Univerzita Karlova v Praze

Přírodovědecká fakulta

Studijní program:

Molekulární biologie, genetika a virologie

Bc. Jan Röslein

Mutační a substituční tempo u sexuálních a klonálních 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, 2016

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.

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 Heidelbergu dne 10. 8. 2016

Podpis:

NÁZEV:

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

AUTOR:
Jan Röslein

KATEDRA (ÚSTAV):
Ústav živočišné fyziologie a genetiky AVČR, v.v.i.

VEDOUCÍ PRÁCE:
Mgr. Karel Janko, Ph.D.

ABSTRAKT:

Předmětem diplomové práce je testování několika hypotéz o evoluci asexuálního rozmnožování u modelové skupiny ryb rodu sekavec a jeho udržení při vzájemné kompetici sexuálních a asexuálních forem, čímž se dotkne jedné z nejstarších a zároveň nevyřešených otázek biologie. Konkrétně se práce zabývá otázkou akumulace nesynonymních mutací, jejichž zrychlená akumulace v genomu klonálních linií teoreticky vede ke zvýšené extinkci oproti sexuálně se množícím populacím (tzv. teorie Mullerovy rohatky a Kondrashovovy sekerky). Diplomová práce je založena na sekvenačních datech normalizované cDNA tkáně oocytů a jater, která posloužila jako základní matrice pro vytvořený transkriptom a dat nenormalizované cDNA sekvenace (RNAseq), jež posloužily k validaci získaných polymorfismů, detekci diferenciální exprese a alelově specifické exprese (ASE) hybridních biotypů. Diplomová práce tedy nastiňuje široké spektrum hypotéz týkajících se evoluce hybridních linií rodu sekavec, dále též následků, vlivů polyplidizace a hybridizace na transkriptom. Výsledky evoluce hybridních linií poukazují na zvýšenou akumulaci nesynonymních mutací v genomu hybridních linií v závislosti na jejich stáří, která nicméně prozatím nevede k znatelné degeneraci hybridních linií. Získaný pohled na transkriptom oocytů mezi asexuálními a sexuálními jedinci odpovídá zjištěním recentních poznatků a odhaluje velké množství nevysvětlených fenoménů, které mohou mít funkční selekčně pozitivní či negativní vliv na gynogentické jedince. Z hlediska ASE se hybridní jedinci jeví vyváženě v závislosti na složení genomu.

KLÍČOVÁ SLOVA:
Evoluce sexu, gynogeneze, akumulace škodlivých mutací, sekvenování nové generace, alelově specifická exprese, hybridizace

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

AUTOR:
Jan Röslein

DEPARTMENT:
Ústav živočišné fyziologie a genetiky AVČR, v.v.i.

SUPERVISOR:
Mgr. Karel Janko, Ph.D.

ABSTRACT:
Subject of this thesis is to test several hypotheses about the evolution of asexual reproduction in model group of fish family *Cobitis* and its mutual competition among sexual and asexual forms, which touches one of the oldest unresolved issues of biology. Specifically, the work deals with the accumulation of non-synonymous mutations, which accelerated accumulation in the genome of clonal lineages theoretically leads to increased extinction compared with sexually reproducing populations (so-called. The theory of Muller's ratchet and Kondrashov´s hatchet). This thesis is based on a normalized cDNA sequencing data from oocytes and liver tissue, which has served as a base matrix (generated based on non-normalized cDNA data) for transcriptome sequencing (RNAseq). Consequently the RNAseq data have served as validation for acquired polymorphisms, detection of differential expression of allele-specific expression (ASE) hybrid biotypes. This diploma thesis balances among the edges of vast spectrum of hypotheses regarding the evolution of the genus hybrid lineages, also consequences and the impact of polyploidisation, hybridization on transcription. Results concerning evolution of hybrid lines pointed out increased accumulation of non-synonymous mutations in the genome of the hybrid lineages in dependence on their age, which for the time being, however, does not lead to a noticeable degeneration in hybrid lineages. The overall view on the oocyte transcriptome between sexual asexual species reveals several differences either correlating with recent findings or points out unexplained phenomena, which may or may not have functionally positive or negative selective influence in gynogenetic individuals. ASE in hybrid genomes appear to be balanced, depending on the composition of the genome.

KEYWORDS:
Evolution of sex, gynogenesis, accumulation of deleterious mutations, next generation sequencing, allele specific expression, hybridization

Obsah

1 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 vzniklé 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 existence 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 divergovanými 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 lokalizovány do specifických oblastí v genomu, či výrazně roztroušeny (rev. v (Seehausen et al., 2014)) (viz e.g. (Parchman et al., 2013)).

Ať již však ke speciaci dochází jakkoli, zdá se, že speciační proces má charakter kontinua, kdy je míra reprodukční izolace mezi druhy úměrná jejich generické distanci – 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 navrhl 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 meiózy, 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 meiózy.

1.1 Asexualita

Termín asexualita je používán především ve spojení s eukaryotními organismy, jelikož výlučně eukaryota přešly 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, jež často vedou 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 pak 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.

1.1.1 Mechanismy vzniku neredukovaných gamet

Změny v meióze vedoucí k asexuálnímu rozmnožování byly studovány u rostlin, a to velmi detailně, 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 vzniku univalentů, což vede k disbalancované segregaci anafázi I a disbalancované segregaci v anafázi II, především pokud také dojde k předčasné ztrátě kohezinového komplexu 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 komplexu) bránící 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 mírně liší. Krytosemenné rostliny po první telofázi I zůstávají ve společné 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.

1.1.2 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 pohlavní buněk je mnoho. Gynogeneze bývá spojena z polyploidií, a to linií tvořenými pouze samicemi nebo hermafrodity. Gynogeneze byla nalezena u kmenů *Chordata*, *Mollusca*, *Annelida*, *Arthropoda*, *Rotifera* a *Platyhelmites* (Beukeboom and Vrijenhoek, 1998).

1.1.3 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 energie 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 dává hostiteli velkou výhodu v obraně před parazity (Flegr, 2007)(Salathe et al., 2008). Naopak u hybridů, asexuálů, by se dal očekávat díky heteroznímu efektu, odlišného 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 jedince, 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 má nekonečnou efektivní velikost (drift nehraje roli), pak průměrný počet mutací na chromosomu bude nepřímo úměrný mutační rychlosti. U klonálních linií 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 ke 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 separaci 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).

1.2 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, a 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 je doloženo, 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 trihybridních 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 k

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 k inkorporacím jeho genomů. Tak došlo k tomu, že nejstarší známá asexuální hybridní linie patrně vznikla před cca 350 tis. lety jako diploidní EN forma, ale postupem času dala vzniknout 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 meiózy endoduplikují – z diploidních, nebo triploidních oogonií se stanou tetra- nebo hexaploidní oogonie - a ty pak vstoupí do „normální" meiózy s rekombinací a segregací. Protože však k tvorbě bivalentů dojde jen mezi sesterskými chromosomy vzniklými endoduplikací, 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 i 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.

