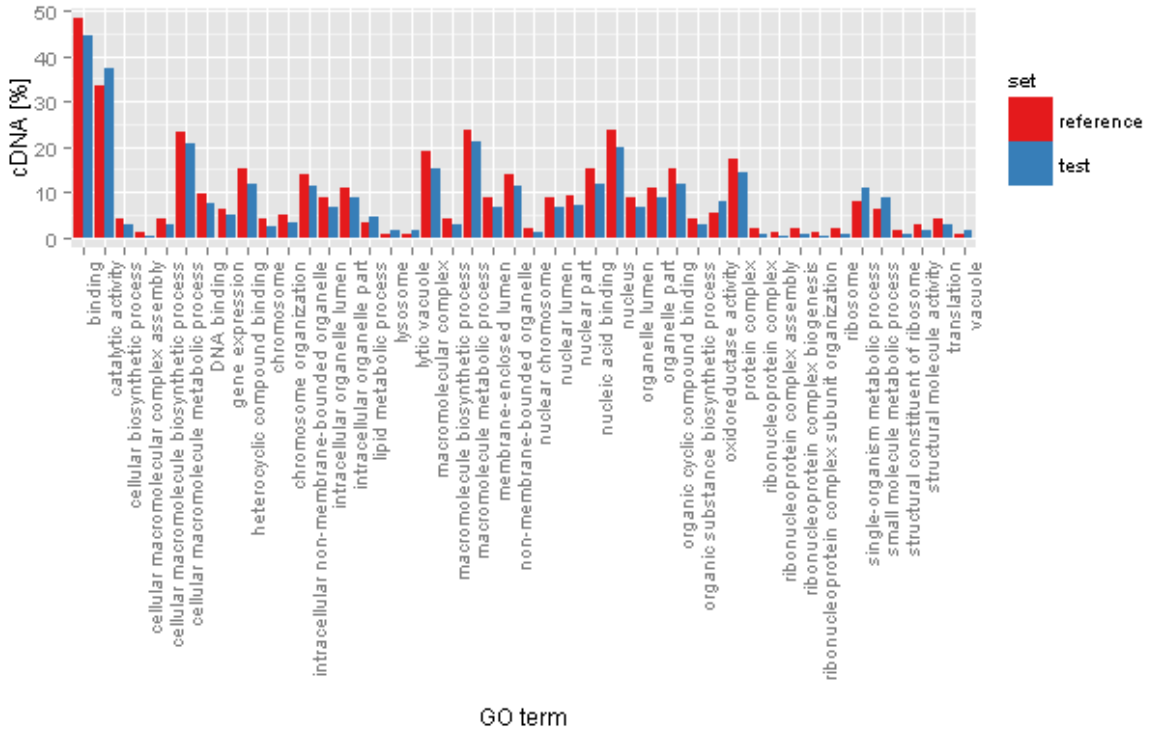


Graf č. 44: Heatplot – zobrazení míry exprese jednotlivých DE genů vzniknuvších komparací **druhů *tt* a *ee*** vzorků **jater (s intenzitou modré barvy stoupá míra exprese – normalizovaná data, bílá - modrá**). Dendogramy na horizontální ose vyjadřují výsledek shlukovací analýzy korelační koeficientů mezi geny, na vertikální ose vyjadřuje podobnost mezi vzorky.

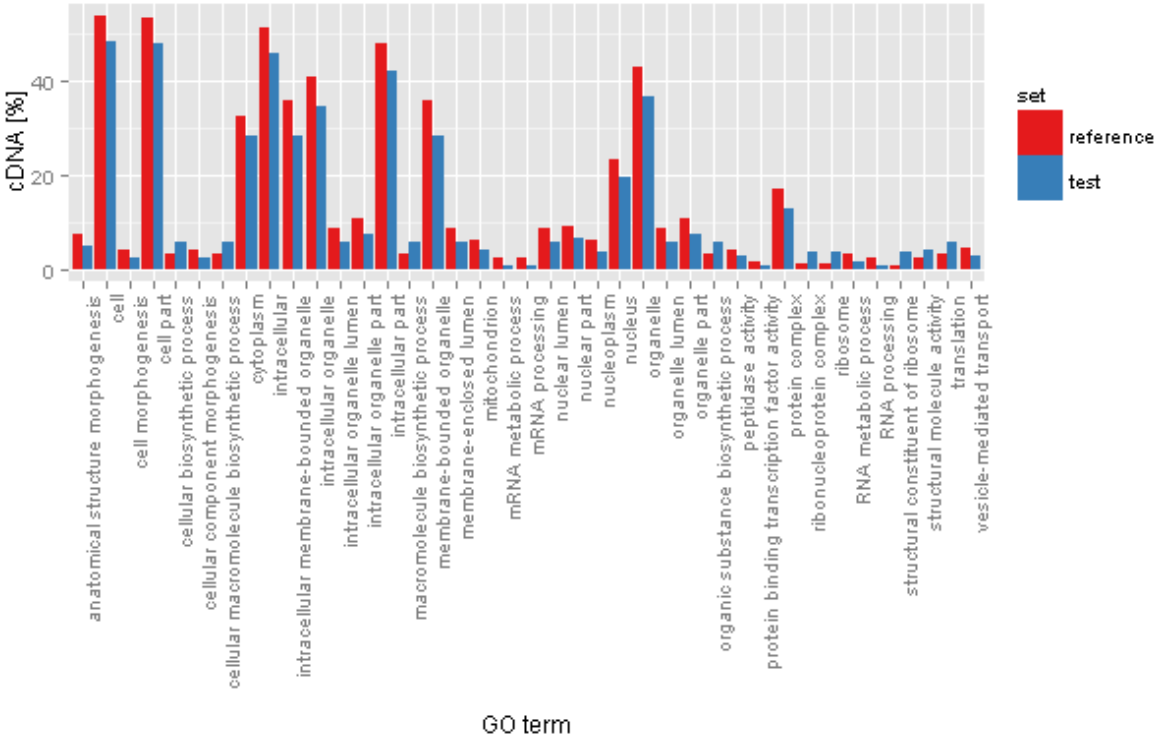


Graf č. 45: Heatplot – zobrazení míry exprese jednotlivých DE genů vzniknuvších komparací **druhů *tt* a *ee*** vzorků **oocytů (s intenzitou modré barvy stoupá míra exprese – normalizovaná data, bílá - modrá)**. Dendogramy na horizontální ose vyjadřují výsledek shlukovací analýzy korelační koeficientů mezi geny, na vertikální ose vyjadřuje podobnost mezi vzorky.

Sekundárním cílem analýzy DE genů je vyjádřit rozdíly mezi druhy *tt* a *ee* na úrovni tkáně jater a oocytů. A to nejen z důvodu identifikace těchto rozdílů z pohledu zaměření na geny hrající roli ve vývoji gynogenetického embrya, ale také pro komparaci výsledků mezidruhových rozdílů s recentní literaturou. Byly nalezeny rozdíly v metabolické aktivitě mezi druhy a potažmo i hybridy. Získané výsledky mohou přímo sloužit k vzájemnému srovnání dat.



Graf č. 46 : Explorace nabohacených GO term identifikátorů v setu DE genů získaných porovnáním druhů ***tt*** a ***ee* jater**. Procento výskytu v referenčním setu anotovaných genů je znázorněn červeně, zatímco procento GO term v testovaném setu je znázorněn modře

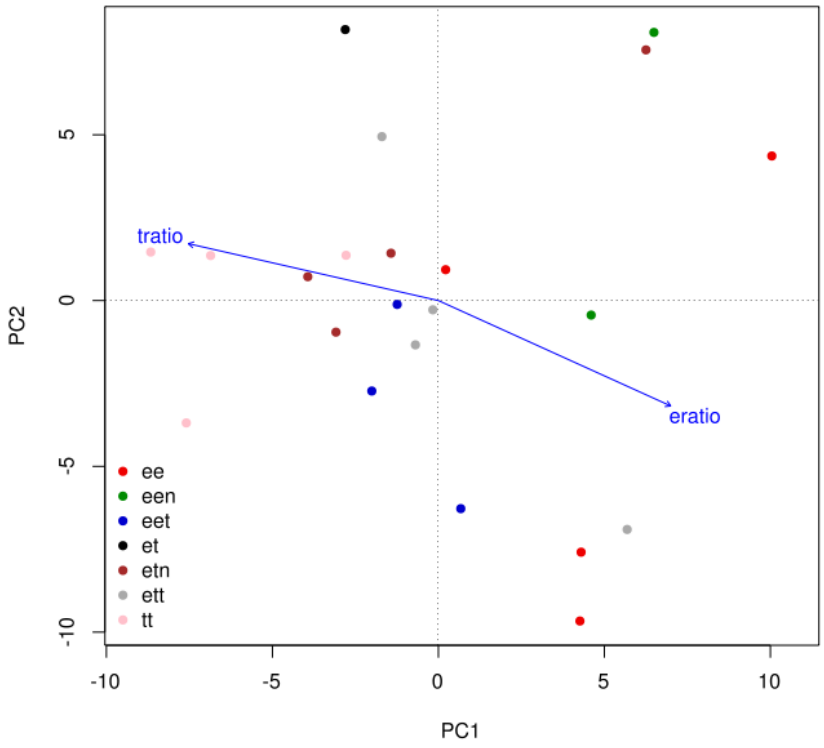


Graf č. 47: Explorace nabohacených GO term identifikátorů v setu DE genů získaných porovnáním druhů ***tt*** a ***ee* oocytů**. Procento výskytu v referenčním setu anotovaných genů je znázorněno červeně, zatímco procento GO term v testovaném setu je znázorněno modře.

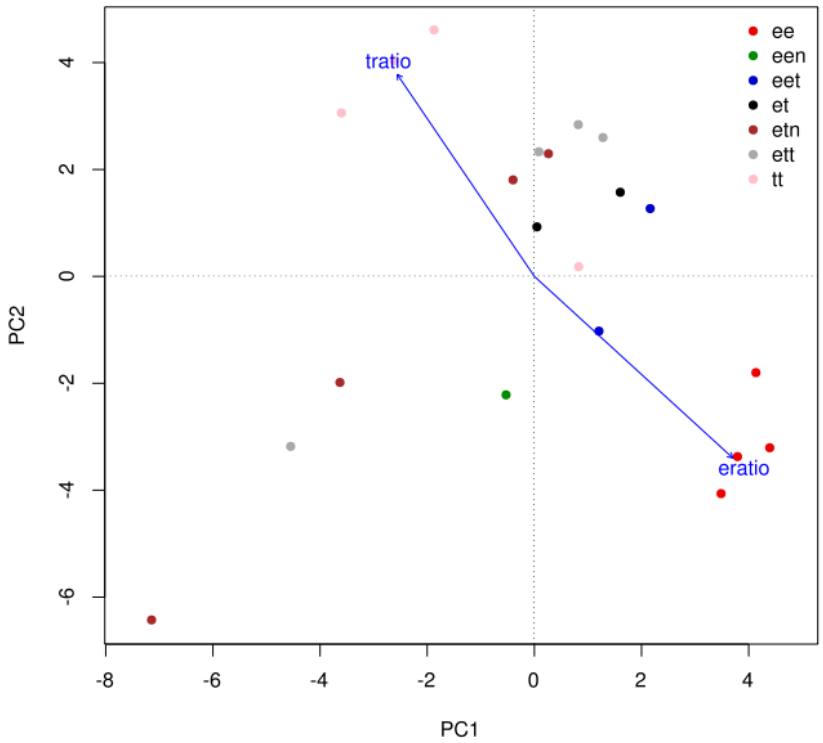
21.1 Imprinting hybridních genomů

Jak vyplývá z předchozích zjištění diferenciální exprese, na hybridní jedince nemá polyploidie vliv, jaký bychom očekávali, nenacházíme významný rozvrat genové exprese, ani postižení konkrétních signálních a metabolických drah. Vyvstává otázka, nakolik si hybridní jedinci, ať di- či polyploidní zachovávají úroveň exprese rodičovských druhů, zda se podobají spíše jednomu z rodičovského druhu, nebo zda dojde k intermediárnímu projevu globální genové exprese.

Pro testování těchto možných hypotéz navrhl Mgr. Karel Janko, Ph.D. a Mgr. Ladislav Pekárik, Ph.D *nafitování* vektorů každého vzorku na ordinaci rda (ordinální shluková analýza – redundační analýza) modelu gradientu *tt* a *ee*, provedeno 1000 permutací (přístup permutační annovy). V rda grafu č. 48, 49 je zobrazen každý analyzovaný vzorek na základě exprese všech genů. Koordináta bodů mezi komponentami vyjadřuje podobnost mezi vzorky. Enviromentální vektor exprese druhů *ee* a *tt nafitovaný* na data je znázorněn modrou čarou, jelikož některé vzorky obsahují navíc haplotyp *n*, nesvírá tato přímka mezi druhy úhel 180°. Tyto grafy vzorků oocytů a jater tedy říkají, jak si jsou kvantitativně podobní jednotliví hybridi definovaných genotypů s rodičovskými druhy *ee* a *tt*. Nfitování modelu směru *tt* i *ee* je signifikantní v obou případech na α 0.001.

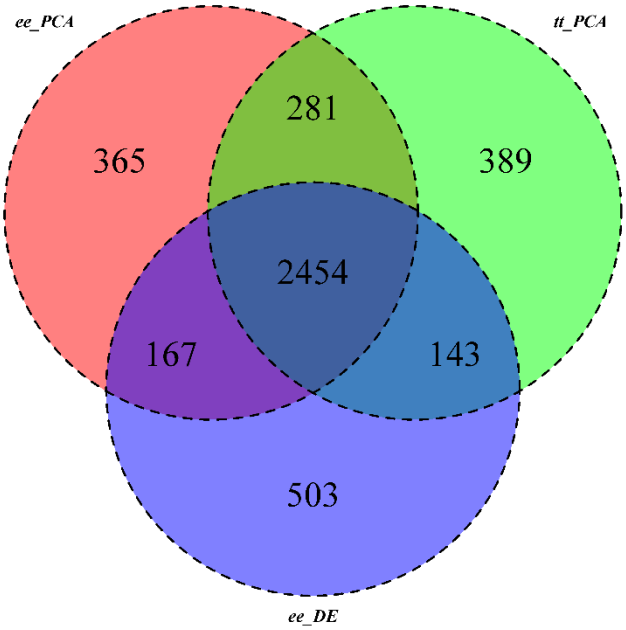


Graf č. 48: *Nafitovaný* environmentální model podle druhů *tt* a *ee* (tratio, eratio) na rda shlukovou analýzu vzorků jater. PCA1 vysvětluje 21.13% variability, PCA2 13.88% (Mrg. Ladislav Pekárik, Ph.D.,2015).