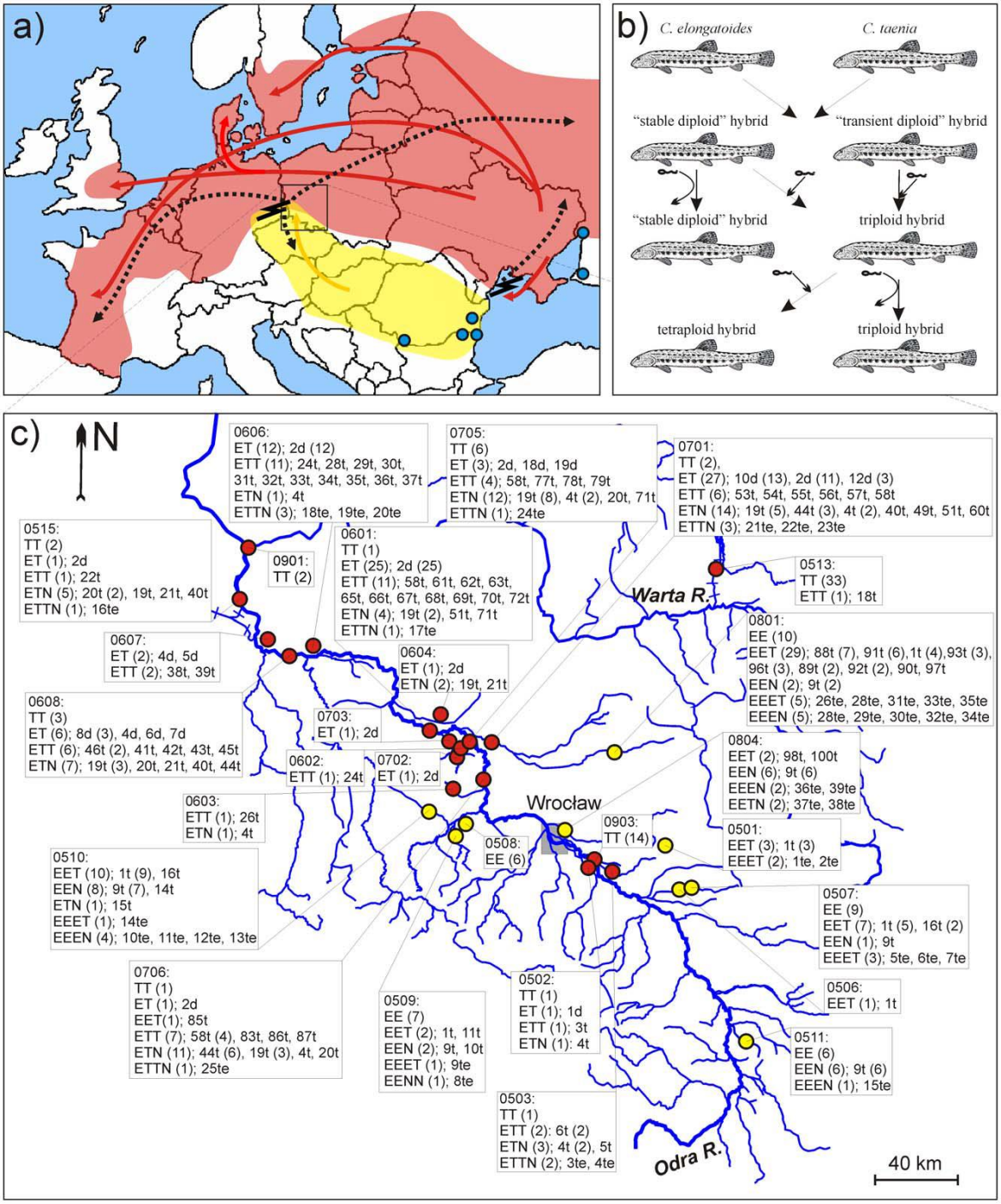


Schéma č. 1: Na obrázku a) je znázorněno geografické holocenní šíření druhů *C. elongatoides* a *C. taenia*. Na obrázku b) je schématicky vyobrazeno, jakým způsobem dochází k hybridizaci mezi druhy *C. elongatoides* , *C. taenia* a *C. elongatoides*. Schémc) je mapa středního toku Odry a jejích přítoků 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).

1.3 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). K polyploidizaci došlo v evoluci mnohokrát a často vedla k neofunkcionalizaci genů, rapidní změně komplexity organismu, jelikož jedna z genových kopií se stala redundantní a vedla k úniku před specializovanými parazity (Weiss-Schneeweiss et al., 2013).

Hybridizace je zejména, ale nejen u rostlin, velmi významný evoluční podnět. V případě kompatibility genomů, transkriptomů dochází k 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 nárůstu přínosné genetické variability nových kombinací alel (Birchler et al., 2010).

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

Zejména v případech triploidizace, jak se například děje vnesením třetího haplotypu do genomu diploidního hybrida při procesu gynogenetické reprodukce, mohou nastávat vážnější problémy, pomineme-li problémy meiózy,

V případě gynogenetického rozmnožování triploidních jedinců existuje více scénářů meiózy. Mají ale společné rysy, kdy se 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, tzn. ztrátu jejich koheze, rekombinace – narušením funkce kinetochoru a rekombinázy ad. (Qi et al., 2006). V anafázi I nesmí dojít k redukčnímu dělení.

1.4 Metody studia transkriptomu – diferenciální exprese a imprinting

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 přijatelnou chybovostí (illumina), kterou lze statisticky řešit. 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, po analýzu sestřihových variant, malých RNA, a další. Nepřeberné možnosti designů experimentů, díky vysoké reprodukcibilitě 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* délky okolo 50 bp porovnáním s výsledky *assembly* podle referenčního templátu genomu (Grabherr et al., 2011a). 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 mRNA (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, které zvyšují 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 komplexity RNA) a další (Wilhelm and Landry, 2009). Příklady zmíněných kontrol jsou mnohdy opomíjeny a mohou vést ke 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 využívají 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 matematických teorémů (Dilworth, 1950). Reprezentovatelnost sestřihových 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 znázorněn na Obr. č. 2.

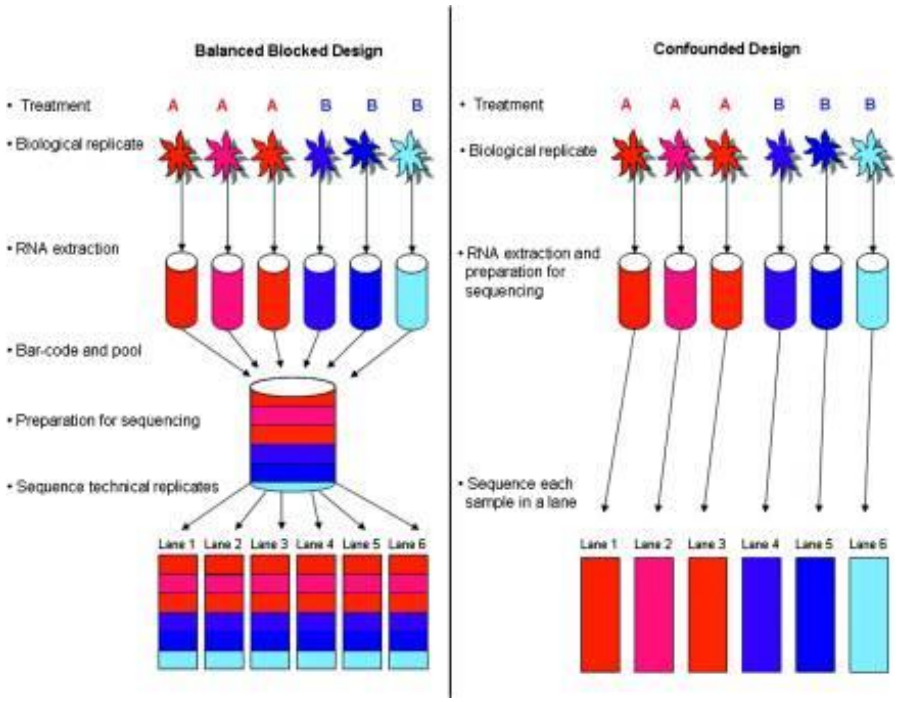


Schéma. č. 2: Znázornění používaných designů RNAseq experimentů; převzato z (Auer and Doerge, 2010).

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 M values*). 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; pochopitelně nejsou však vhodné pro data, kde naopak očekáváme velké rozdíly v expresi.

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 negativně binominální model, který byl získán aproximací poissonovského modelu. Pro využití statistické síly biologických replikátů byl navržen model společné disperze (Auer and Doerge, 2010).

Analýza disbalance paternálních alel vyžaduje poněkud delikátnější přístup díky celé řadě příčin. První z nich, která se naštěstí netýká našeho přístupu, je biologický zdroj variability způsobený sestřihovými variantami (Cho et al., 2014). Problematické jsou též haplotypy překlenuty jediným *readem*, obzvláště pak takové SNP, které se nacházejí v blízkosti inzerce/delece (indel), protože v těchto místech *aligmnet* obsahuje velké množství chyb (Li, 2011a). Klasickým statistickým přístupem k detekci alelické disbalance je binomiální, či chi-kvadrát test, kde u diploidního organismu testujeme odchylku od 0,5 poměru alel. Bohužel pro nízké, či hodně vysoké hodnoty exprese je P hodnota nadhodnocena. Značná technická variabilita mezi SNP pozicemi může vést ke shlukování spíše dle experimentálního designu nežli z biologické variability. V případě užití klasické frekvenční statistiky volíme spíše odstranění problematických regionů, pro minimalizaci chyb I řádu. Možným řešením je opět implementace vhodnějších modelů (např. poisson-gama) a bayesiánské statistiky (León-Novelo et al., 2014). To vše je možno kombinovat s iterativními maximum likelyhood metodami jako např. EM (*expectation–maximization*) algoritmus (Teare et al., 2006).

2 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, 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.

3 Materiál a metody

Rád bych čtenáře upozornil, že v metodách se mohou vyskytovat pasáže, které se zdánlivě podobají výsledkům. Jedná se o výsledky kontrol, které jsou nezbytnou součástí nejen bioinformatických analýz, bez nichž by nebylo možné se dopracovat výsledku - kvůli mnoha technickým problémům, lidským chybám i biologickým faktorům, které je třeba v každém kroku sledovat.

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: Vzorky 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 končící L, či o jsou vzorky normalizované cDNA specificky jedné tkáně; pokud toto označení chybí, jedná se o směsný vzorek normalizované 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

3.1 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.

3.1.2 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 mikrozkumavky, 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 peletu je přidáno 1 ml 75% etanolu (1 ml etanolu na 1ml Trizolu).
13) Rozpuštění RNA peletu 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 vyschlý pelet nelze snáze rozpustit)
16) RNA rozpuštěna v 10-20µl ddH2O (*RNase-free-water*) - dle velikosti pelety.
17) Roztok několikrát „propipetován".
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.

3.1.3 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. 4: 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 vlákna 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 ekvimolárního 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 single-end (SE).

3.1.4 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. 2. RNA integrita vztažena k 18S a 28S *peaku* byla testována Bioanalyzer 2100 (Agillent 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).

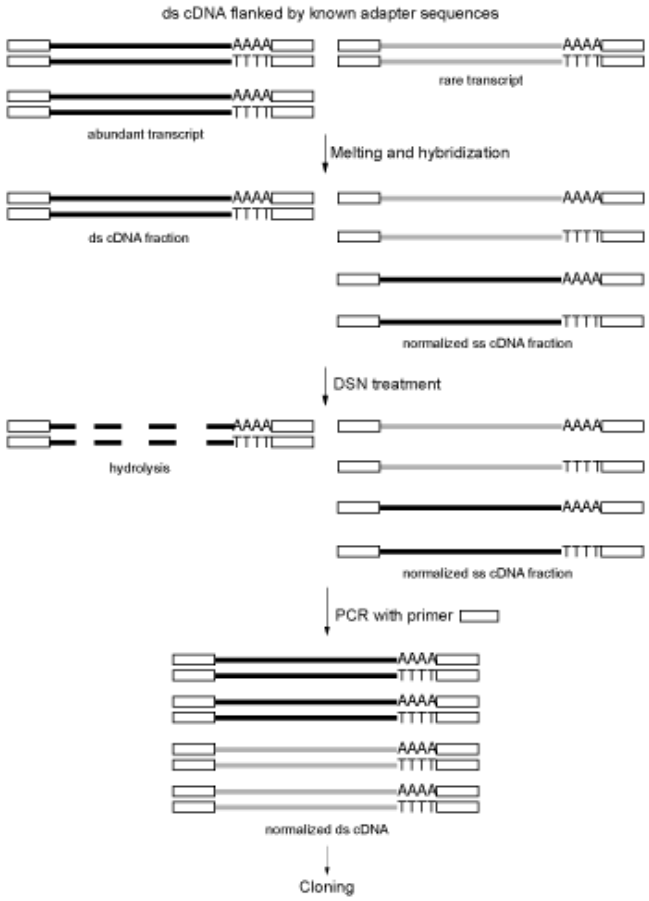


Schéma č. 2: Znázornění principu cDNA normalizace (zrdoj: http://www.evrogen.com/technologies/normalization.shtml)

3.2 Kompozice referenčního transkriptomu

Hlavním požadavkem na referenční sekvenci je obsáhnout maximální množství exprimovaných genů, jejich sestřihových variant a to nejvyšší možné délky. Z toho 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. 5) byly zbaveny sekvencí a konců s nízkou kvalitou pomocí Trimmomatic softwaru (Bolger et al., 2014). Technické PCR multiplikáty byly odstraněny z datasetu *readů* užitím cdhit-454softwaru (Li and Godzik, 2006). Výsledných 1886536 *readů* – 648620753 párů bazí bylo selektováno k tvorbě cDNA assembly – referenční sekvence (provedeno: Mgr. Jan Pačes, Ph.D.).

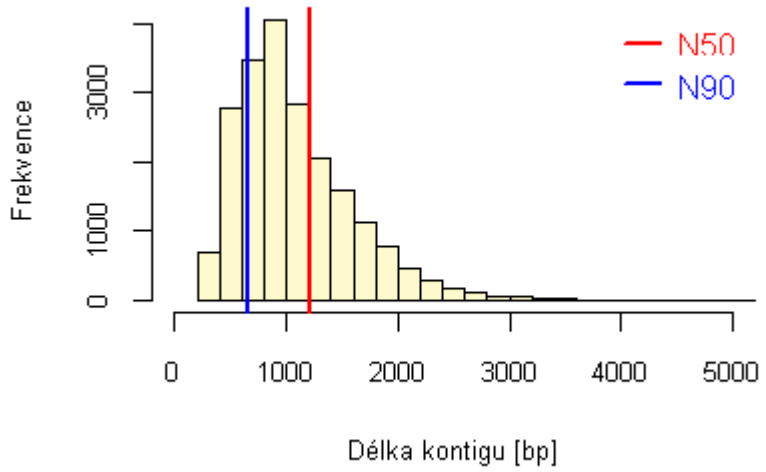
Jedinec	Typ tkáně	Knihovna	Počet bazí	Počet *readů*
co05	oocyty	co05	97492888	364919
co06	oocyty	co06	66110724	243092
cab04	játra	cab04L	103716091	285342
cab05	oocyty	cab05o	150231825	374684
	játra	cab05L	91235506	208142
cab06	játra	cab06L	54725340	154445
cab10	játra	cab10L	96836134	255912

Tab. č. 5: Tabulka jedinců tvořící referenční cDNA transkriptom se základní deskripcí; K referenci byl přiložen appedix zrekonstruovaný dle illumina sekvenačních dat

3.2.1 *Assembly* transkriptomu

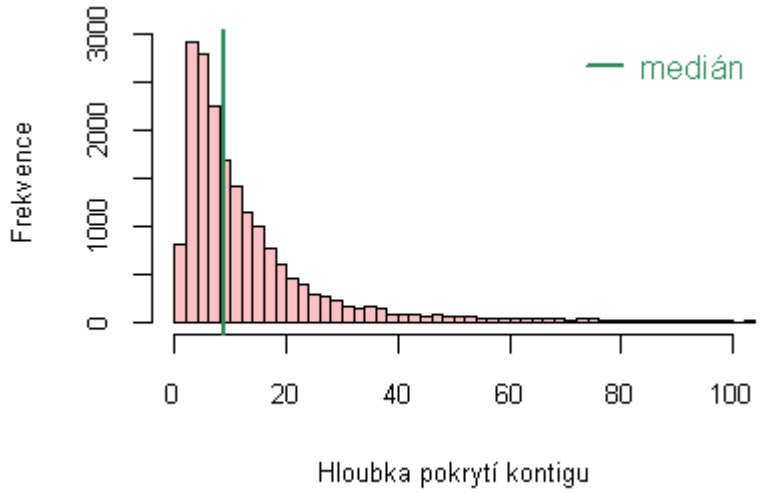
Vlastní *assembly* 454 dat bylo provedeno Newbler softwarem (verze: 2.6 20110517_1502, Roche) modifikovaným pro analýzu 454, tak illumia datových setů, s parametry: 20 bp minimální délka *readu*, minimální překryv *readů* 40 bp, minimální identita překrývajících se *readů* 90 %, minimální délka všech 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 bází 32540805 a N50 1 rovno 1291 viz tab. č. 5, přičemž 95.39 % sekvenovaných bází 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* 454 dat provedeno Mgr. Jan Pačes, Ph.D.).