Graf č. 49: *Nafitovaný* environmentální model podle druhů *tt* a *ee* (tratio, eratio) na rda shlukovou analýzu vzorků oocytů. PCA1 vysvětuje 27.36% variability, PCA2 15.29% (Mrg. Ladislav Pekárik, Ph.D.,2015).

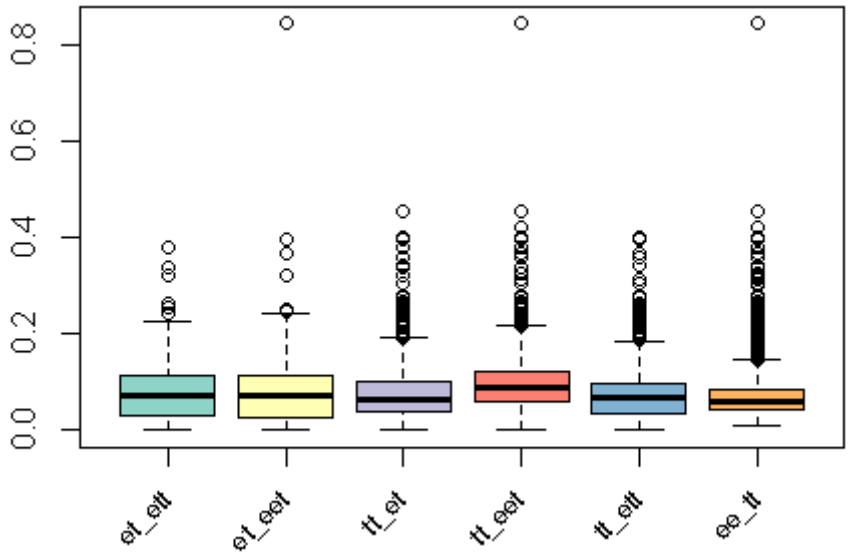
Trigonometrickým přepočtem lze ze získaných expresních výsledků získat též příspěvek jednotlivých genů vzhledem k *nafitovanému* gradientu a to jak *ee*, tak i *tt*. Geny byly seřazeny podle nové hodnoty os a seřazeny podle svého významu vůči ose gradientu (absolutní hodnota "přerotované" koordináty genu). Význam tkví především v tom, že působí jako další vnitřní kontrola diferenciální exprese. DE geny jsou expresně taktéž vychýleny, můžeme tedy srovnat průnik množin mezi geny vychylující osu gradientu *tt*, nebo *ee* a DE geny mezi vzorky druhů *tt* a *ee*. Z grafu Vennova diagramu č. 49 vyplývá, že DE geny mezi definovanými skupinami druhů přispívají taktéž k diferenci mezi druhy osy gradientů.



Graf č. 49: Vennův diagram tři množin: geny přispívající ke gradientu osy *ee*, geny přispívající ke gradientu osy *tt* a množina všech nalezených DE genů mezi druhy *tt* a *ee*.

Z grafu č. 49 lze vypozorovat, že většina genů DE genů zjištěných mezi skupinami druhů *tt* a *ee* náleží do sjednocení množin s geny vysvětlující směr, příspěvek k environmentálně *nafirovanému* gradientu na rda expresní data.

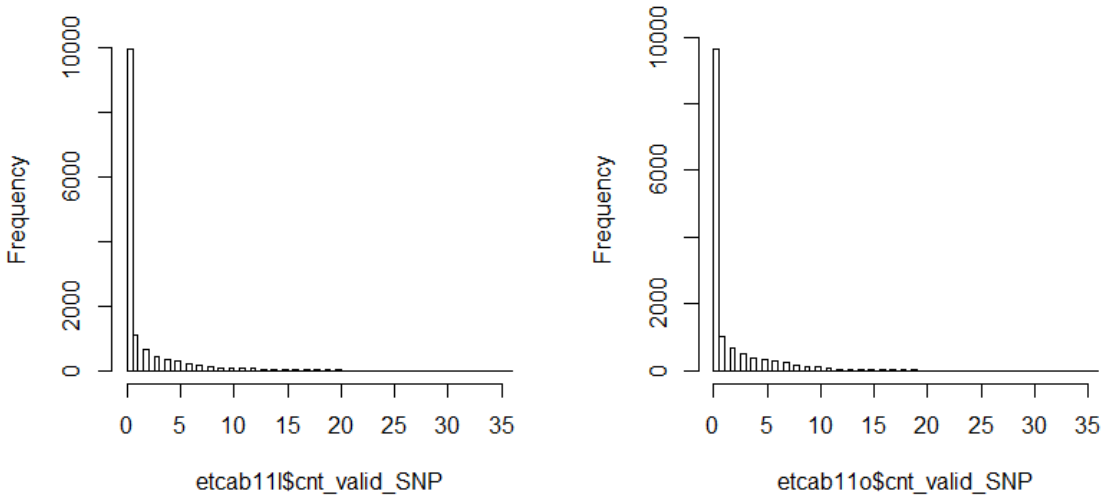
Další statistickou podporou intermediárnosti hybridů z hlediska celkové úrovně transkripce je srovnání distribucí nových, rotovaných kooridinát genů získaných spojením informace DE genů mezi vybranými srovnáními hybridů a rodičovských druhů vycházejících z grafu č. 35. Jak bylo naznačeno, počet DE genů vzrůstá s nárůstem rozdílných haplotypů v genomu hybrida. V grafu č. 50, 51 srovnáváme rotované hodnoty koordinát DE genů, které jsou výsledkem srovnání *tt et*, *tt ett*, *tt eet*, *et ett*, *tt eet*, totéž srovnání bylo provedeno naopak vůči druhu *tt*. Očekávaným výsledkem je výskyt DE genů nejvzdálenějších haplotypů zásadněji se vychylujících environmentální gradient *ee*, *tt*. Očekávali bychom zde posun k vyšším hodnotám u skupin haplotypově nejvzdálenějších, tedy především mezi druhy.



Graf č. 50: Vyjádření distribucí absolutních hodnot rotovaných kooridnát vůči *nafitovanému* gradientu *ee* DE genů vyobrazených skupin

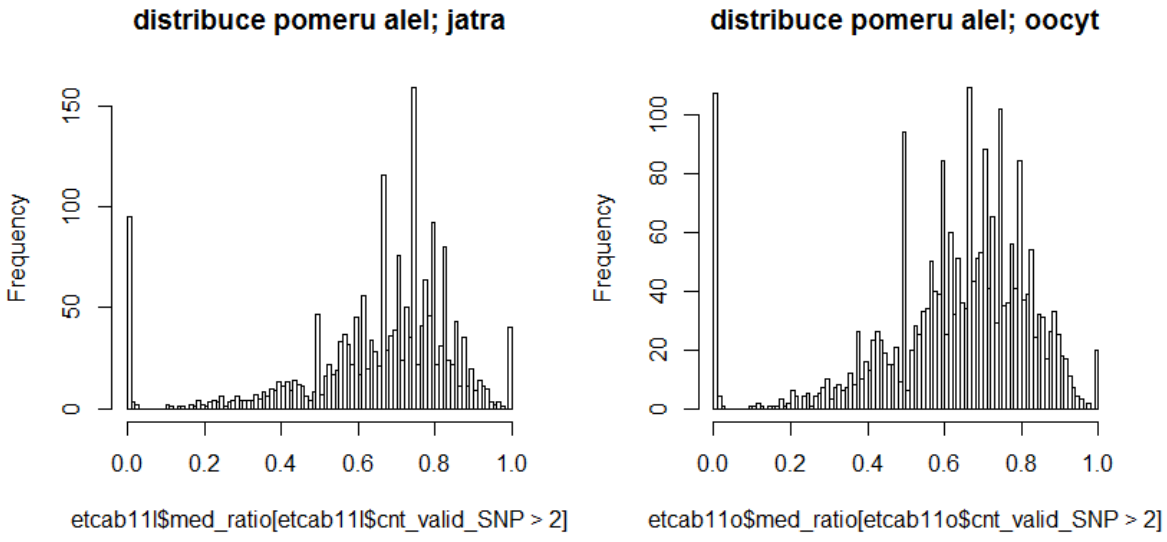
V následující část výsledků se bude upínat k otázce alelové exprese u hybridních jedinců, čili je jeden z genomů dominantní ve smyslu regulace exprese, tj. je jeden z genomů hybrida transkripčně umlčen, a pokud takové geny existují, jaký je jejich biologický význam.

Pro další pochopení výsledků je důležité zmínit, že SNP vhodné k testování alelické disbalance neslo jen jisté procento genů, které se navíc lišilo mezi jedinci v závislosti na typu použité tkáně, úspěšnosti sekvenování, či míře exprese daného genu (pochopitelně v málo exprimovaných genech jsme nemohli úspěšně detekovat žádné SNP, protože počet *readů*, z nichž by se dal rekonstruovat stav daného vzorku, byl příliš malý). Navíc, samozřejmě naše schopnost testovat alelickou disbalanci závisí na mutačních rozdílech mezi geny, pokud v daném loci žádný diagnostický SNP nebyl nalezen, nebyli jsme schopni tento test provézt pro daný gen. Jak ukazuje obrázek



Graf č. 51: histogram rozložení testovatelných SNP mezi studovanými geny na příkladu diploidního hybrida (cab11 jaterní, resp. oocytární tkáň)

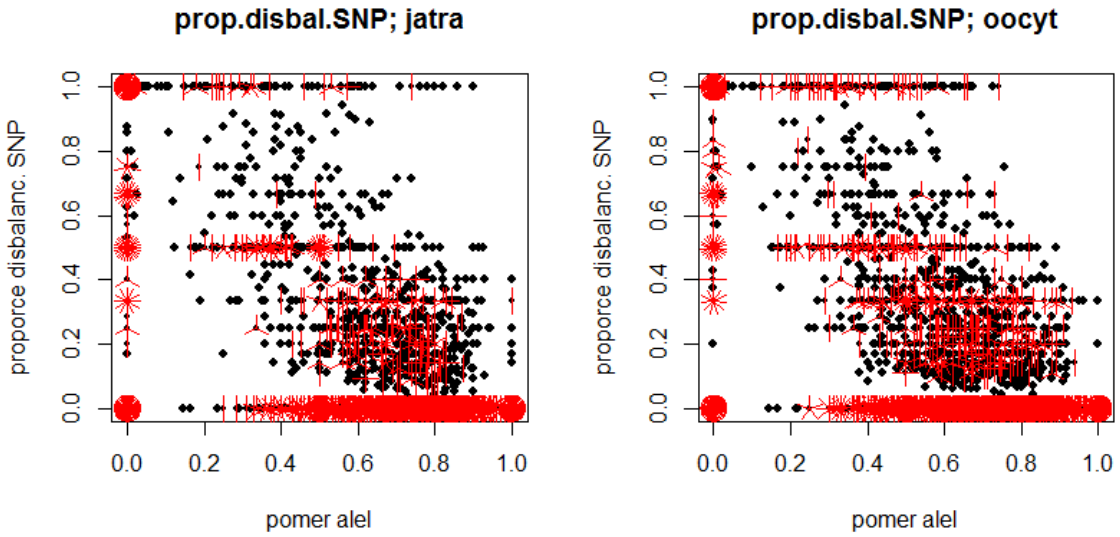
Množství detekovatelných SNP pozic v genech exponenciálně klesá a například 3 a více detekovatelných SNP pozic mělo jen cca 2000 genů (byly rozdíly v počtech takto způsobilých genů mezi vzorky, ale celkově se jejich počty pohybovaly víceméně mezi 1500 a 2800 geny s třemi a více detekovatelnými SNP). Jelikož tedy veškerá naše vypovídací schopnost závisí na počtu detekovatelných SNP, omezili jsme další studium alelické exprese jen na ty geny, které měly více než 2 detekovatelné SNP. Graf č. 52



Graf č. 52: rozložení mediánu poměrné exprese obou alelických variant v rámci detekovatelných SNP mezi geny u diploidního hybrida. Hodnota 1 znamená, že obě varianty byly stejně exprimovány, 0 znamená, že jedna varianta téměř neexistovala a lokus se jevil jako homozygotní.

ukazuje, že mezi geny existovala významná variabilita co do vzájemného poměru exprese obou alelických variant. Řada genů měla poměr vyvážený, ale existoval nezanedbatelný počet genů s očividně silně disbalancovanou expresí, kdy jeden genom byl téměř unikátně exprimován.

V rámci každého vzorku jsme pomocí chi kvadrát testu a následné sekvenční FDR korekce určili ty konkrétní SNP, které se na hladině celkové 5%ní pravděpodobnosti odchylovaly od očekávané balance buď ve prospěch genomu *C. elongatoides* nebo *C. taenia*. Graf č. 53



Graf č. 53: grafy znázorňují, jak souvisí poměr disbalancovaných SNP s celkovou úrovní poměrné exprese jednotlivých alel vyjádřenou mediánem všech detekovatelných SNP. Na příkladu diploidního hybrida

se ukazuje, že poměr signifikantně disbalancovaných SNP k celkovému počtu detekovatelných SNP v daném genu souvisí s celkovou mírou disproporce mezi alelami. Toto je triviální zjištění, ale za povšimnutí stojí fakt, že i v případě extrémně disbalancovaných genů (medián poměru alel roven nule, tedy prakticky homozygotní stav) jsme měli řadu genů, u nichž jsme nedetekovali žádný signifikantně disbalancovaný SNP. Tento paradoxní stav nastal tím, že naše schopnost určit SNP jako signifikantně disbalancovaný samozřejmě přímo souvisí i s pokrytím daného místa sekvenačními čteními.

Výše zmíněný popis výsledků byl důležitý proto, abychom si uvědomili limitace RNAseq v analýze alelické exprese. V následujícím textu se zaměřím na výsledky samotné detekce signifikantně disbalancovaných genů. Připomínám, že nadále budu u každého vzorku pracovat jen s těmi geny, které měly 3 a více detekovatelných SNP. Za

signifikantně disbalancovaný gen pak budu považovat pouze ten, který nesl alespoň jeden SNP u nějž byla pomocí chi kvadrát testu zamítnuta nulová hypotéza po provedení sekvenční FDR korekce. Naše analýza ukázala, že ze všech analyzovatelných genů jich měly jednotlivé vzorky signifikantně disbalancovaných mezi 30% a 55% (a to v závislosti na daném zvířeti a typu tkáně). U většiny genů jsem byl schopen jednoznačně určit, kterým směrem je jejich exprese vychýlena. V takových případech měly všechny signifikantně vychýlené SNP jasnou afinitu buď k jednomu, nebo ke druhému rodičovskému druhu. Avšak u nezanedbatelného procenta signifikantně vychýlených genů jsem našel konfliktní situaci, kdy alespoň jeden SNP byl v konfliktu se zbylými vychýlenými. Takovýchto genů, které dále nazývám problematickými, bylo u jednotlivých vzorků mezi 15% a 22%. Situace u nich zpravidla vypadal tak, že většina SNP ukazovala na disbalanci k jednomu druhu, zatímco jeden SNP byl v konfliktu. Bohužel jsem také nalezl nemálo genů, kdy poměr konfliktních SNP byl vyvážený.