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

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



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

Takto připravená *de novo assembly* ale nicméně nereflektuje poslední požadavek analýz – detekovat transkripty, které jsou exprimovány díky možné deregulaci transkripce zapříčiněné polypoidizací nebo hybridizací. Z tohoto důvodu byla po jednotlivých jedincích provedena programem Trinity v2.0.2 (Grabherr et al., 2011b) *de novo assembly* a to nejen hynridních jedinců, ale i čistých druhů. Na všech vygenerovaných transkriptech delších než 300 bp byla provedena shlukovací analýza – Markovo řetězové shlukování programem mcl (Enright et al., 2002) na základě hodnot *bitscore*, která reflektují váhu podobnosti mezi dvěma sekvencemi (minimální hodnota, kdy byly sekvence považovány za podobné, činí 80). Výstupem shlukování při nastavené granularitě rovné 2 je formát, kde je každý signifikantní shluk reprezentován názvy genů v řádku. Abychom mohli stanovit, zda jsou dané transkripty specifické například pro skupinu asexuálních jedinců, bylo nutné nejprve stanovit jejich počty ve shluku, čehož bylo dosaženo regexp knihovnou v programovacím jazyce python3.4, schéma č. 1

```
#!/usr/bin/python3.4
import re
cnt = 0
lom300tf = cab02 = []
# regexp hledání transkriptů z původní referenční sekvence a nově
složených jedinců z oocytů a jater
p0 = re.compile("lom300tf_[ics][0-9]+")
p1 = re.compile("cab02_TR[0-9a-zA-Z_]+")
soubor = open('clusters.txt', encoding='utf-8')
# nalezení všech transkriptů, podle definovaného tvaru, zápis jejich
jména a počtu do nového listu (pro ilustraci neuvedeny všichni jedinci)
for i in soubor:
    cnt += 1
    m0 = re.findall(p0,i)
```

```
    m0_cnt = len(re.findall(p0,i))
    m1 = re.findall(p1,i)
    m1_cnt = len(re.findall(p1,i))
    if m0:
        lom300tf.insert(cnt,m0_cnt)
    else:
        lom300tf.insert(cnt,0)
    gen_F = list(set(gen))
    zapis = open("clst%s.fas" % cnt, mode = "w")
        SeqIO.write(gen_F,zapis,"fasta")
```

Schéma č. 1: Znázornění hledání *kontiguů* konkrétních jedinců z výstupu shukovací analázy

Nyní, kdy pro každého jedince máme počet transkriptů ve shluku podobných sekvencí, můžeme se v databázi dotázat, zda byly přítomny pouze transkripty asexuálních, či sexuálních jedinců, a vybrat posléze nejdelší transkript - nejdelší sestřihovou variantu ve shluku. Toho bylo dosaženo níže uvedeným přístupem v jazyce python, schéma č. 2.

```
# Výběr nejdelšího kontigu ze shluku je proveden aplikaci knihovny Bio
files = glob.glob("*.fas")
for i in files:
    maximum = 0
    handle = open(i, "rU")
    for ii in SeqIO.parse(handle, "fasta"):
        l = len(ii.seq)
        if l > maximum:
            maximum = l
            to_handle = ii
            jmeno = ii.id
    zapis1 = open("%s.fasta" % jmeno, mode="w")
    SeqIO.write(to_handle, zapis1, "fasta")
zapis1.close()
# Výběr hybrid specifických kontiguů
cur.execute("select transcripts from animals_lom_mcl_raw a
                join animals_lom_mcl b using (cluster)
             where (
                b.lom300tf + b.cab04 + b.cab05 + b.cab06 + b.cab10 +
                b.cab18 + b.cab19 + b.cab20 + b.cab21) = 0 and (b.cab02
                + b.cab03 + b.cab07 + b.cab08 + b.cab09 + b.cab11 +
                b.cab13 + b.cab15 + b.cab16 + b.cab17 + b.cab22 +
                b.cab23 + b.cab24 + b.cab25) >= 1"
            )
asex = cur.fetchall()
for y in glob.glob('*.fasta'):
    fasta2 = SeqIO.index(y, "fasta")
    for yy in fasta2:
        for yyy in asex:
            if fasta2[yy].id in str(yyy):
                zapis2 = open("hybrid/%s.fasta" % fasta2[yy].id,
                              mode="w")
                SeqIO.write(fasta2[yy], zapis2, "fasta")
```

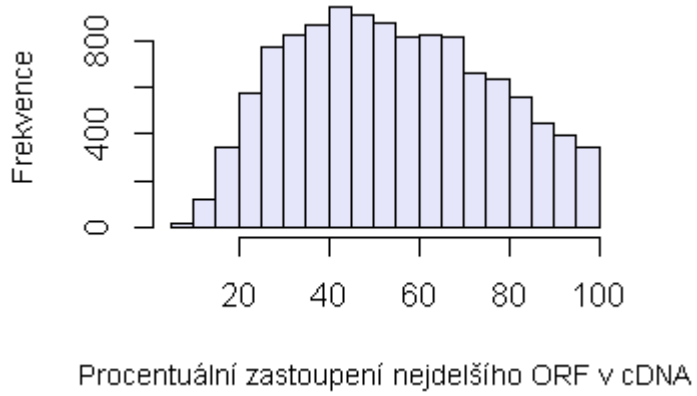
Schéma č. 2: Popis výběru nejdelší fasta sekvence z definovaných skupin s testem toho, zda je podmnožinou asexuálních *kontiguů*.

3.2.2 Dodatečné korekce transkriptomu

Reference musela být v několika postupných krocích vyčištěna, neboť byly artificiálně vygenerovány duplikáty cDNA. 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 geny vzniklých spojením skrze sekvenčně parciálně homologické sekvence. Díky těmto krokům byly postupně získáno 5 verzí referenční transkriptomů. Basální popis změn v transkriptomu je uveden v tab. č. 6.

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 bází, 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 *kontiguů* odstraněno, neboť po separaci nedosáhly limitu 300 bp, rovněž bylo odstraněno 698 původních chimérických *kontiguů*.

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. U extra *kontiguů* z illumina dat nedošlo ke změně orientace, neboť nebyly použity k analýze Müllerovy rohatky; tyto sekvence obsahují větší procento chyb a přes 550 z nich má signifikantní shodu s *transpozabilními* elementy.
7) Finální verze byla dodatečně podrobena shlukovací analýze mcl, viz výše a zredukována na 18258 transkripů, abychom s určitostí vyloučili problémy s mapováním sekvenčně velmi podobných transkriptů. K těmto transkriptům byl připojen dodatek transkriptů z *de novo assembly* rna-seq, který se alespoň v jedné kopii vyskytoval pouze u asexuálních jedinců, nebo nebyl přítomen v původní referenční sekvenci a vyskytoval se pouze u sexuálních jidinců. Finalní počet transkripů tedy činí **21508**.
8) 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; oligomery morfolino k inhibici translace ad.)

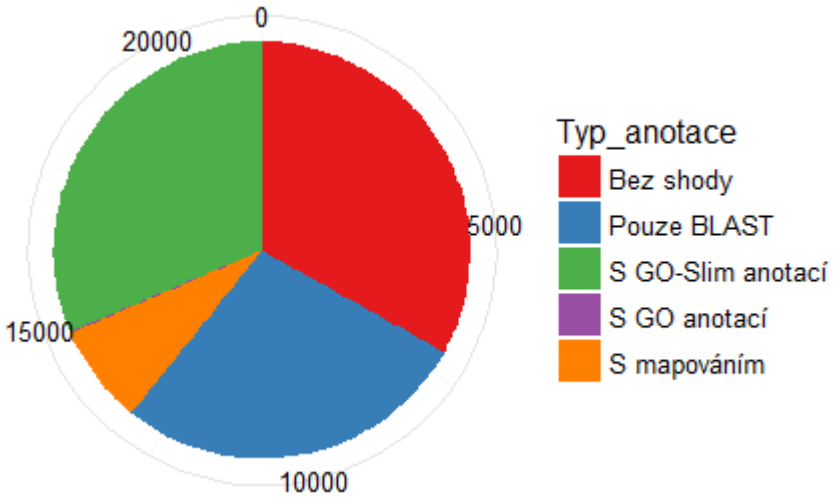


Graf č. 3: Histogram distribuce procentuálního zastoupení nejdelšího ORF v cDNA sekvenci s mediánem 52,96.

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
Lom300_ex	21508	1056.13	22714201	1215	612

Tab. č. 6: Deskriptivní statistika jednotlivých setů referenčních sekvencí

3.2.3 Anotace 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*)

Přibližně třetina cDNA sekvencí nebyla anotována, viz graf č. 4. Primární důvod tak vysokého počtu genů bez anotace je zapříčiněn tím, 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).

3.3 Mapování RNAseq sekvencí

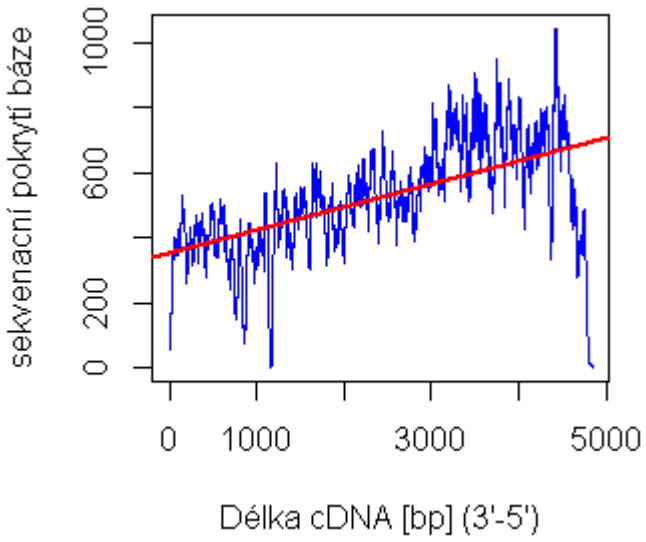
K mapování SE 50 bp *readů* RNAseq dat (HighSeq 2000) bylo přistoupeno k několika rozdílným algoritmickým přístupům: BWA (verze 0.7.12-r1039) (*Burrow wheel alignment*) (Li and Durbin, 2009), Bowtie2 (Degner et al., 2009), novoalign (V3.01.02) (Ruffalo et al., 2012) a Mosaik (verze 2014-03-26) (Lee et al., 2014b). Mosaik se jeví jako vhodný nástroj k detekci SNP vzhledem k preciznosti mapování, tedy nakonec i k analýze diferenciální exprese. Programy byly spuštěny se základními parametry, v případě programu Mosaik byly povoleny 4 *missmatch* báze a délka *hash* byla nastavena na 15 bp. Pro analýzu SNP byl zvolen pro svou nejvyšší přesnost Mosaik. V tabulce č. 7 je vyjádřen rozdíl v mapování hybrida biotypu *ett* a druhu *tt* pro srovnání výsledků *mapping* programů určených pro *mapování* dat RNAseq na transkriptom.

	tt				***eet***		
mapper	**Pokrytí**	**std Pok.**	**map. [%]**	**mapper**	**Pokrytí**	**Pok. std**	**map. [%]**
bwa	22.54	200.37	44.00	**bwa**	36.09	258.718	43.92
bowtie2	22.65	200.41	44.91	**bowtie2**	36.2958	258.565	44.89
novoalign	22.77	201.36	47.14	**novoalign**	8.46986	91.7364	48.18
mosaik	22.9355	201.904	46.26	**mosaik**	37.2447	263.233	46.92

Tab. č. 7: Srovnání Procenta *namapovaných readů* a pokrytí se standartní odchylkou, při základním nastavení, vůči finální referenci transkriptom *C. taenia*. Srovnání divergence

Ready byly zbaveny již adaptérových sekvencí (EMBL, genecore servisní centrum), nicméně nebyly odstraňovány nekvlitní počáteční, či koncové báze, protože byly poměrně vysoké kvality a pokles kvality bazí byl zanedbatelný viz graf. č. 6.

Díky nízké hladině exprese značného počtu mRNA a nedokonalosti reverzní transkripce (absence pozic směrem k 3' konci cDNA viz kapitola 4.4.2, graf. č. 5) došlo k složení pouze 3' UTR a pouze krátké časti kódující sekvence, nicméně „vysoký" počet ne*namapovaných* sekvencí souvisí také s kvalitou *readů* a výskytem SNP v proximální blízkosti.

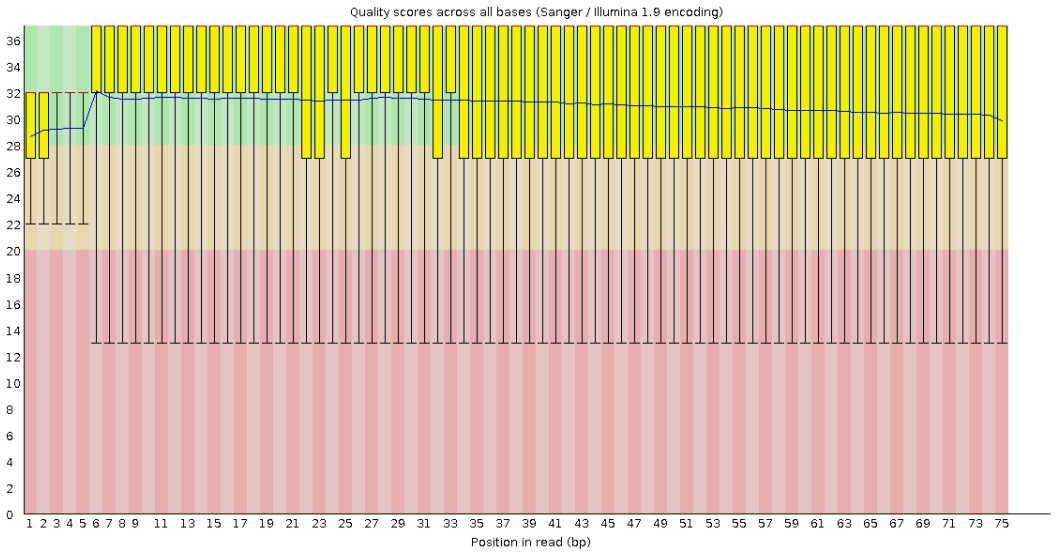


Graf č. 5: Sekvenační hloubka jednotlivých bazí napříč cDNA kódující gen komplementu C4-2 s délkou 4843 bp; (vzorek nemusí být reprezentativní pro celý dataset; podobný trend lze ale pozorovat i u mnoha jiných náhodně vybraných cDNA). Červená přimka je znázorněním lineárního modelu bez konfidenčního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

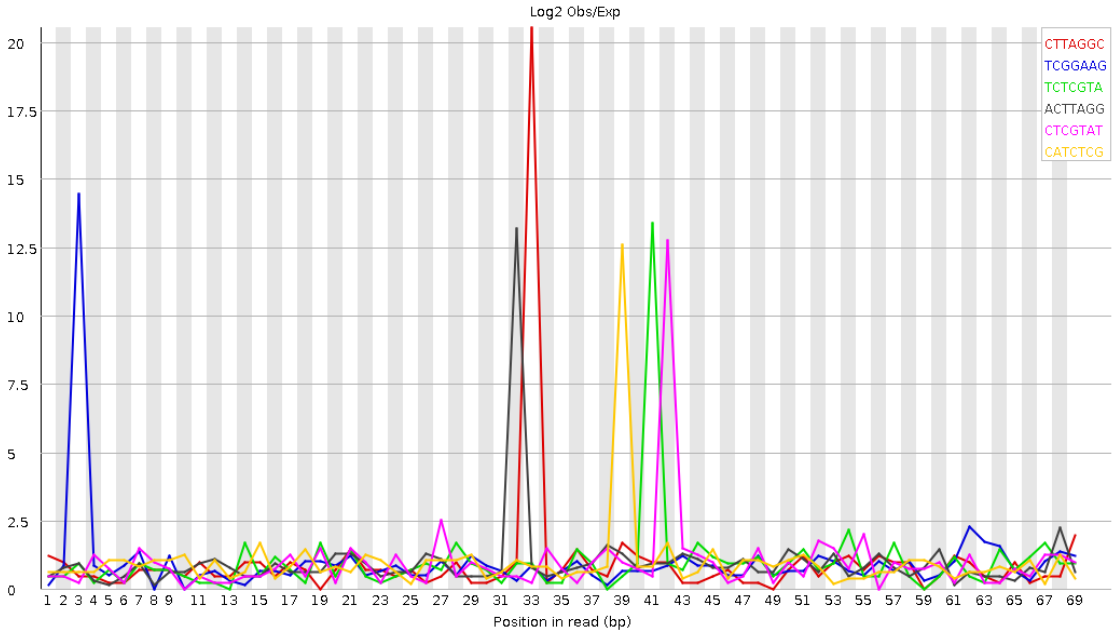
3.4 Kontrola kvality sekvenačních dat

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 bází při začátcích syntézy *readů*. Naštěstí takovýto *bias* neafektuje výsledek