Série následujících obrázků ukazuje na průniku mezi pěti vzorky jater, jak se jednotlivá zvířata shodovala, či lišila co do nevyváženosti alelické exprese genů. Aby bylo možno nějakým způsobem průniky ukázat, zvolil jsem 5 reprezentativních zvířat v jaterní tkáni, na nichž výsledky demonstruji. Vyšší počet zároveň zobrazovaných vzorků by již byl velmi nepřehledný. Zvolených 5 zvířat obsahuje 2 diploidní hybridy (cab11 a cab23), jednotho triploida genové kompozice *ett* (cab16) a dva triploidy genové kompozice *eet* (cab08 a cab09). Výběr také zahrnuje různé směry hybridizace, neboť maternálním předkem zvířete cab09 byl *ee*, zatímco maternálním předkem ostatních zvířat byl *tt*. Toto nám teoreticky umožňuje studovat i vliv atenuace exprese podle maternálních/paternálních předků. Veškerá následující srovnání pěti vybraných zvířat jsem provedl na podmnožině genů, které splňovaly tu podmínku, že **u všech zvířat zároveň** měly 3 a více detekovatelných SNP pozic (celkem 1007 reprezentativních genů).

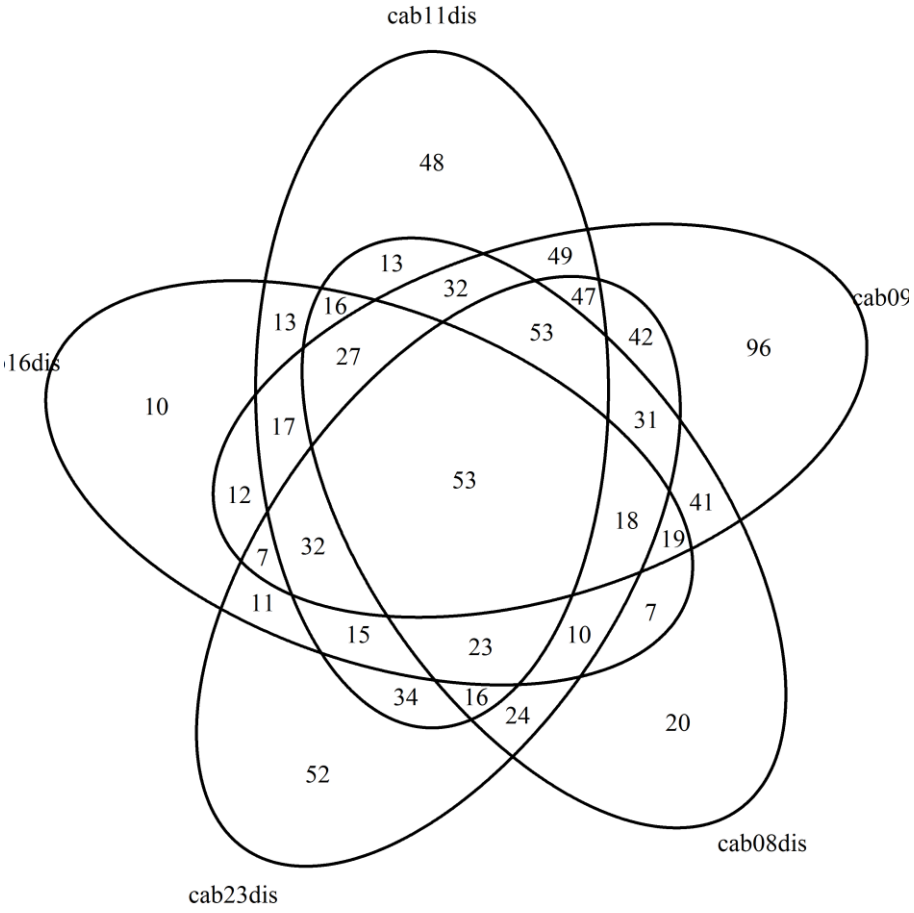
Z následující tabulky č. 8 je patrné, kolik SNP bylo detekováno jako průkazně vychýlených a kterým směrem u jednotlivých pěti vzorků.

vzorek	biotyp	Počet cDNA s validními SNP	ASE směr *ee*	ASE směr *tt*	problem. cDNA
cab11L	*et*	1007	124	364	100
cab23L	*et*	1007	139	329	119
cab08L	*eet*	1007	330	73	82
cab09L	*eet*	1007	525	51	126
cab16L	*ett*	1007	103	187	49

Tab. č. 8: Srovnání počtu všech analyzovatelných SNP, validních SNP a alelově specifických genů exprimovaných buď genomem *ee* nebo *tt*. Případy, kdy došlo ke konflitku disbalance v rámci genů, byly označeny jako problematické.

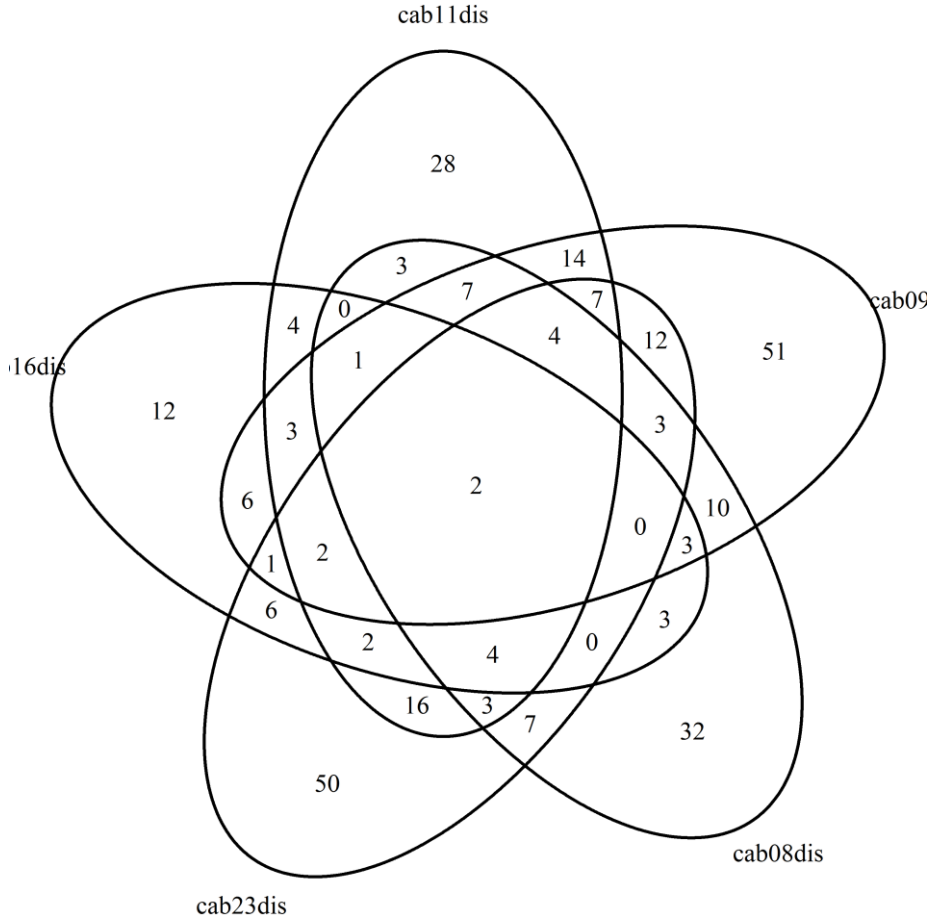
Z tabulky je patrné, že například diploidní hybridi a triploid typu *ett* měli tendenci k častější disbalanci ve prospěch genomu *tt*, zatímco oba triploidi typu *eet* častěji nadexprimovali genomy druhu *ee*.

Ukázalo se, že řada genů, u nichž byla zjištěna disbalance, je sdílená mezi vzorky, ale nemálo genů také jevilo disbalanci jen u jednoho, či několika málo vzorků, zatímco u jiných disbalance alel potvrzena nebyla, jak patrno z obrázku



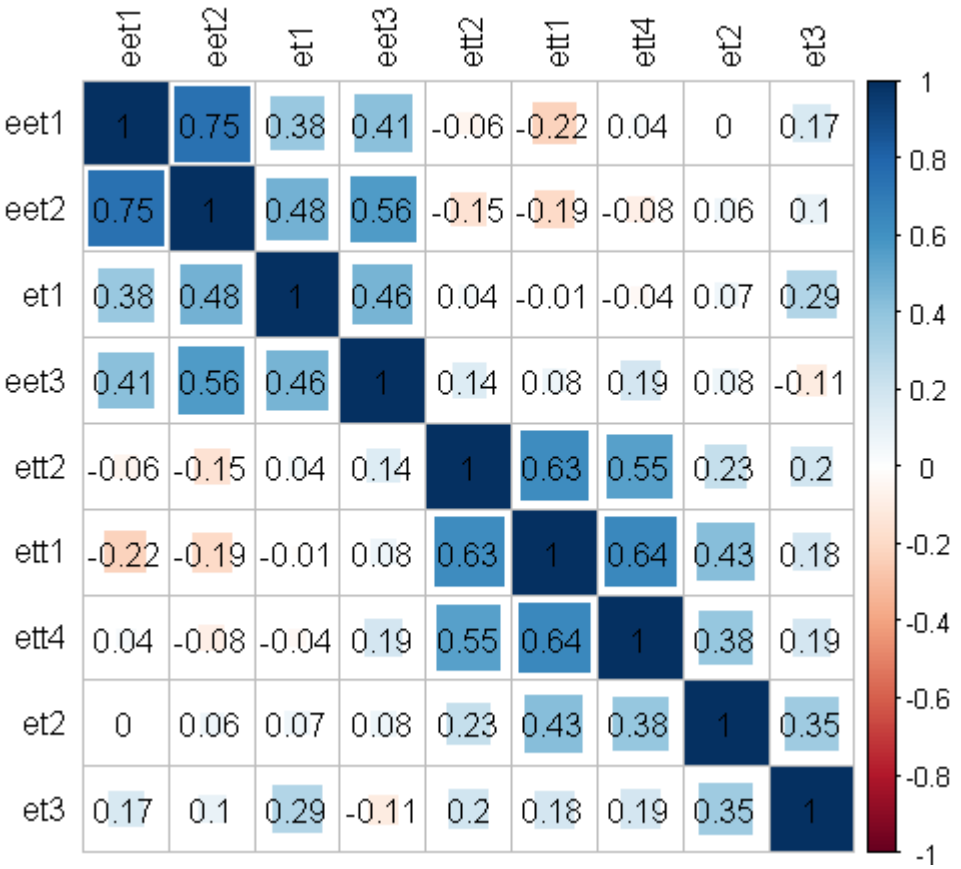
Graf č. 54: Vennův diagram znázorňující počty genů u nichž byla zjištěna disbalance a které byly sdíleny mezi jednotlivými zvířaty.

Počty problémových genů dosahovaly u vzorků do 20% a jejich shodnost, či rozdílnost mezi jednotlivými zvířaty je znázorněna na obrázku

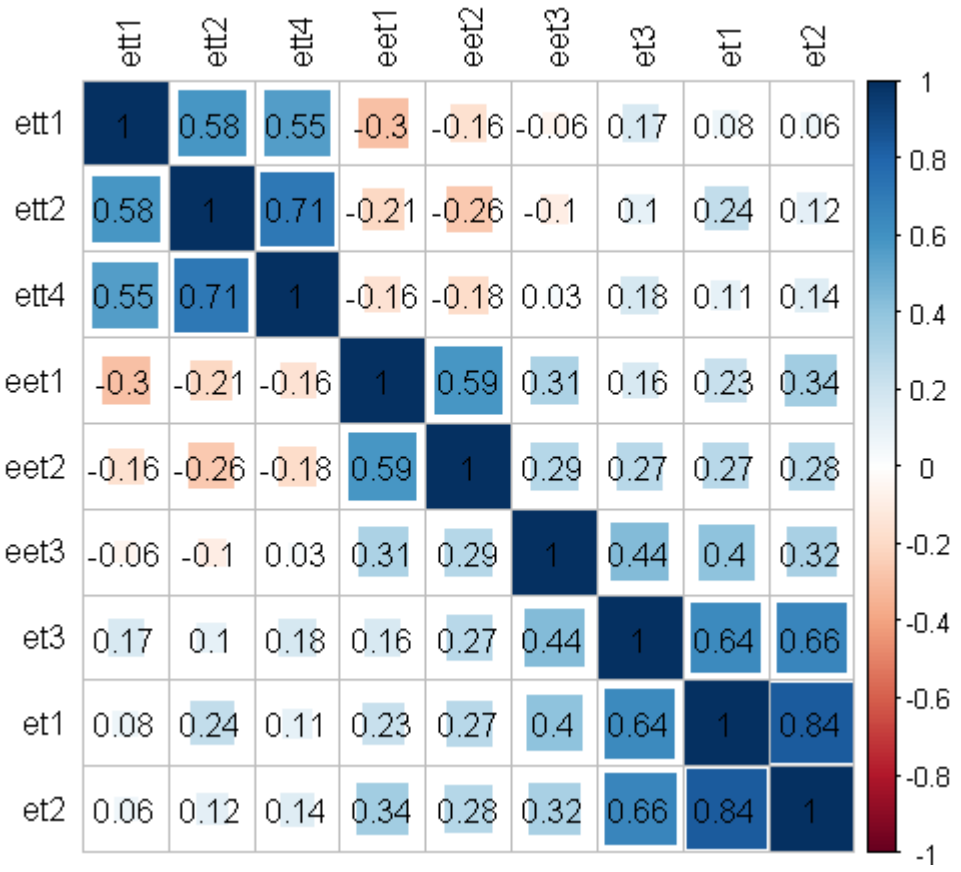


Graf č. 55: Vennův diagram znázorňující počty problematických genů, u nichž byla zjištěna disbalance oběma směry a současně ty, které byly sdíleny mezi jednotlivými zvířaty.

Shodnost mezi všemi vzroky na základě mediánu poměru alternativní a hlavní alely, ne disbalance, je znázozněna v grafu č. 56, 57 vyjádřena jako neparametrická spearmanova korelace všech genů, které obsahovaly alespoň 3 a více validních SNP v genu. Cílem je tedy srovnat podobnost "homo- či hetrozygotnosti" mezi jednotlivými typy hybridů. Toto vyjádření nám nic neříká o disbalanci alel ve prospěch jednoho rodičů, pouze podává informaci o podobnosti genů z hlediska exprese.

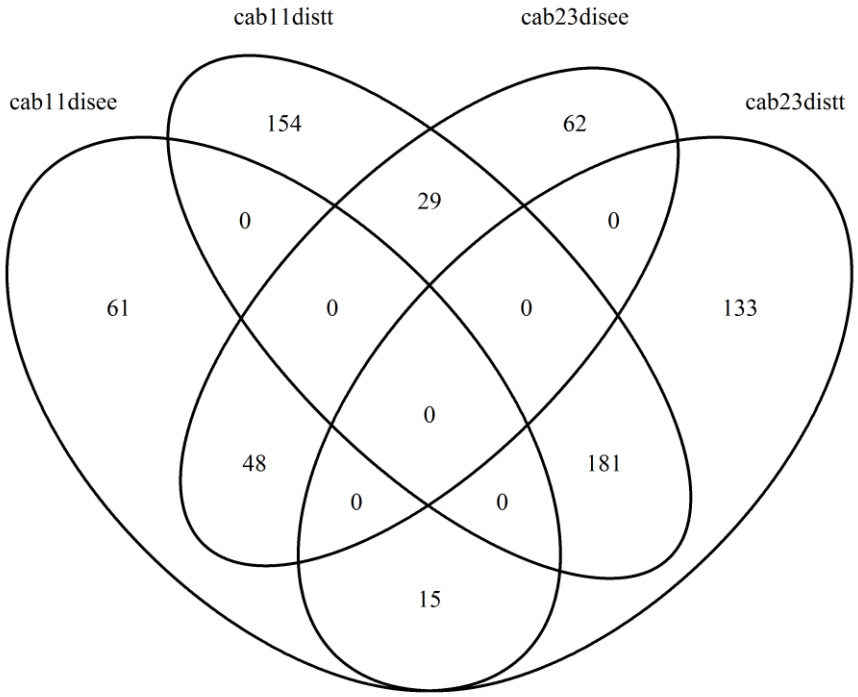


Graf č. 57: Korelační matrice mezi hodnotami mediánu všech alelově disbalancovaných genů pro jedince genotypu *et*, které splňují podmínku počtu validních SNP >=3 vzorků jater, korelační spearmanův koeficient je vyjádřen barevnou škálou.



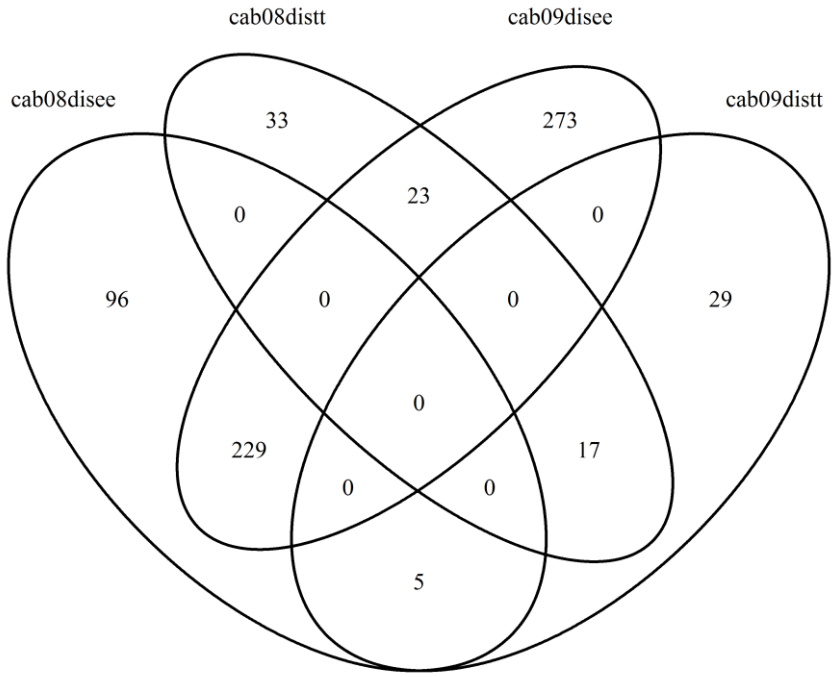
Graf č. 58: Korelační marice mezi hodnotami mediánu všech alelově disbalancovaných genů pro jedince genotypu *et*, které splňují podmínku počtu validních SNP >=3 vzorků oocytů (18 disbalancovaných genů), korelační spearmanův koeficient je vyjádřen barevnou škálou.

Způsob, jakým byl konzistentně, či nekonzistentně disbalancován daný gen, ukazují



Graf č. 59: Vennův diagram znázorňující počty genů u nichž byla zjištěna disbalance směrem k elongatoides (označeno jako ee) anebo taenia (označeno jako tt) a které byly sdíleny mezi dvěma diploidními hybridy

a



Graf č. 60: Vennův diagram znázorňující počty genů u nichž byla zjištěna disbalance směrem k elongatoides (označeno jako ee) anebo taenia (označeno jako tt) a které byly sdíleny mezi dvěma triploidními hybridygenomové konstituce ett.

Je patrno, že řada genů není mezi jedinci konzistentní: ačkoliv se jeví jako průkazně disbalancované u obou srovnávaných zvířat, tak v každém z nich jinak. Zajímavé však je, že takováto inkonzistence se týká genů nadexprimovaných ve prospěch *ee* u diploidních hybridů, zatímco u triploidů typu *ett* je to naopak (tam jsou inkonzistentní geny nadexprimované ve prospěch druhu *tt*). Opačný směr disbalance byl u obou skupin spíše konzistentní mezi vzorky. Připomínám, že u diploidních hybridů jsou častěji nadexprimovány geny ve prospěch *tt* a ty byly spíše konzistentní a u triploidů typu *eet* je to mu stejně v opačném gardu.

22.1 Degenerace hybridních linií - Müllerova rohatka

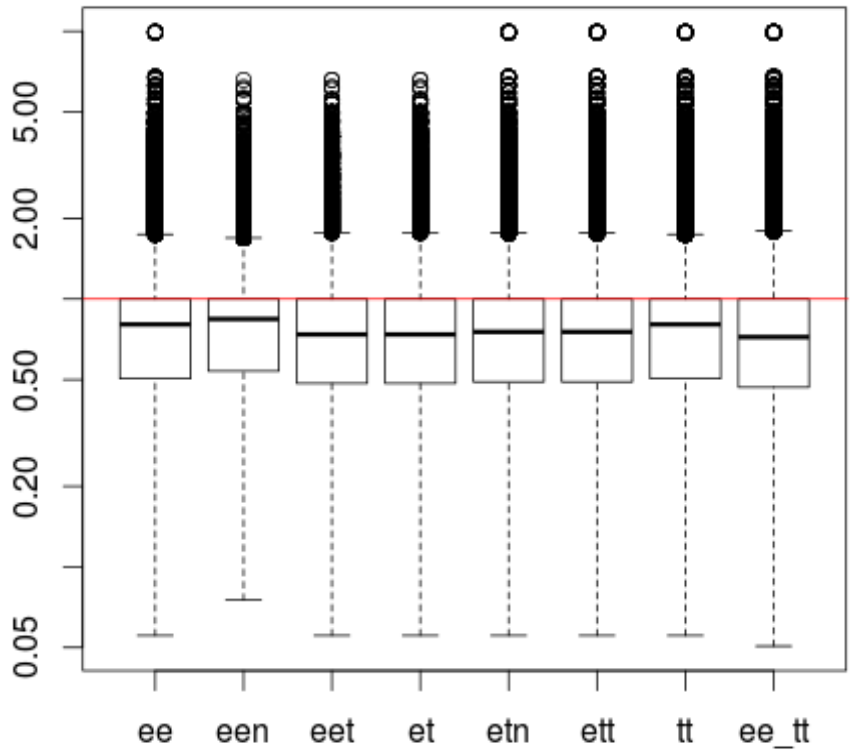
V poslední řadě budou prezentovány ústřední výsledky této práce – odpověď na otázku jak jsou ovlivněni hybridní jedinci z pohledu evolučního, jakým tempem dochází v genomu hybridů reprodukujících se pouze klonální cestou k akumulaci nesynonymních mutací (měření poměru nesynonymních / synonymních mutací z párové komparace), neboť

nedochází k rekombinaci mezi genomy rodičů a tudíž ani k náhodnému, efektivnějšímu odstraňování selekčně nevýhodných, nesynonymních mutací.

Z výsledků SNP RNA sekvenování je patrné vychýlení dN/dS poměru in silico vytvořených et hybridů směrem k nižším hodnotám. U párového testování rovnosti průměru pořadovým Wilcox testem zamítáme nulovou hypotézu o shodě průměrů s P hodnotou < 2.2e-16 (chi-kvadrát = 18328.63, df = 7), viz tab. č. 9; v případě užití neparametrické jednocestné anovy - Kruskal-Wallis pořadový test se P hodnota také blíží 0 (chi-kvadrát = 18328.6252, df = 7. Uměle vytvořený hybrid vykazuje nejnižší hodnotu mediáni dS/sN. Rodičovské druhy se mezi sebou signifikantně neliší takéž se srovnání vůči in silico vytvořeneému hybridov.

	ee	***een***	***eet***	***ee_tt***	***et***	***etn***	***ett***
een	<2e-16						
eet	<2e-16	<2e-16					
ee_tt	<2e-16	<2e-16	<2e-16				
et	<2e-16	<2e-16	1	<2e-16			
etn	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16		
ett	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	1	
tt	1	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16

Tab. č. 9: Párový Wilcoxnův test, neparametrické testování rovnosti průměrů



Graf č. 61: Distribude poměru dN/dS z párového porovnání vůči uměle vytvořenému hybridovi ee_tt.

Diskuse

Datapoužitá v analýzách jsou technicky různě zpracovaná. Používáme normalizovaná 454 sekvenovanou cDNA data pro tvorbu reference, nenormalizovanou cDNA sekvenovanou technologií ilumina a gDNA získaných hybridizací na navržené sondy ze 454 normalizovaných dat. Před samotným skládáním referenčního transkriptomu a mapování sekvencí byla kontrolována kvalita *readů*. Hlavním problémem referenční sekvence je nedostatečná hloubka sekvenování, která vyústila v produkci mnoha tisíc nekompletních cDNA – pouze UTR, či UTR s velmi krátkým úsekem kódující sekvence, které pak nebylo možno anotovat. Část problémů týkající se artificiálního spojení sekvencí na základě podobné sekvence DNA – chimérická DNA vzniklá procesem skládání cDNA, byla vyřešena díly anotaci (*alignmentů* blast *hitů*). V případě kde nemáme anotaci, nebo se blastx pozice *alignmentu* překrývají, nelze chimérické geny rozdělit.

Procesem *assembly* byly do některých genů, zejména oblastí s nízkým pokrytím a vysokou komplexitou, vneseny úseky nenáležící k cDNA, které konsekventně způsobily posun čtecího rámce. Vyřešeny byly též problémy s duplikovanými sekvencemi procesem *assembly* a především design experimentu. Na referenční sekvenci byly mapovány hybridní typy a různé druhy, aby měl každý druh stejnou šanci být *namapován*, byly do reference v místech jedno nukleotidových polymorfismů vloženy abiguidní pozice N. cDNA, ve kterých při opakovaném mapování vznikaly nové polymorfismy, byly odstraněny. V konečné fázi zůstaly tedy velmi spolehlivé referenční sekvence vhodné pro následné analýzy, nicméně díky nízké sekvenační hloubce 454 sekvenování s velkou spoustou nekompletních, neanotovaných sekvencí.

SNP byly získány v případě 454 standartním způsobem, byl užit software vhodný pro mapování cDNA na transkriptom, nevyřešené pozice ale musely být dodatečně upraveny. Data nenormalizovaných RNAseq sekenovaných technologií illumina byly zpracovaný SNP N pozic nestandardně, jelikož pro daný problém neexistuje nástroj, kterým by se daly SNP získat. Z toho důvody byly aplikovaný statisticky adekvátní nástroje pro detekci, které ale opomíjely část nevysvětlené variability ve vzorcích. Nejzadnější překážkou jsou naopak data genomické DNA. Postup mapování genomické DNA na transkriptom je stále ojedinělý a není softwarově vyřešen, proto byly SNP v pozicích kolem míst sestřihu – mezer v *alignmentu* odstraněny, abychom se zbavili falešně pozitivních výsledků.

Během analýzy expresních dat RNAseq bylo užito několik přístupů pro kontrolu kvality a konzistence dat pro maximální věrohodnost dat. Pro vlastní RNAseq expresní data byly navrženy primery pro vybrané *house-keeping* geny a deferenciálně exprimované geny pro stanovení korelace mezi relativní expresí zjištěnou qPCR a RNAseq; bohužel popsané srovnání není obsahem této práce.

Samotnou příčinu vzniku gynogenetického rozmnožování nejsme schopni z dat diferenciální exprese identifikovat a to jak bylo řečeno z triviálního důvodu; zaměřili jsme na analýzu rozdílů mezi oocyty 6. vývojového stádia nikoliv zárodečné tkáně vaječníků, sledujeme pouze výsledek tohoto děje, který podle všeho nastává již v době diferenciace primordiálních zárodečných buněk. Nicméně design naší studie umožňuje testovat genové projevy s asexualitou související. Proto je například zajímavé, že mezi DE geny skupiny sexuálních a asexuálních jedinců oocytů byly nalezen u skupiny asexuálních jedinců podexprimovaný gen pro rekombinázu 8 a nabohacený cyklin A2 a E1, protože předpokládáme, že mRNA by měla být po ukončení meiózy I odbourána.