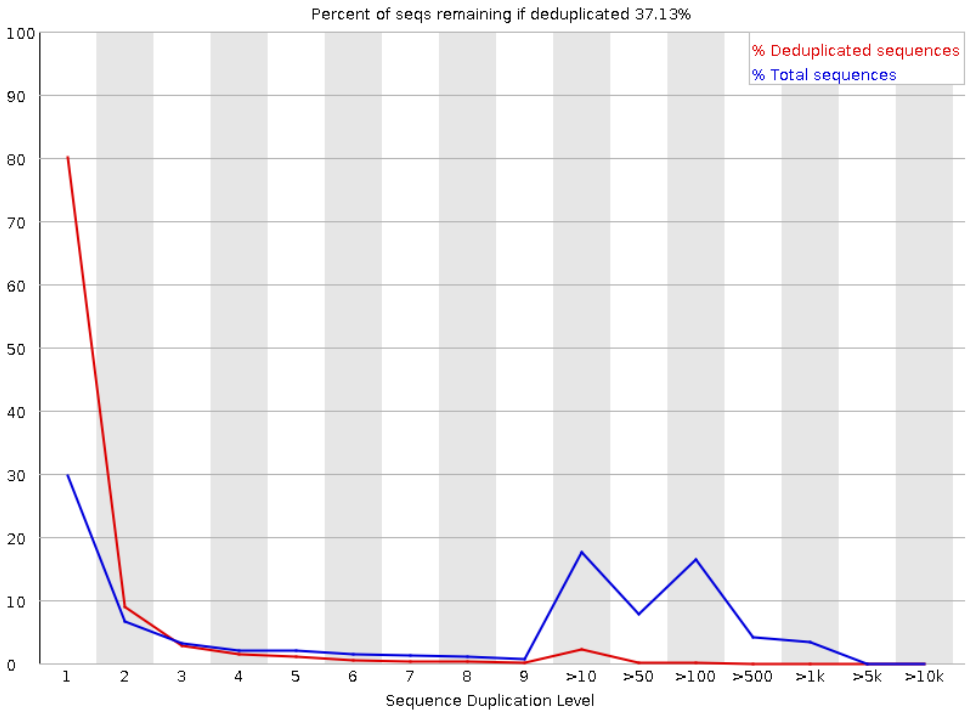
expresí; za příčinou *biasu* poziční kompozice bazí může stát nenáhodná fragmentace (tepelným šokem v případě našich dat), selekcí *randomizačních* primerů (nebyly aplikovány) nebo chybným, příliš horlivým zásahem bioinformatika při odstranění sekvencí adaptéru fragmentu.



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)



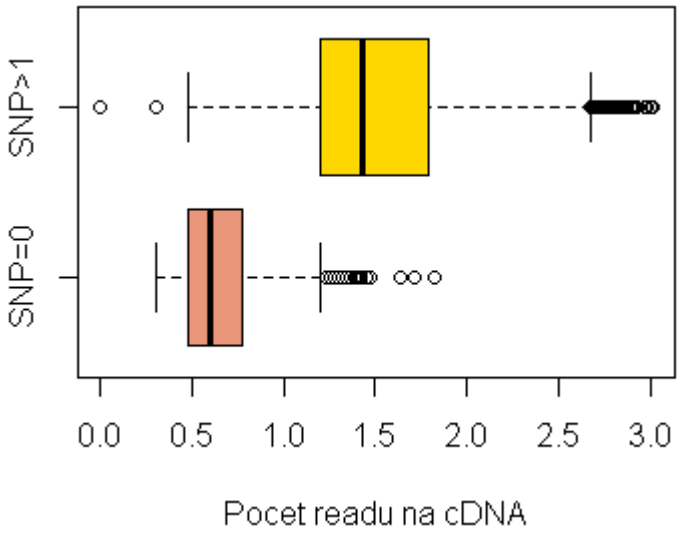
Graf č. 8: Znázornění míry duplikovaných sekvencí v rámci souborů *readů* RNAseq vzorku jater (počet neunikátních sekvencí by neměl překročit 20 %)

3.5 Analýza polymorfismů 454 normalizovaných RNAseq dat

Sekvenační data jsme mapovali BWA softwarem (na základní nstavení), protože je vhodný i pro 454 data (ta byla převedena ze standartního formátu sff do fastq softwarem sff2fastq v. 0.9.2). SNP byly získány cestou samtools (v. 1.2) / bcftools (call, v. 1.2-151-g7357020) / vcfutils (VarFilter v. 0.9) (Li, 2011b). SNP, které nesplňovaly následující podmínky byly odstraněny: sekvenační hloubka DP >= 10, kvalita mapování MQ >= 20 a pozice, které se vyskytovaly 3 bp v blízkosti inzerce či delece (indel). Sada těchto SNP spolu se SNP illumina sekvenování posloužila k vytvoření konsenzuální sekvence, viz kapitola č. 3.7. Programem BWA byla data *remapována* na refrenci se zamaskovanými polymorfismy a SNP byly opět získány softwarem bcftools call s parametrem -m (*multiallelic caller*), protože je schopen řešit i disbalanci v expresi a vzácené alelické varianty. *Remapování* je nezbytným krokem pro zpřesnění SNP datasetu, zejména pak pro řešení heterozygotních pozic i původně invariantních pozic (vzhledem k sekvenci *C. taenia*). Mosaik je také možno použít na 454 data, v tomto případě ale preferujeme rychlost BWA, protože mimo maskování polymorfismů v referenci a kontroly pro nás SNP 454 dat nemají další užitek.

Před touto analýzou byly data navíc mapovány modifikovaným programem Newbler – GSmapper; na základě *remapování* 454 dat byly odstraněny tzv. pseudoparalogní sekvence (objevovaly se velké nových SNP) (Mgr. Jan Pačes, Ph.D., 2014).

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 zřetelné, že hloubka sekvenování 454 normalizované RNA nebyla dostatečná pro identifikaci velké části prezentovaných polymorfismů daných jedinců. Nízká sekvenační hloubka je také spojena se zkrácením 3' konců rererenční sekvence viz kapitola č. I. Přes tyto problémy bylo identifikováno 394072 SNP.

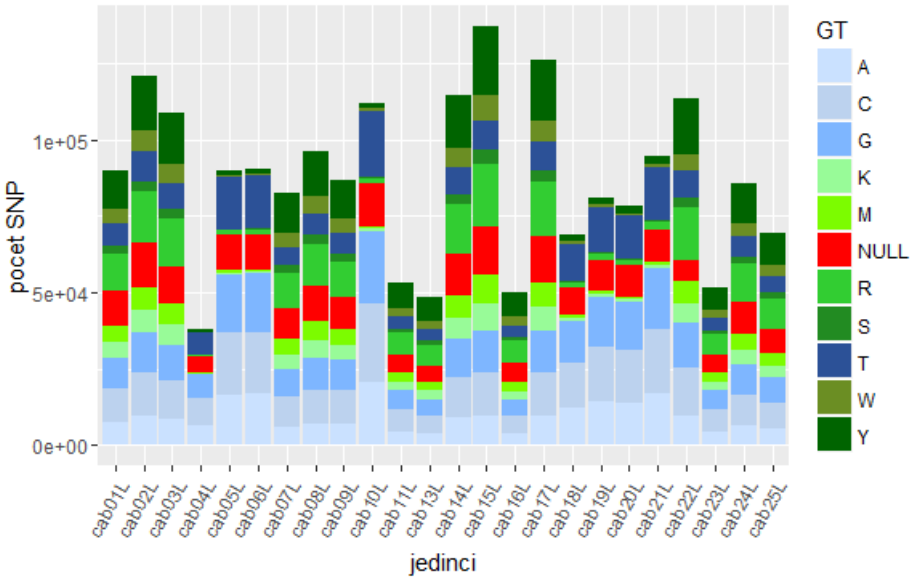


Graf č. 9: Srovnání distribuce počtu *readů* na cDNA mezi datasety: počet *readů* na *kontigu* bez nalezených SNP a počet *readů* na *kontigu* obsahující více než 1 SNP. Data jsou transformovány log10+1.