Jak vyplývá z grafu č. 28, 26 překryv s geny skupiny di- vůči polyploidním vzorkům je značný, bohužel ani odfiltrování této skupiny není dostatečné. Například geny v referenci mají mnoho sestřihových variant, které v mohou v některých případech odpovídat realitě vyskytujíc se v obou skupinách – polyploidie a gynogeneze, po získání rozdílu množin odstraníme pouze část sestřihových variant (ne-li duplikované geny vzniklé procesem assembly cDNA), viz varianty cytoskeletrárních proteinů – nemůžeme s jistou určit, zda se jedná o následek polyploidního šoku nebo se jedná o funkční podmínku pro umožnění gynogenetického rozmnožování studovaného hybridního komplexu. V případě kdy navrátíme do *datasetu* vzorky *et* oocytů (*datasety* nejsou rovny do počtu identických jedinců, pouze různé tkáně), počty DE jedinců jsou zredukovány a jsme schopni s vyšší jistou navrhovat hypotézy týkající se funkčních podmínek gynogenetické reprodukce. Aplikací fisherova testu bez korigovaných P hodnot pro zjištění nabohacených GO termínů se výsledky shodují experimentu s hybridním komplexem karas-kapr (nepoužili taktéž korigované statistiky pro mnohonásobné testování) (Li et al. 2014). Na základě srovnání DE genů uvedené v tomto článku je s našimi daty pozorovatelná velká shoda, kterou bohužel nelze vyjádřit kvantitativně, jelikož autoři použili jiný softwarový přístup k anotaci genů. Jedná se o geny transkripčích faktorů, ubiqvitin ligáz, signálních genů ras

dráhy, genů spojených s extracelulární matrix, transferázy metabolismu lipidů a genů imunitní odpovědi.

Tato shoda mezi DE geny je velmi pozoruhodná vemeni v úvahu, že se jedná o zcela nezávislý vznik gynogenetických forem. Bylo již mnohokrát dokázáno, že s rostoucí divergencí druhů se též mění globální exprese morna (a to i v rámci rodu), přičemž většina diferenciální exprese nemá selekčně benefiční význam a ani nijak nepřispívá ke klonálnímu rozmnožování (Sartor 2006). Navíc je téměř nemožné predikovat význam sporadických mutací v místech transkripčních faktorů. Ačkoliv v místech sekvenčních motivů vysoce konzervovaných míst pro cis transkripční elementy lze připsat větší podíl na celkové změně genové exprese než cis-elementům s nižší konzervovaností (Tirosh et al. 2008). Na genovou expresi má také významný vliv tandemové genové duplikace genů, významné evoluční změny (evoluce genovou duplikací teorie Ohno 1970), navíc vedoucí k velkým expresím rozdílům (Blekhman et al. 2009). Hlavním faktorem, který je uváděn v souvislosti s ovlivněním genové exprese u hybridů, je označen samotný polyploidní stav vnášející *bias* mezi cis, trans promotory a regulačními motivy.

V případě srovnání výsledků s rostlinnými apomiktickými modely nacházíme shodu v nabohacení genů u klonálně se reprodukujících jedinců pro regulaci buněčného cyklu, cyklinů, metylačních kompexů, chromodomény a histonových variant (v našem případě H3.3 – podexprimován u hybridů) (Galla et al. 2015). Obzvláště výsledky DE genů pro epigenetického umlčení, restrukturalizaci chromatinu jsou velmi překvapivé, naopak byly u gynogenetických jedinců tkáně oocytů, tak jater nalezeny přibližně 6x více transkriptů pro retrotranspozon jockey. Čekali bychom spíše demetylační komplexy a ztrátu heterochromatinizovaných loci. Pozn.: Histon H3.3 je spojen s chromatinem transkripčeně aktivních genů.

Jedním z dalších fenoménů studovaných hybridních, klonálně se reprodukujících typů je deregulace epigenetického aparátu, především pak demetylace transponovatelných elementů. Především "mladé hybridní linie" zažívají doslova explozi aktivity transponovtelných elementů. Časem, vlivem selekce - výhodných mutací a genové konverze je jejich činnost utlumena. Aktivní retrotranspozony, retroelementy, mohou vést k zásadním a nečekaným změnám exprese, ať včlenění cDNA do nového regulační oblasti, či inzercí do funkčního genu, regulační oblasti vedoucí k jejímu vyřazení. Dávná aktivace retrotranspozonů byla nalezena i u starých linií rodu *rotifera*, spekuluje se, že aktivita transpozonů může v určitých stádiích vývoje hybridní linie vnášet diverzitu, kterou ztratili

spolu s opuštěním sexuální formy reprodukce (Mark Welch and Meselson 2000). Existují i důkazy, že inzercí retrotranspozonu do rekombinázy 8, hrající zásadní roli v rekombinaci sesterských alel v meióze, vznikly některé asexuální linie rodu *Daphnia* (Eads et al. 2012). Geneze aktivity transponovatelných elementů je spojena s hybridizací, polyploidizací. Transponovatelné elementy mohou být aktivovány stresem – různé environmentální faktory. Dráhy spojené s indukcí stresu a aktivací transponovatelných elementů se velmi podobají. Tento je byl nazván jako polyploidizační šok (Guerreiro 2014).

Bohužel tato práce nemůže k problematice transponovatelných elementů jako následek hybridizace příliš přispět, protože reference, ze které byla expresní data získána je složena pouze z druhu *tt*, kde nepředpokládáme exces transponovatelných elementů, nemůžeme je tedy ani "složit", ačkoliv pro testování hybridního máme data evolučně "starých" i "mladých" hybridní linie i hybridů P1 generace artificiálního křížení.

V oocytech klonálních hybridů byla detekována snížená hladina mRNA syntázy progesteronu a naopak zvýšená hladina pro substyp P450 odbourávající, metabolizující progesteron. Význam progesteronu tkví v maturaci oocytu, nicméně o vlivu na gynogenetickou reprodukci není příliš známo. Jediným organismem kde bylo spojení produkce progesteronu sledováno ve spojitosti s asexuální reprodukcí je známa u rodu *Rotifera*. Bylo zjištěno, že hladina progesteronu má významný, negativní vliv na růst asexuální linie. Progesteron aplikovaný v rané fázi embrya může indukovat zvrat pohlaví ve prospěch samců. Pro asexuálně rozmnožující samice je tedy selekčně velmi výhodné produkci progesteronu blokovat (Snell and DesRosiers 2008). Při fertilizaci embrya je progesteron nezbytný také pro rozpad karyolemy a přechod oocytu do druhé fáze meiotického cyklu. Nevyřešenou otázkou je nárůst mRNA SRY regionu u hybridních, klonálních jedinců. Recentně neexistují publikace, které by tento výsledek, nebo možný artefakt mohl vysvětlit.

Mezi geny, které mají logickou přímou spojitost s gynogenetickým rozmnožováním je glykoprotein zona pelucida bránicí vzniku polyspemického embrya. Některé části tohoto glykoproteinového komplexu byly podexprimovány, jiné přesně naopak. Otázkou je zda může spermie vniknout v případě gynogenetického rozmnožování do oocytu a přispět svou genetickou informací (vznik neživotaschopného zárodku), či je spermie vždy pouze aktivátorem pro rýhování neoplozené zygoty. Dochází-li k inkorporaci a zániku takových embryí, je přestavba zony pelucidy selekčně nevýhodná a vzniká následkem hybridizační transkripční disregulace, v opačném případě zona pelucia ztrácí svůj význam a nepodléhá

negativní selekci. Mezi DE geny mohouc objasnit procesy vedoucí ke gynogenezi jsou přestavby cytoskeletu. U hybridních jedinců byl detekovány velké změny hladin RNA několika typů aktinu, konexinu, fibronektinu, myosinu ad. spolu s Ran GTP proteiny "umožňující" tyto přestavby. Inhibitor polymerizace aktinu cytochalasin B může indukovat gynogenetickou aktivaci oocytu a to i u vyšších savců, ačkoliv embrya zanikají nejpozději ve stádiu blastocysty. Změny cytoskeletu mohou tedy přímo bránit vstupu spermie do oocytu (Lee et al. 2014a).

Aktivace embryonálního vývoje silně závisí na oscilaci vápenatých a draselných iontů v cytoplasmě po vniknutí spermie. Protein, který je zodpovědný za aktivaci *sperm-specific* fosfolipázy C – PLC zeta , není sice identifikován, zato je detailně popsán proces aktivace tohoto receptoru. PCL zeta je lokalizován neobvykle v cytoplasmě, kde interaguje s vezikuly PIP2 (sekundární přenašečem DAG) spouštějící elevaci Ca^{2+} z těchto vezikulů (Nomikos 2015), který následně spouští kortikální granulovou exocytózu, nutnou pro dokončení meiotického cyklu a první dělení zygoty. Detailněji, fosfolipázy katalyzují rozpad fosatidylionositol4,5bisfosfátu na IP3 (ionositol-1,4,5 trifosfátu) a diacylglycerol (DAG). DAG aktivuje proteinkinázu C a IP3 elevaci vápníku granul endoplasmatického retikula. Další cesty fertilizace zahrnují i aktivaci D1b (fosfolipáza D1b) a uvolnění fosatidové kyseliny (PA), která se váže na Src kinázu skrze SH3 a SH4 domény, která pak záhy aktivuje fosfolipázu C aktivující elevaci vápníku. mRNA Src kinázy a epidermální růstový faktor byly taktéž nabohacena ve vzorcích oocytů hybridních jedinců. Elevace vápníku způsobuje otevření chloridových kanálů – depolarizace membrány zabraňující vniku další spermie (Stith 2015). Nabohacená mRNA pro proteiny vápenatých a draselných kanálů ve vzorcích hybridních oocytů, včetně aktivačních drah fertilizace hybridů zůstává neobjasněnou otázkou, jež by si zasloužila hlubší analýzu.

Tímto bych rád přešel od diferenciální exprese k popisu exprese globální úrovně. Z grafů č. 48, 49 PCA s *nafitovanými* gradienty druhové exprese, vyplývá, že hybridní typy se podobají rodičovskému druhu v závislosti na složení genotypů, tzn., že počet haplotypů v genomu jednoho rodiče je přímo úměrný podobnosti s rodičovským druhem. Např. diploidní hybrid se jeví expresně průměrný, intermediární mezi rodičovskými druhy. Mgr. Ladislav Pekárik, Ph.D. studoval mimo jiné environmentální mikrohabitatové preference hybridů; opět se na gradientu genotypů *tt* a *ee* ukázalo, že hybridi okupují mimrohabitatové niky v závislosti na genotypu. Gradient závislosti genotypu je taktéž vysoce průkazný z

pohledu morfologické analýzy (Mgr. Miroslav Pertýl Ph.D., Mgr. Janek Kotusz, Ph.D., 2015, nepublikováno.

To, že hybridní jedinci se projevují expresně průměrně mezi rodičovskými druhy, také implikuje, že ani jeden rodičovských genomů není globálně inaktivován – heretabilně metylován, což bývá často nalézáno u hybridních komplexů jako řešení transkripční inkompatibility genomů (Wu et al. 2013). Nejzajímavějším zjištěním o imprintingu hybridních genomů je geneze nového unikátního vzoru imprintingu u stejných i různých typů hybridních jedinců nepřispívající k celkovému epigenetickému vzoru. Tyto menší rozdíly mezi hybridními liniemi se mohou stát zdrojem fenotypové varibilitity hybridních linií. Epigenetická modifikace jsou přičítány spíše hybridizaci nežli ploidizaci (Salmon and Ainouche 2010).

V této části prozkoumáme, zda existují imprintované geny v hybridních jedincích. Jakým způsobem změny korelují mezi jednotlivými typy hybridů a jaké funkční kategorie genů byly atenulací zasaženy. Naše studie odhalila stopy víceméně výrazné a velice rozsáhlé alel-specifické exprese mnoha genů. Než se pokusím naše data interpretovat, bude jistě vhodné diskutovat silné i limitacující stránky našeho přístupu. Za významný přínos považuji to, že aplikace NGS umožnila skutečně „large-scale“ studii téměř celého transkriptomu u většího počtu zvířat, což by normálně možné nebylo. Fakt, že dobře známe fylogenezi studovaných druhů, nám navíc umožnil kvalitní design studie s možností odhalení velkého počtu druhově specifických SNP. Za výraznou výhodu považuji to, že jsme naši studii provedli na referenci s již známým velkým počtem variabilních pozic. Tyto pozice byly konzistentně nahrazeny v referenci písmenem „N“, čímž jsme v podstatě omezili možnost, že *ready* z některých alel se budou mapovat na referenci efektivněji a tím artificiálně vnášet do dat disbalanci.

Potíže však také byly nezanedbatelné, mnohé vycházely ze samotné podstaty dat, jiné byly a v podstatě typické pro mnohé nemodelové organismy. Za prvé, kvalitní pročtení druhově diagnostických SNP jsme získali jen u relativně malého procenta studovaných genů (cca 20%). Tento fakt je dán hlavně tím, že mnohé geny měly nízkou coverage způsobenou jejich relativně nízkou mírou exprese. Vzhledem k tomu, že tytéž geny mohly býti dost různě exprimovány mezi zvířaty (viz předchozí kapitola), tak se ještě váce snížil počet genů, u nichž jsme mohli testy alel-specifické exprese porovnat mezi vzorky. Řešením by jistě mohlo být zvýšení sekvenačního úsilí, ale to by mělo jen lineární efekt a jelikož rozložení expresní intenzity mezi geny je zhruba lognormální, i při

několikanásobně vyšších finančních nákladech bychom si příliš nepolepšili. Další možností je tak jako u datasetu použitého pro 454 sekvenování použít normalizaci, která exponenciálně zvedne relativní pokrytí málo exprimovaných genů. Nicméně tím bychom do dat mohli vnést velký bias, takže ani tuto variantu jsme neuvažovali.

Hlavním problémem vyhodnocení dat je jejich sekvenační hloubka. V navazující studii bude nezbytné počítat s normalizovanou sekvenční hloubkou např. metodou RPKM (McManus et al. 2010), (Bell et al. 2013). Nelze tedy jednoznačně definovat korelace mezi danými geny, diference mezi jednotlivými typy hybridů. Ze statistického hlediska byla data analyzována zcela adekvátně, ačkoliv mohlo být přitoupeno k sofistikovenjšímu stanovení navržením posteriori pravděpodobností vycházejícího z bayesiánského modelu (Skelly et al. 2011), či využití bayesiánských sítí jako EMASE (*Expectation Maxmization algorithm*) (Munger et al. 2014) pro zajištění vyžší robustnosti testování. Bohužel je nutné podotknout, že hodnoty korigované P hodnoty chi kvadrát testu nejsou spočteny adekvátním způsobem, měli bychom korigovat P hodnoty ne na základě všech SNP, ale pouze těch, u kterých jsme si jisti jejich původem z rodičovských druhů a zároveň SNP, které jsou druhově specifické; tím se dostáváme na frakci původního množství SNP a testování přijme nulovou hypotézu o balancovanosti alel mnohem častěji.

Za další důležitý bod považuji to, že u řady genů jsme našli konfliktní signál disbalance oběma směry. Jakkoliv se toto může zdát paradoxní, část takovýchto genů je vysvětlitelné technickými vlastnostmi dat. Často jsme totiž viděli, že většina SNP v problematických genech má disbalanci jedním směrem a jen jeden, či málo SNP je s nimi v konfliktu. Toto se dá vysvětlit tak, že naše pokrytí přirozené variability rodičovských genomů je celkem děravé a zkrátka jsme některý SNP mylně považovali za diagnostický a on ve skutečnosti nebyl.

Obzvláště důležité je si uvědomit, že některé geny, které se nám jeví jako extrémně vychýleny tak, že jsou v podstatě homozygotní po jednom druhu, nemusí být disbalancovány vůbec. Je totiž možné, že u těchto genů došlo ke konverzi a sekundární ztrátě heterozygotnosti, což se dá zjistit jedině tak, že osekvenujeme patřičný úsek gDNA a potvrdíme, či vyvrátíme homozygotnost, jelikož nezbytné mít na paměti, že v těchto extrémních případech nejsme schopi rozlišit mezi stavem dvou alel v genomu, kdy je jedna z nich umlčena, nebo nedošlo k jevu známému jako genové konverze, tedy fyzickému nahrazení jedné alely podle templátu druhé alely. Případy, kdy se jedná o genové konverze,

jsme schopni detekovat pouze z genomické DNA; bohužel tato analýza nebyla v této studii zatím provedena.

Za další problém lze považovat to, že v mnoha genech jsme prostě žádný vhodný SNP nenašli – přeci jen jsme pracovali s kódujícími sekvencemi, které jsou často pod negativní selekcí, a tudíž množství mutací v takovýchto genech může být malé.

Přes masivní sekvenační úsilí se tak vlastně ukazuje, že genů vůbec vhodných pro náš test je jen zlomek z celkového počtu. Pokud chceme srovnávat více typů hybridů najednou, tak se toto mnžství ješrě snižuje kvůli nekompletnímu průniku mezi zvířaty. Proto se domnívám, že hledání nějakého konkrétního genu, či typu genů, který je/jsou expresně disbalancován/y, nemá valný smysl, protože u většiny genů tento test vůbec nejsme schopni provézt. Spíše má smysl se ptát po obecných trendech, což naše data jasně umožňují.

Domníváme se totiž, že z analyzovaných dat, které by měly být náhodně vybrány, pokud nebudeme předpokládat korelaci mezi sekvenační hloubkou a alel specifckou expresí, je náš soubor výsledků ASE je reprezentovatelný pro stanovení celkové úroveň ASE. Obecně můžeme prohlásit, že značné procento genů (kolem 50%) má průkazně vychýlenou transkripci jedním, či druhým směrem.

Ačkoliv metylom v této analýze studován nebyl, identifikovali jsme DE, které jasně poukazují na změny v umlčování genů. Jelikož je známo, že metylovaný cytozin podléhá rychleší mutační rychlosti je možné vypočítat poměr mezi CT a GA SNP a tím naznačit, zda případě našeho modelového hybridního, může docházet k vyšší, či nižší míře epigenetické regulace (Yebra and Bhagwat 1995). Je známo, že stres vyvolaný polyploidním, hybridizačním šokem může vyvolat metylaci konkrétních genů, jako podobně jako odpověď na environmentální stres, kteréžto mohou být po mnoho generací děděny. V budoucnu plánujeme též sekvenovat metylomy hybridních linií rozdílného stáří, od P1 generace po linie staré až 350 tis let.

Úroveň alelických disbalancí mezi studovanými hybridními modely obecně se značně liší. Procentuální zastoupení disbalancovaných alel se může pohybovat od 73 % naleyených například u kukuřice do pouhých 10 % alelicky disbalancovaných genů u myši (Zhuang and Adams 2007),(Cowles et al. 2002). Bohužel odvozovat závěry týkající se korelace mezi druhovou distancí hybrid v závislotsi na změně počtu, či vychýlení disbalancovaných alel není reálné, jelikož studie se zaměřují zejména na F1 generace ne hybridní linie našeho stáří.

Z našich výsledků je také vypozorovatelné, že poměr disbalancovaných alel závisí na složení hybridního genomu. A to opět způsobem aditivním. Mutace v cis regulačních místech mění např. sílu promotoru, enhanceru, či stabilitu mRNA. Změny v trans regulaci jsou často globálního chrakteru, protože se mění afinita transkripčních faktorů k sekvenčně dependentním loci. Naopak pokud se v cis elementu objeví mutace, postižený gen vykazuje nevyvážený poměr exprese z obou alel. U mutací vzniklých v trans elementech rodičů nedochází v hybridním genomu k disbalanci mezi rodičovskými alelami (Shen et al. 2012). Jelikož pozorujeme relativně vyváženou ASE genomů hybridních jedinců, usuzujeme, že v našich datech převažují zejména efekty trans regulace transkripce. Testovat zda námi detekované rozdíly spadají do kategorie cis či trans regulace prozatím není možné, neboť nám chybí informace F1 hybridů. Regulačnímu vlivu cis a trans elementů jsou zcela stejně vystaveny rodičovské druhy, tak hybridní jedinci (bohužel naši hybridi akumulovali mutace, které mohly vést k novým deregulacím transkripce), proto je teoreticky možné detekovat trans složku alelické disbalance srovnáním rodičovských druhů (Bell et al. 2013). Naše data v tomto ohledu odpovídají zjištěním hybridních kompexů ryb rodu *Poeciliidae* (Shen et al. 2012) nebo hybridů rýže (Zhai et al. 2013).

Umlčení jednoho z rodičovských genomů, bývá také závislé na typu tkáně, vývojového stádia (Adams 2007). Ačkoliv jsme měli možnost vzájemně srovnat pouze tkáň oocytů a jater, nenašli jsme výzarný rozdíl, nebo dokonce opačný trend v disbalanci genů. Musím ale podotknout, že bez normalizace sekvenační hloubky není možné spolehlivě stanovit závěry podobností tkání ani jednotlivých hybridních komplexů.

Vzhledem k tomu, že jsme se celou dobu potýkali s daty pocházejícími z převážně kódujících oblastí, je přírozené se také ptát, zda nalezneme podporu pro obecně citovanou teorii Müllerovy Rohatky. V tomto kontextu zdůrazňuji, že design mé práce je pro takovýto test mimořádně vhodný: ovzorkovali jsme rodičovské sexuální druhy, získali jsme též data z různých typů hybridů a především jsme v jejich rámci měli k dispozici jak asexuální klony evolučně mladé, tak i evolučně dosti staré, u nichž se dá efekt akumulace nesynonymních mutací očekávat především. Testy rohatky se klasicky prováděly tak, že se ze sekvenčních dat jednoho každého lokusu (obvykle pocházejícího z mtDNA, Paland and (Lynch et al. 2008), (Neiman et al. 2010) byly udělány dva fylogenetické stromy, jeden ze všech SNP a druhý jen z kódujících a poté bylo statisticky testováno, zda větve vedoucí ke klonálním liniím mají signifikantně delší větve ve druhém typu stromu. Takovýto test u nás

ovšem nepřicházel v úvahu, neboť my jsme pracovali s di-, či polyploidními lokusy, které navích u asexuálů byly značně heterozygitní díky jejich hybridnímu původu a tudíž konstrukce takovýchto fylogenetických matic nebyla možná. Proto jsme zvolili přístup (Pellino et al. 2013), kdy jsme v podstatě testovali, zda Ka/Ks poměr se liší mezi typy zvířat.

Výsledek analýzy nenaznačuje žádný výrazný nárůst aminokyselinových záměn u asexuálních linií bez ohledu na to, zda se jedná o klony mladé či staré. Tento fakt poněkud kontrastuje s neskutečnou popularitou „mutational“teorií o sexu, neboť existence Müllerovy rohatky a podobných teorií, či jejich derivátů je snad povinně zmiňována v každém článku zabývajícím se asexualitou. Nedávná práce (Neiman et al. 2010) navíc ukazovala na mtDNA lokusu, že k vyššímu tempu akumulace nesynonymních mutací může docházet i u relativně mladých klonů. Naše data naopak nenaznačují, že by sekavčí klony, třeba i 350 tisíc let staré, měly mít nějaký výraznější problém s akumulací škodlivých mutací. Alespoň tey ne ve srovnání s jejich sexuálními protějšky. Jsem si samozřejmě vědom toho, že naše data i analýzy mohou mít řadu objektivních i subjektivních potíží (což se ostatně dá říci o každé studii), ale v následujícím textu ukáži, že negativní výsledk má jasné biologické opodstatnění.

Především bych rád podotknul, že studované sekavčí klony jsou často dominantími formami sekavců a úspěšně konkurují svým sexuálním protějškům (Janko et al. 2012). Náš negativní výsledek je tedy v souladu s terénními daty, jinak si těžko představit, že by mutacemi postižený klon tak úspěšně po stovky tisíc let kompetoval se sexuálními druhy a jinými mladšími klony. Na rozdíl od prací dokazujících existenci rohatky, řada prací ji také nepotvrdila (Pellino et al. 2013) a (Guex et al. 2002) navíc přímo testovala rozdíly ve fitness mladých a starých klonů žab rodu Pelophylax a nepotvrdila je, což je také v rozporu s teoretickými předpoklady.

(Janko et al. 2011) použil originální populačně genetickou metodu pro detekci stop Müllerovy rohatky ze sekvenčních dat aseuálních komplexů jako sekavec a jiných a ukázal, že data jsou obecně v rozporu s očekávanými vzory a navíc se ukázalo, že řada asexuálních taxonů ani nejeví vyšší tempo extinkce ve srovnání se sexuálními druhy (Liu et al. 2012). Naše data tedy zapadají do rostoucího množství evidence, že asexuálové nijak zvláště nesynonymní mutace neakumulují. Tuto negativní evidenci, podpořenou našimi daty pak (Janko et al. 2008) neinterpretuje tak, že Müllerova rohatka, či jiné podobné procesy neexistují. Naopak, není důvod se domnívat, že by například sekavci s jejich

striktně asexuálním rozmnožováním neměli postupně hromadit mutace. Spíše se však zdá, že časový interval, po který klony existují, je příliš krátký, aby se tento proces projevil. (Janko et al. 2012) razí teorii, že většina klonů je z populace odstraněna driftem a jinými procesy dříve, než se u nich vůbec mohou „long-term costs of asexuality“ vůbec projevit, což moje data podporují.

Souhrn

Moje práce měla od svého počátku velice dynamický průběh s řadou změn podle toho, jak se postupně ukazovaly probmémy a nové otázky související se složitostí analyzovaných dat. Právě tato komplexita způsobila, že nebylo možné se zaměřit na jedinou otázku například testu Müllerovy rohatky, protože práce musela postupovat v jednotlivých hierarchických krocích, jež se musely také zpětně validovat. Nakonec jsem tedy přispěl odpověďmi k několika okruhům otázek, ale na druhou stranu jsem si plně vědom faktu, že řada závěrů je stále předčasná a bude vyžadovat odatečné analýzy před publikací. Částečně se tím budu zabývat v navazujícím PGS.

Každopádně moje práce umožnila tvorbu a validaci relativně kvalitního referenčního transkriptomu nemodelového, leč vědecky významného, organismu, na nějž teprve poté bylo možno věrohodně mapovat získané sekvence a testovat obsáhlejsí hypotézy.

Určil jsem také řadu genů, které mají jednoznačně diferencovanou expresi mezi jednotlivými formami sekavců, a dokonce zjistil některé obecnější prvky, které jsou podobné i u jiných asexuálních organismů, u nichž přitom asexuality vznikla zcela nezávisle na našem organizmu. Na druhou stranu jsem však také našel řadu genů, jejich exprese se sice také jasně lišila mezi sexuálními a klonálními formami, ale které u žádných jiných organismů doposud nepadly v podezření, že by mohly s asexualitou souviset. Tím jsem v podstatě otevřel pole pro následné cílenější studie, které mohou studovat biologickou validitu těchto kandidátních genů.

Transkripční analýza dále ukázala intermediaritu studovaných hybridů a jasný „gene dose“ efekt obecné úrovně transkripce, což je v úzké korelaci s morfologickými i ekologickými daty mých kolegů. Naznačuje to, že víceméně lineární efekt genové dávky se u sekavčích hybridů a polyploidů projevuje na celé ontogenetické škále od genotypu, přes expresi genů, morfologickou plasticitu až po interakce s okolním prostředím.

Ukázalo se dále, že sice na celotranskriptomové úrovni rozhodně neexistuje nějaká systematická tendence k umlčení jednoho rodičovského genomu a hybridi víceméně exprimují oba genomy (až na výjimky, kdy jsem našel také jednotlivé geny exprimované výhradně alelami jediného rodičovského druhu), na druhou stranu je však velice rozšířená over exprese jedné rodičovské alely oproti druhé u mnoha genů. Směr vychýlení exprese se kupodivu zdál odlišný mezi jedniotlivými typy hybridů, což dále může přispívat k pozorovanému lineárnímu gradientu podobností hybridů a di- polyploidů k jejich rodičovským druhům.

Za významný nakonec považuji fakt, že jem v práci nepotvrdil výraznější tendenci asexuálů k akumulaci nesynonymních záměn, což sice odporuje obecnému očekávání, ale zapadá do rostoucí řady jiných podobných důkazů.

Myslím si, že moje práce patří k nejkomplexnějším genomickým studiím, které na asexuálních organismech byly podniknuty a tomu také odpovídá nejen její rozsah, ale závažnost zjištění. Výsledky, které jsem zde prezentoval, budu dále rozvíjet do formy několika publikací.