3.6 Kontrola SNP RNAseq v pozicích identifikovaných 454 daty

Při kontrole dat vycházejme z hypotézy, že hybridi museli vzniknout následkem mezidruhového křížení, a tedy i jejich heterozygotnost by měla být mnohonásobně vyšší. U starších klonálních linií by ale měla heterozygotnost klesat, pokud není ovlivněna pozitivní selekcí. Heterozygotností míníme pouze poměr homozygotních a heterozygotních alel, nikoliv heterozygozitu vycházejícího z Hardy-Weinbergovy rovnováhy. V grafu č. 10 znázorňujeme poměry homozygotních / heterozygotních alel u jedinců všech analyzovaných biotypů včetně rodičovských druhů; graf č. 10 znázorňuje průměrné hodnoty pro veškeré jedince daných biotypů. Z grafů je patrné, že hodnoty

heterozygotnosti se mezi hybridy a rodičovskými druhy značně liší; můžeme tedy říci, že i zvolený přístup je schopen detekce velké částí heterozygotích pozic v transkriptomu.



Graf č. 10: Zastoupení bází v datasetu SNP všech analyzovaných jedinců jaterních vzorků; biotypy jsou popsány v tabulce č.2,3; Odstíny zelené označují heterozygotní báze, odstíny modré naopak homozygotní, červená frakce (*NULL*) značí pozice, které nedosáhly *trashold* sekvenační hloubky a kvality. Genotypy (GT) jsou uvedeny pod IUPAC zkratkou.

3.7 Sestavení „konsenzuální“ reference pro analýzu diferenciální exprese

Polymorfismy v referenci je nutno maskovat, protože bychom v opačném případě dosáhli nerovnoměrného *namapování* jednotlivých druhů, a to pochopitelně dle fylogenetické distance mezi nimi. Namísto všech ambiguídních bází byly do reference vneseny pouze pozice „N“, které reprezentují pozici, kde se byť jediný vzorek lišil od reference druhu *C. taenia*. SNP pozice byly získány aplikací postupu samtools/bcftools/vcftools (Li, 2011c). VCF (*variant standart format*) byl převeden vcftools do tabulkového formátu a zpracováván opět v databázi MySQL; na každé variantní pozici kteréhokoliv druhu byla jednoduchým python skriptem přepsána tato báze v referenci na “N”. Nejpodstatnější je, že byla použita nejen 454 normalizovaná data, která tvořila téměř polovinu všech SNP (2555623 – podmnožina druhů (*tt,ee,nn*)), ale i čtení illumina rnaseq dat a to opět pouze jedinců, kteří byli analyzováni (vzdálené druhy jako *C. strumicae* nebyly do “konsenzuální” reference zahrnuty). Do reference bylo zaneseno celkem 506648 SNP – „N“ SNP v místech polymorfismů.

3.8 Identifikace a validace druhově specifických SNP

Z MySQL transformace souhrnné tabulky SNP jedinců do binárního formátu je patrné (schéma č. 7), 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;
```

Schéma č. 3: 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 52202 takových SNP z celkového počtu 506648. Nicméně, získané druhově specifické SNP je nutno validovat nejlépe genomickými daty, neboť geneze těchto dat je podepřena pouze čtyřmi vzorky pro druh *ee* a 8 vzorky pro druh *tt*.

3.9 Detekce SNP z RNAseq illumina dat vůči transkriptomu se zamaskovanými SNP

Illumina data byla *remapována* na „N“ referenci softwarem Mosaik pro maximalizaci přesnosti; délka hash 15 bp, 4 povolnné SNP na *read*. SNP byly učeny opět samtools/bcftools/vcfutils s téže parametry. Oproti rychlému přístupu pro detekci polymorfismů vůči nemodifikované referenci byl ale přidán krok navíc. Byly provedena deplece duplikovaných *readů* (dle sekvence a pozice v alignmnetu), a to programem picard (v. 1.140) (McKenna et al., 2010); odstranění optických duplikátů je zásadní zejména pro ASE, na statistiku DE nemá zásadní vliv. V posledním kroku byly z VCF verze 4.1 (*variant call format*) (Danecek et al., 2011) vyextrahovány potřebné informace: název cDNA, pozice SNP, sekvenační hloubka na pozici, alternativní alely oproti referenci a kvalita mapování MQ. Nové detekované polymorfismy nebyly dále užity.

3.10 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. 2 je uveden postup získání DE genů balíkem DESeq/edgeR a konsekventně i nabohacených GO termínů těchto genů.

3.10.1 Extrakce počtu *namapovaných readů* na cDNA RNAseq vzorků

Pro získání informace, kolik je *namapovaných readů* s kvalitou mapování >= 20 a navíc pouze těch, které byly namapovány unikátně na jeden "gen" – *kontigu*, bylo užito software bedtools multicov v2.25.0 (Quinlan and Hall, 2010). U ~550 genů byl rozsah, ve kterém se počítaly počty *readů* na cDNA zkrácen na rozsah nejlepšího blast *hitu* vůči proteinové databázi transpozabilních element, viz kapitola č. 3.10.1.

3.10.2 Identifikace diferenciálně exprimovaných genů

V prvním koku jsou získané počty *readů* na cDNA normalizovány, jelikož 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 kratších *readů* při klastrování fragmentů na destičce illumina platformy (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 neexistuje velké množství DE genů. Bohužel tato premisa v našem případě rozhodně neplatí u srovnání mezidruhového, kde očekáváme velké rozdíly – volíme zde metodu *upperquartile*, kde je celkový počet *readů* na transkript dělen celkovým množstvím *readů* v knihovně, to celé vynásobeno hodnotou 75 % kvantilu řádku/transkriptu (Bullard et al., 2010).

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. 4. Fyzický příklad postupu získání DE genů mezi skupinami asexuálně a sexuálně se rozmnožujícími jedinci je uveden s popisy ve schématu č. 5. Ve schématu č. 6 je uveden postup, jakým bylo provedeno kontrolní znázornění *heatplot* a shlukovací analýza genů uvedených ve výsledcích sekce diferenciální genové exprese.

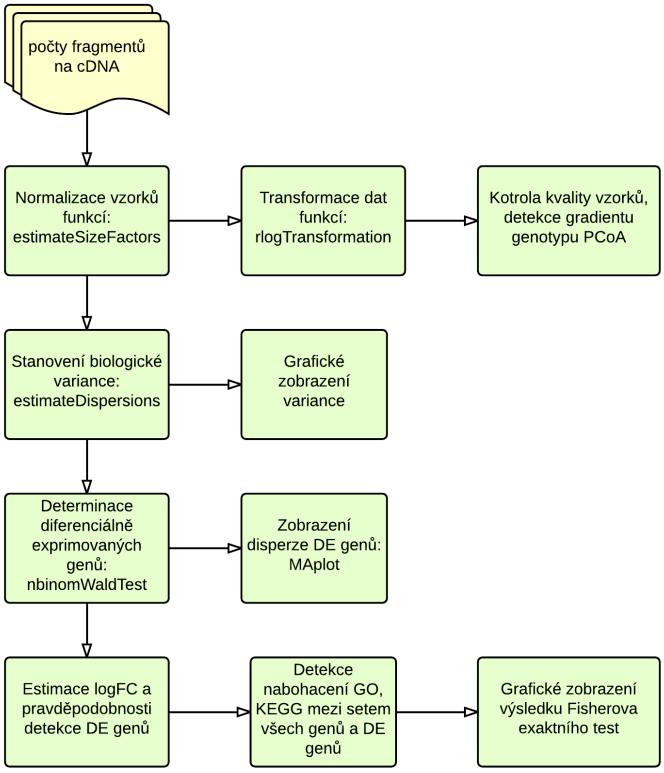


Schéma č. 4: Znárornění postupu získání DE genů