Seznam užitých zkratek

A	adenin
blast	eng - *basic local alignment search tool*
blastn	eng - *basic local alignment search tool* nukleotidové sekvence vůči nukleotidové sekvenci
blastx	eng - *basic local alignment search tool* přeložených nukleotidových sekvencí do proteinu v 6 čtecích rámcích vůči proteinové sekvenci
C	cytosin
PCA	mnohorozměrná analýza – analýza hlavních komponent
PcoA	mnohorozměrná analýza – analýza hlavních koordinát
MDS	mnohorozměrná analýza – multidimenzionální škálování (v případě užití euklieánské distance je výsledek identický s PCA)
cDNA	komplementární DNA - pocházející z mRNA
CpG	„ostrovy“ cytosinu a guaninu – regulační fce, promotory
DE	diferenciálně exprimované geny v rámci definovaných skupin
DNA	deoxy-ribonukleová kyselina
VCF	*variant call format*
DTT	di-thio treitol
ee	*Cobitis elongatoides*
eet	triploidní hybrid druhů C. elongatoides a C. taenia
et	diploidní hybrid druhů C. elongatoides a C. taenia
etn	triploidní hybrid druhů C. elongatoides, C. taenia a C. tanaetica
ett	triploidní hybrid druhů C. elongatoides a C. taenia
G	guanin
gDNA	genomická DNA
GI	Identifikátor genu
GO	genová ontologie (kontrolovaný slovník přiřazující genům známé atributy molekulární funkce, buněčné lokalizace a biologických procesů)
KEGG	eng - *Kyoto Encyclopedia of Genes AND Genomes*, databáze signálních a metabolických drah
lncRNA	dlouhé nekdující RNA

LOH	eng – *loss of heterozygosity*, ztráta heterozygotnosti
lom300tt	játra oocyty 300 bp limit C. taenia (454 normalizovaný transkriptom)
N	ambiguidní báze (A, T, C, či G)
ORF	eng – *open reading frame*, otevřený čtecí rámec
PCR	eng – *polaymerase chain reaction,* polymerázová řetězová reakce
qPCR	kvantitativní řetězová reakce
PE	eng - *pair end*, sekvenování fragmentu ssDNA v obou směrech
Q	Phered skóre: záporný dekadický logatitmus pravděpodobnosti přečtení chybné báze
RNAseq	sekvenování nenormalizované RNA skrze reverzeně tanskribovanou cDNA
SE	eng - *single end,* sekvenování fragmentu ssDNA v jednom směru
SeqCap	sekvenování vzorku se sníženou komplexitou aplikací hybridizačních sond cílených na gDNA
SNP	jedno-nukleotidové polymorfismy
T	tymin
tt	*Cobitis tenia*
UTR	netranslatované regulační oblasti na 3’ a 5’ konci mRNA (fce - lokalizace mRNA, účinnost translace a stabilita mRNA)
454	Pyrosekvenování (syntézní sekvenační metoda založená na bioluminiscenci pyrofosfátu vznikajícího při polymeraci)

Bibliografie

Adams, K. L. 2007. Evolution of duplicate gene expression in polyploid and hybrid plants. J. Hered. 98:136–141.

Andersen, C. L., J. L. Jensen, and T. F. Ørntoft. 2004. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res. 64:5245–5250.

Anders, S., and W. Huber. 2010. Differential expression analysis for sequence count data. Genome Biol. 11:R106.

Auer, P. L., and R. W. Doerge. 2010. Statistical Design and Analysis of RNA Sequencing Data. Genetics 185:405–416.

Bajgain, P., B. A. Richardson, J. C. Price, R. C. Cronn, and J. A. Udall. 2011. Transcriptome characterization and polymorphism detection between subspecies of big sagebrush (Artemisia tridentata). BMC Genomics 12:370.

Beaudet, A. L., and Y. H. Jiang. 2002. A rheostat model for a rapid and reversible form of imprinting-dependent evolution. Am. J. Hum. Genet. 70:1389–1397.

Bell, G. D. M., N. C. Kane, L. H. Rieseberg, and K. L. Adams. 2013. RNA-Seq Analysis of Allele-Specific Expression, Hybrid Effects, and Regulatory Divergence in Hybrids Compared with Their Parents from Natural Populations. Genome Biol. Evol. 5:1309–1323.

Bengtsson, B. O. 2009. Asex and Evolution: A Very Large-Scale Overview. Pp. 1–19 *in* I. Schön, K. Martens, and P. Dijk, eds. Lost Sex. Springer Netherlands, Dordrecht.

Beukeboom, L. W., and R. C. Vrijenhoek. 1998. Evolutionary genetics and ecology of sperm-dependent parthenogenesis. J. Evol. Biol. 11:755–782.

Birchler, J. A., H. Yao, S. Chudalayandi, D. Vaiman, and R. A. Veitia. 2010. Heterosis. Plant Cell 22:2105–2112.

Blekhman, R., A. Oshlack, and Y. Gilad. 2009. Segmental Duplications Contribute to Gene Expression Differences Between Humans and Chimpanzees. Genetics 182:627–630.

Bohlen, J., A. Perdices, I. Doadrio, and P. S. Economidis. 2006. Vicariance, colonisation, and fast local speciation in Asia Minor and the Balkans as revealed from the phylogeny of spined loaches (Osteichthyes; Cobitidae). Mol. Phylogenet. Evol. 39:552–561.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma. Oxf. Engl. 30:2114–2120.

Bolnick, D. I., and T. J. Near. 2005. Tempo of hybrid inviability in centrarchid fishes (Teleostei: Centrarchidae). Evol. Int. J. Org. Evol. 59:1754–1767.

Burge, S. W., J. Daub, R. Eberhardt, J. Tate, L. Barquist, E. P. Nawrocki, S. R. Eddy, P. P. Gardner, and A. Bateman. 2012. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. gks1005.

California, I. J. A. U. of. 2008. Clonality : The Genetics, Ecology, and Evolution of Sexual Abstinence in Vertebrate Animals: The Genetics, Ecology, and Evolution of Sexual Abstinence in Vertebrate Animals. Oxford University Press, USA.

CARMAN, J. G. 1997. Asynchronous expression of duplicate genes in angiosperms may cause apomixis, bispory, tetraspory, and polyembryony. Biol. J. Linn. Soc. 61:51–94.

Charif, D., and J. R. Lobry. 2007. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. Pp. 207–232 *in* D. U. Bastolla, P. D. M. Porto, D. H. E. Roman, and D. M. Vendruscolo, eds. Structural Approaches to Sequence Evolution. Springer Berlin Heidelberg.

Chen, Z. J. 2010. Molecular mechanisms of polyploidy and hybrid vigor. Trends Plant Sci. 15:57–71.

Choleva, L., K. Janko, K. De Gelas, J. Bohlen, V. Šlechtová, M. Rábová, and P. Ráb. 2012. Synthesis of clonality and polyploidy in vertebrate animals by hybridization between two sexual species. Evol. Int. J. Org. Evol. 66:2191–2203.

Choleva, L., Z. Musilova, A. Kohoutova-Sediva, J. Paces, P. Rab, and K. Janko. 2014. Distinguishing between Incomplete Lineage Sorting and Genomic Introgressions: Complete Fixation of Allospecific Mitochondrial DNA in a Sexually Reproducing Fish (Cobitis; Teleostei), despite Clonal Reproduction of Hybrids. PLoS ONE 9:e80641.

Comai, L. 2005. The advantages and disadvantages of being polyploid. Nat. Rev. Genet. 6:836–846.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinforma. Oxf. Engl. 21:3674–3676.

Cowles, C. R., J. N. Hirschhorn, D. Altshuler, and E. S. Lander. 2002. Detection of regulatory variation in mouse genes. Nat. Genet. 32:432–437.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

De Muyt, A., L. Pereira, D. Vezon, L. Chelysheva, G. Gendrot, A. Chambon, S. Laine-Choinard, G. Pelletier, R. Mercier, F. Nogue, and M. Grelon. 2009. A High Throughput Genetic Screen Identifies New Early Meiotic Recombination Functions in Arabidopsis thaliana. Plos Genet. 5:e1000654.

d'Erfurth, I., S. Jolivet, N. Froger, O. Catrice, M. Novatchkova, M. Simon, E. Jenczewski, and R. Mercier. 2008. Mutations in AtPS1 (Arabidopsis thaliana Parallel Spindle 1) Lead to the Production of Diploid Pollen Grains. PLoS Genet 4:e1000274.

Dillies, M.-A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, F. Jaffrézic, and French StatOmique Consortium. 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief. Bioinform. 14:671–683.

Dilworth, R. P. 1950. A Decomposition Theorem for Partially Ordered Sets. Ann. Math. 51:161–166.

Dubois, A. 2011. Species and "strange species" in zoology: Do we need a "unified concept of species"? Comptes Rendus Palevol 10:77–94.

Eads, B. D., D. Tsuchiya, J. Andrews, M. Lynch, and M. E. Zolan. 2012. The spread of a transposon insertion in Rec8 is associated with obligate asexuality in Daphnia. Proc. Natl. Acad. Sci. U. S. A. 109:858–863.

Flegr, J. 2007. Úvod do evoluční biologie. Vyd. 1. Academia, Praha.

Fontaneto, D., E. A. Herniou, C. Boschetti, M. Caprioli, G. Melone, C. Ricci, and T. G. Barraclough. 2007. Independently evolving species in asexual bdelloid rotifers. PLoS Biol. 5:e87.

Galla, G., H. Vogel, T. F. Sharbel, and G. Barcaccia. 2015. De novo sequencing of the Hypericum perforatum L. flower transcriptome to identify potential genes that are related to plant reproduction sensu lato. BMC Genomics 16:254–275.

Gavery, M. R., and S. B. Roberts. 2012. Characterizing short read sequencing for gene discovery and RNA-Seq analysis in Crassostrea gigas. Comp. Biochem. Physiol. Part D Genomics Proteomics 7:94–99.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652.

Guerreiro, M. P. G. 2014. Interspecific hybridization as a genomic stressor inducing mobilization of transposable elements in Drosophila. Mob. Genet. Elem. 4:e34394.

Guex, G.-D., H. Hotz, and R. D. Semlitsch. 2002. Deleterious alleles and differential viability in progeny of natural hemiclonal frogs. Evol. Int. J. Org. Evol. 56:1036–1044.

Hansen, K. D., S. E. Brenner, and S. Dudoit. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 38:e131.

Howard, R. 1994. Selection Against Deleterious Mutations and the Maintenance of Biparental Sex. Theor. Popul. Biol. 45:313–323.

Janko, K., J. Bohlen, D. Lamatsch, M. Flajshans, J. T. Epplen, P. Ráb, P. Kotlík, and V. Slechtová. 2007. The gynogenetic reproduction of diploid and triploid hybrid spined loaches (Cobitis: Teleostei), and their ability to establish successful clonal lineages--on the evolution of polyploidy in asexual vertebrates. Genetica 131:185–194.

Janko, K., M. A. Culling, P. Ráb, and P. Kotlík. 2005. Ice age cloning--comparison of the Quaternary evolutionary histories of sexual and clonal forms of spiny loaches (Cobitis; Teleostei) using the analysis of mitochondrial DNA variation. Mol. Ecol. 14:2991–3004.

Janko, K., P. Drozd, and J. Eisner. 2011. Do clones degenerate over time? Explaining the genetic variability of asexuals through population genetic models. Biol. Direct 6:17.

Janko, K., P. Drozd, J. Flegr, and J. R. Pannell. 2008. Clonal Turnover Versus Clonal Decay: A Null Model for Observed Patterns of Asexual Longevity, Diversity and Distribution. Evolution 62:1264–1270.

Janko, K., and J. Eisner. 2009. Sperm-dependent parthenogens delay the spatial expansion of their sexual hosts. J. Theor. Biol. 261:431–440.

Janko, K., J. Kotusz, K. De Gelas, V. Šlechtová, Z. Opoldusová, P. Drozd, L. Choleva, M. Popiołek, and M. Baláž. 2012. Dynamic Formation of Asexual Diploid and Polyploid

Lineages: Multilocus Analysis of Cobitis Reveals the Mechanisms Maintaining the Diversity of Clones. PLoS ONE 7:e45384.

Johnson, S. G., and E. Bragg. 1999. Age and Polyphyletic Origins of Hybrid and Spontaneous Parthenogenetic Campeloma (Gastropoda: Viviparidae) from the Southeastern United States. Evolution 53:1769–1781.

Juchno, D., and A. Boroń. 2006. Age, reproduction and fecundity of the spined loach Cobitis taenia L. (Pisces, Cobitidae) from Lake Klawój (Poland). Reprod. Biol. 6:133–148.

Kondrashov, A. S. 1993. Classification of Hypotheses on the Advantage of Amphimixis. J. Hered. 84:372–387.

Kondrashov, A. S. 1988. Deleterious mutations and the evolution of sexual reproduction. Nature 336:435–440.

Kondrashov, A. S. 1994. Muller's Ratchet under Epistatic Selection. Genetics 136:1469–1473.

Koonin, E. V. 2005. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 39:309–338.

Kotusz, J., M. Popiołek, P. Drozd, K. De Gelas, V. Šlechtová, and K. Janko. 2014. Role of parasite load and differential habitat preferences in maintaining the coexistence of sexual and asexual competitors in fish of the Cobitis taenia hybrid complex. Biol. J. Linn. Soc. 113:220–235.

Langmead, B., K. D. Hansen, and J. T. Leek. 2010. Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol. 11:R83.

Lee, K., C. Wang, L. Spate, C. N. Murphy, R. S. Prather, and Z. Machaty. 2014a. Gynogenetic Activation of Porcine Oocytes. Cell. Reprogramming 16:121–129.

Lee, W.-P., M. P. Stromberg, A. Ward, C. Stewart, E. P. Garrison, and G. T. Marth. 2014b. MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. PLoS ONE 9:e90581.

Li, C.-Y., J.-T. Li, Y.-Y. Kuang, R. Xu, Z.-X. Zhao, G.-Y. Hou, H.-W. Liang, and X.-W. Sun. 2014. The Transcriptomes of the Crucian Carp Complex (Carassius auratus) Provide Insights into the Distinction between Unisexual Triploids and Sexual Diploids. Int. J. Mol. Sci. 15:9386–9406.

Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25:2078–2079.

Liu, H.-M., R. J. Dyer, Z.-Y. Guo, Z. Meng, J.-H. Li, and H. Schneider. 2012. The Evolutionary Dynamics of Apomixis in Ferns: A Case Study from Polystichoid Ferns. J. Bot. 2012:e510478.

Li, W., and A. Godzik. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinforma. Oxf. Engl. 22:1658–1659.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36:96–99.

Lovén, J., D. A. Orlando, A. A. Sigova, C. Y. Lin, P. B. Rahl, C. B. Burge, D. L. Levens, T. I. Lee, and R. A. Young. 2012. Revisiting Global Gene Expression Analysis. Cell 151:476–482.

Lynch, M., A. Seyfert, B. Eads, and E. Williams. 2008. Localization of the Genetic Determinants of Meiosis Suppression in Daphnia pulex. Genetics 180:317–327.

Mark Welch, D., and M. Meselson. 2000. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. Science 288:1211–1215.

McManus, C. J., J. D. Coolon, M. O. Duff, J. Eipper-Mains, B. R. Graveley, and P. J. Wittkopp. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Res. 20:816–825.

Mihola, O., Z. Trachtulec, C. Vlcek, J. C. Schimenti, and J. Forejt. 2009. A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase. Science 323:373–375.

Moore, T., and D. Haig. 1991. Genomic Imprinting in Mammalian Development - a Parental Tug-of-War. Trends Genet. 7:45–49.

Morishima, K., H. Yoshikawa, and K. Arai. 2008. Meiotic hybridogenesis in triploid Misgurnus loach derived from a clonal lineage. Heredity 100:581–586.

Moritz, C., T. Uzzell, C. Spolsky, H. Hotz, I. Darevsky, L. Kupriyanova, and F. Danielyan. 1992. The material ancestry and approximate age of parthenogenetic species of Caucasian rock lizards (Lacerta: Lacertidae). Genetica 87:53–62.

Munger, S. C., N. Raghupathy, K. Choi, A. K. Simons, D. M. Gatti, D. A. Hinerfeld, K. L. Svenson, M. P. Keller, A. D. Attie, M. A. Hibbs, J. H. Graber, E. J. Chesler, and G. A. Churchill. 2014. RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. Genetics 198:59–73.

Nadeau, N. J., A. Whibley, R. T. Jones, J. W. Davey, K. K. Dasmahapatra, S. W. Baxter, M. A. Quail, M. Joron, R. H. ffrench-Constant, M. L. Blaxter, J. Mallet, and C. D. Jiggins. 2012. Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. Philos. Trans. R. Soc. Lond. B Biol. Sci. 367:343–353.

Nakatani, M., M. Miya, K. Mabuchi, K. Saitoh, and M. Nishida. 2011. Evolutionary history of Otophysi (Teleostei), a major clade of the modern freshwater fishes: Pangaean origin and Mesozoic radiation. BMC Evol. Biol. 11:177.

Neiman, M., G. Hehman, J. T. Miller, J. M. Logsdon, and D. R. Taylor. 2010. Accelerated mutation accumulation in asexual lineages of a freshwater snail. Mol. Biol. Evol. 27:954–963.

Nomikos, M. 2015. Novel signalling mechanism and clinical applications of sperm-specific PLC zeta. Biochem. Soc. Trans. 43:371–376.

Novaes, E., D. R. Drost, W. G. Farmerie, G. J. Pappas, D. Grattapaglia, R. R. Sederoff, and M. Kirst. 2008. High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics 9:312.

Otto, S. P., and J. Whitton. 2000. Polyploid incidence and evolution. Annu. Rev. Genet. 34:401–437.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289–290.

Parchman, T. L., Z. Gompert, M. J. Braun, R. T. Brumfield, D. B. McDonald, J. a. C. Uy, G. Zhang, E. D. Jarvis, B. A. Schlinger, and C. A. Buerkle. 2013. The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. Mol. Ecol. 22:3304–3317.

Pellino, M., D. Hojsgaard, T. Schmutzer, U. Scholz, E. Hörandl, H. Vogel, and T. F. Sharbel. 2013. Asexual genome evolution in the apomictic Ranunculus auricomus complex: examining the effects of hybridization and mutation accumulation. Mol. Ecol. 22:5908–5921.

Pickrell, J. K., J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464:768–772.

Qi, W., Z. Tang, and H. Yu. 2006. Phosphorylation- and Polo-Box–dependent Binding of Plk1 to Bub1 Is Required for the Kinetochore Localization of Plk1. Mol. Biol. Cell 17:3705–3716.

Ramanna, M. S., and E. Jacobsen. 2003. Relevance of sexual polyploidization for crop improvement - A review. Euphytica 133:3–18.

Ravi, M., M. P. A. Marimuthu, and I. Siddiqi. 2008. Gamete formation without meiosis in Arabidopsis. Nature 451:1121–U10.

Rehrauer, H., L. Opitz, G. Tan, L. Sieverling, and R. Schlapbach. 2013. Blind spots of quantitative RNA-seq: the limits for assessing abundance, differential expression, and isoform switching. BMC Bioinformatics 14:370.

Robinson, M. D., and A. Oshlack. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11:R25.

Ruffalo, M., M. Koyuturk, S. Ray, and T. LaFramboise. 2012. Accurate estimation of short read mapping quality for next-generation genome sequencing. Bioinformatics 28:i349–i355.

Russell, S. T. 2003. Evolution of intrinsic post-zygotic reproductive isolation in fish. Ann. Zool. Fenn. 40:321–329.

Salathe, M., R. D. Kouyos, and S. Bonhoeffer. 2008. The state of affairs in the kingdom of the Red Queen. Trends Ecol. Evol. 23:439–445.

Salmon, A., and M. L. Ainouche. 2010. Polyploidy and DNA methylation: new tools available. Mol. Ecol. 19:213–215.

Sartor, M. A. 2006. A new method to remove hybridization bias for interspecies comparison of global gene expression profiles uncovers an association between mRNA sequence divergence and differential gene expression in Xenopus. Nucleic Acids Res. 34:185–200.

Schultz, R. J. 1969. Hybridization, Unisexuality, and Polyploidy in the Teleost Poeciliopsis (Poeciliidae) and Other Vertebrates. Am. Nat. 103:605–619.

Seehausen, O., R. K. Butlin, I. Keller, C. E. Wagner, J. W. Boughman, P. A. Hohenlohe, C. L. Peichel, G.-P. Saetre, C. Bank, A. Brännström, A. Brelsford, C. S. Clarkson, F. Eroukhmanoff, J. L. Feder, M. C. Fischer, A. D. Foote, P. Franchini, C. D. Jiggins, F. C. Jones, A. K. Lindholm, K. Lucek, M. E. Maan, D. A. Marques, S. H. Martin, B. Matthews, J. I. Meier, M. Möst, M. W. Nachman, E. Nonaka, D. J. Rennison, J. Schwarzer, E. T. Watson, A. M. Westram, and A. Widmer. 2014. Genomics and the origin of species. Nat. Rev. Genet. 15:176–192.

Shen, Y., J. Catchen, T. Garcia, A. Amores, I. Beldorth, J. Wagner, Z. Zhang, J. Postlethwait, W. Warren, M. Schartl, and R. B. Walter. 2012. Identification of transcriptome SNPs between Xiphophorus lines and species for assessing allele specific gene expression within F_1 interspecies hybrids. Comp. Biochem. Physiol. Toxicol. Pharmacol. CBP 155:102–108.

Skelly, D. A., M. Johansson, J. Madeoy, J. Wakefield, and J. M. Akey. 2011. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. Genome Res. 21:1728–1737.

Snell, T. W., and N. J. D. DesRosiers. 2008. Effect of progesterone on sexual reproduction of Brachionus manjavacas (Rotifera). J. Exp. Mar. Biol. Ecol. 363:104–109.

Stith, B. J. 2015. Phospholipase C and D regulation of Src, calcium release and membrane fusion during Xenopus laevis development. Dev. Biol. 401:188–205.

Tirosh, I., A. Weinberger, D. Bezalel, M. Kaganovich, and N. Barkai. 2008. On the relation between promoter divergence and gene expression evolution. Mol. Syst. Biol. 4.

Trivedi, U. H., T. CÃ©zard, S. Bridgett, A. Montazam, J. Nichols, M. Blaxter, and K. Gharbi. 2014. Quality control of next-generation sequencing data without a reference. Front. Genet. 5.

Waggoner, B. M., and G. O. P. Jr. 1993. Fossil habrotrochid rotifers in Dominican amber. Experientia 49:354–357.

Wang, C.-J., L.-Q. Zhang, S.-F. Dai, Y.-L. Zheng, H.-G. Zhang, and D.-C. Liu. 2010. Formation of unreduced gametes is impeded by homologous chromosome pairing in tetraploid Triticum turgidum x Aegilops tauschii hybrids. Euphytica 175:323–329.

Wang, C., L. Wei, M. Guo, and Q. Zou. 2013. Computational Approaches in Detecting Non- Coding RNA. Curr. Genomics 14:371–377.

Wang, J., L. H. Ye, Q. Z. Liu, L. Y. Peng, W. Liu, X. G. Yi, Y. D. Wang, J. Xiao, K. Xu, F. Z. Hu, L. Ren, M. Tao, C. Zhang, Y. Liu, Y. H. Hong, and S. J. Liu. 2015. Rapid genomic DNA changes in allotetraploid fish hybrids. Heredity, doi: 10.1038/hdy.2015.3.

Weiss-Schneeweiss, H., K. Emadzade, T.-S. Jang, and G. M. Schneeweiss. 2013. Evolutionary Consequences, Constraints and Potential of Polyploidy in Plants. Cytogenet. Genome Res. 140.

Wilhelm, B. T., and J.-R. Landry. 2009. RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48:249–257.

Wu, R., X. Wang, Y. Lin, Y. Ma, G. Liu, X. Yu, S. Zhong, and B. Liu. 2013. Inter-Species Grafting Caused Extensive and Heritable Alterations of DNA Methylation in Solanaceae Plants. Plos One 8:e61995.

Wu, T. D., and S. Nacu. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26:873–881.

Xing, Y., and Q. Zhang. 2010. Genetic and molecular bases of rice yield. Annu. Rev. Plant Biol. 61:421–442.

Yebra, M. J., and A. S. Bhagwat. 1995. A cytosine methyltransferase converts 5-methylcytosine in DNA to thymine. Biochemistry (Mosc.) 34:14752–14757.

Zhai, R., Y. Feng, X. Zhan, X. Shen, W. Wu, P. Yu, Y. Zhang, D. Chen, H. Wang, Z. Lin, L. Cao, and S. Cheng. 2013. Identification of Transcriptome SNPs for Assessing Allele-Specific Gene Expression in a Super-Hybrid Rice Xieyou9308. PLoS ONE 8:e60668.

Zhang, Q., K. Arai, and M. Yamashita. 1998. Cytogenetic mechanisms for triploid and haploid egg formation in the triploid loach Misgurnus anguillicaudatus. J. Exp. Zool. 281:608–619.

Zhang, Z., and J. Yu. 2006. Evaluation of Six Methods for Estimating Synonymous and Nonsynonymous Substitution Rates. Genomics Proteomics Bioinformatics 4:173–181.

Zhou, X., H. Lindsay, and M. D. Robinson. 2014. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res. 42:e91.

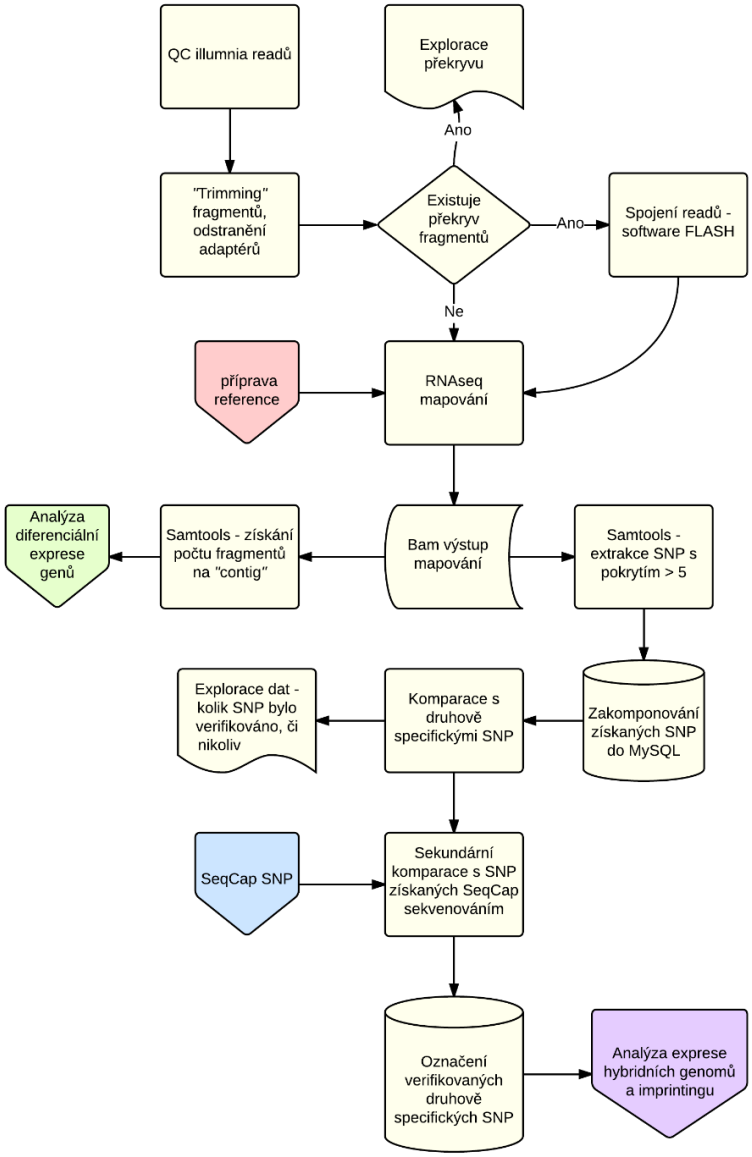
Zhuang, Y., and K. L. Adams. 2007. Extensive allelic variation in gene expression in populus F1 hybrids. Genetics 177:1987–1996.

Zhulidov, P. A., E. A. Bogdanova, A. S. Shcheglov, I. A. Shagina, L. L. Wagner, G. L. Khazpekov, V. V. Kozhemyako, S. A. Lukyanov, and D. A. Shagin. 2005. A method for

the preparation of normalized cDNA libraries enriched with full-length sequences. Russ. J. Bioorganic Chem. 31:170–177.

Zhu, Y. Y., E. M. Machleder, A. Chenchik, R. Li, and P. D. Siebert. 2001. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. BioTechniques 30:892–897.

Přílohy


QC illumnia readů
Explorace překryvu
Ano
"Trimming" fragmentů, odstranění adaptérů
Existuje překryv fragmentů
Ano
Spojení readů - software FLASH
Ne
příprava reference
RNAseq mapování
Analýza diferenciální exprese genů
Samtools - získání počtu fragmentů na "contig"
Bam výstup mapování
Samtools - extrakce SNP s pokrytím > 5
Explorace dat - kolik SNP bylo verifikováno, či nikoliv
Komparace s druhově specifickými SNP
Zakomponování získaných SNP do MySQL
SeqCap SNP
Sekundární komparace s SNP získaných SeqCap sekvenováním
Označení verifikovaných druhově specifických SNP
Analýza exprese hybridních genomů a imprintingu