```
library(sqldf)
library(edgeR)
# Odstranění řádků se sumou readů menší než 20 (zcela neexprimované geny)
cnt_o <- dat_o[rowSums(dat_o)>20,]
# Odstranění jedinců, kteří nejsou přítomni v obou datasetech, nemají
# dostatečnou sekvenační hloubku, či jde o nevhodné vzorky (biotyp)
```

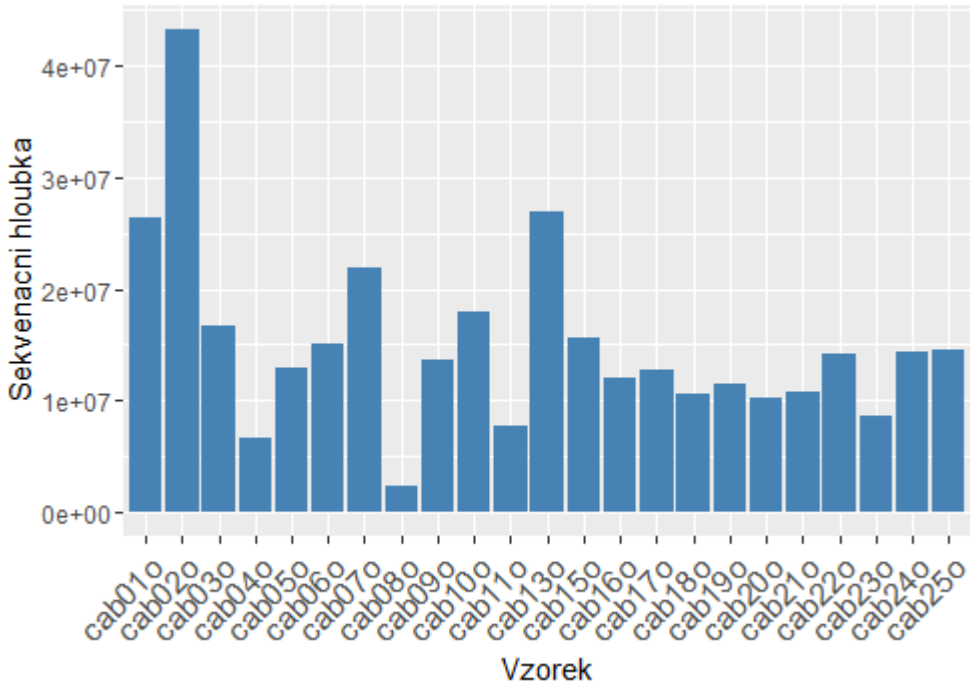
```
in_o_sex_noN <- sqldf("select * from animals_o where name != 'cab23o' and
name != 'cab08o' and biotype != 'etn' and biotype != 'een' and biotype !=
'enn'")
dat_o_sa_noN <- cnt_o[, names(cnt_o) %in% in_o_sex_noN$name]
cds_o_sa_noN <- DGEList(dat_o_sa_noN, group = in_o_sex$sex)
# normalizace dat metodou TMM (trimmed mean median)
cds_o_sa_noN <- calcNormFactors(cds_o_sa_noN)
# Stanovení disperse v rámci skupiny
cds_o_sa_noN <- estimateCommonDisp(cds_o_sa_noN, verbose=TRUE)
# Stanovení disperse mezi skupinami
cds_o_sa_noN <- estimateTagwiseDisp(cds_o_sa_noN)
# Výběr skupin ke komparaci
et_o_sa_noN <- exactTest(cds_o_sa_noN, pair=c("asex","sex"))
top_o_sa <- topTags(et_o_sa_noN, n=nrow(cds_o_sa_noN$counts))$table
# Selekce genů s P hodnotou < 0.05
de_edgeR_o_sa_noN <- top_o_sa[top_o_sa$FDR<0.05]
de_edgeR_o_sa_noN$ref_acc <- rownames(de_edgeR_o_sa_noN)
# Spojení výsledků s náležitou anotací
de_edgeR_o_sa_ann <- sqldf("select ref_acc, logFC, logCPM, PValue, FDR,
annotation from de_edgeR_o_sa_noN 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_noN <- sqldf("select ref_acc, logFC, logCPM, PValue, FDR,
annotation from de_edgeR_o_sa_noN join annotation using (ref_acc) where
annotation != 'NA' order by abs(logFC) desc limit 50")
```

Schéma č. 5: Deskripce fyzického postupu při získání DE genů metodou edgeR

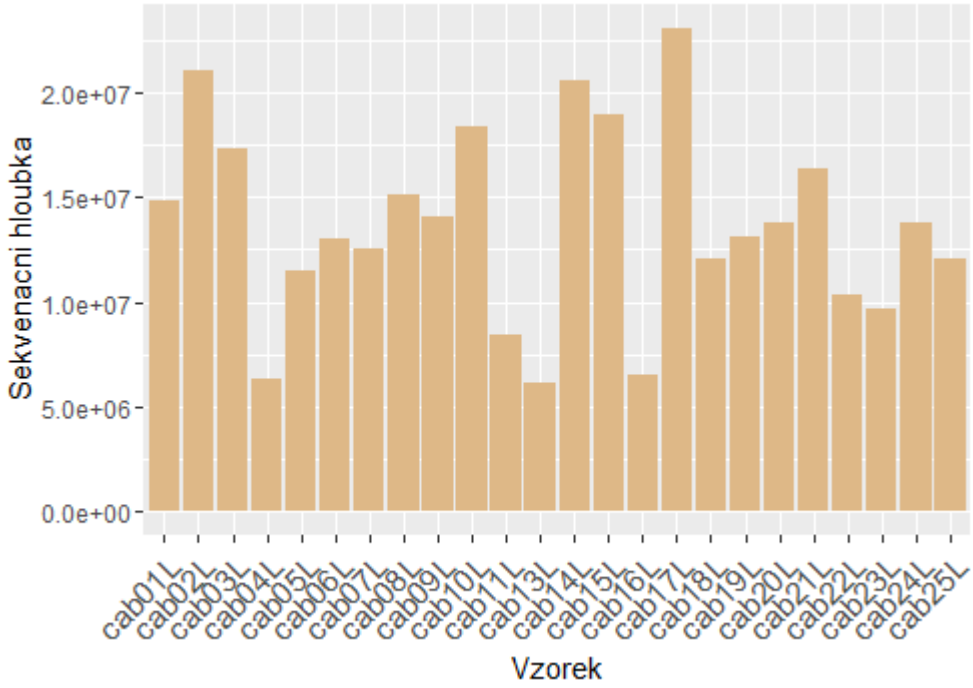
```
# Získání dat pro heatplot2 a shlukovací analýzu spojením dle klíče DE
# genů s TMM normalizovanými daty napříč použitými vzorky
cluster <- merge(cds_o_spec$pseudo.count,de_edgeR_o_spec,by="row.names")
cluster <- cluster[, names(cluster) %in% in_o_spec$name]
# Vygenerování heatplot grafu
library(RColorBrewer)
cols <- colorRampPalette(brewer.pal(10, "RdBu"))(256)
mydatascale <- t(scale(t(data.matrix(cluster))))
# 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$height)/1.5)
mycolhc <- sample(rainbow(256))
mycolhc <- mycolhc[as.vector(mycl)]
# Tvorba heatmap diagramu z dat normalizovaných vzroků použitých k
# analýze DE genů – příklad komparace skupiny "sex" vs "asex" (20 vzorků
# oocytů)
heatmap.2(data.matrix(cluster), Rowv = as.dendrogram(hr), Colv =
as.dendrogram(hc), scale="row", RowSideColors=mycolhc, labRow = '',
cexCol = 1.5,col= cols,labCol = in_o_sex$biotype,trace = "none",keysize =
2,key.title = "")
```

Schéma č. 6: Deskripce aplikovaných statistických metod programu R vyobrazených ve výsledcích týkajícín se diferenciální genové exprese.

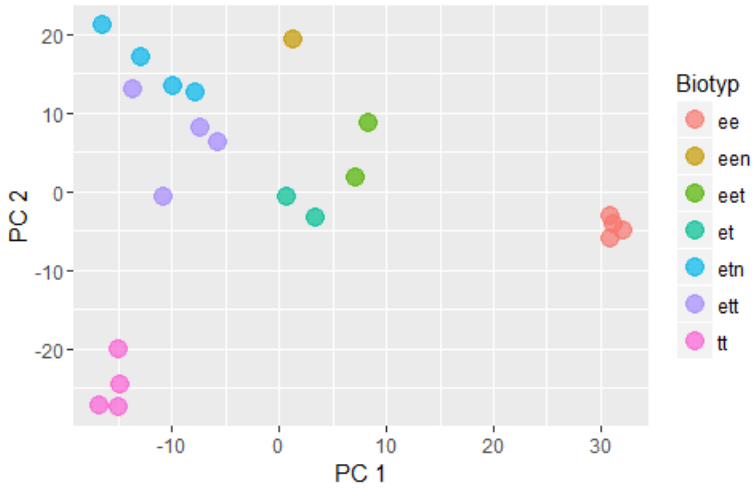
K analýze DE byli vyloučeni ti jedinci, kteří se vyznačovali neúměrně nízkou sekvenační hloubkou vzhledem k sumám počtu *readů* na cDNA ostatních jedinců viz graf č. 11, protože se projevil i *bias* v datech kdy na PCA znázornění. Jedinec této distance na ose PC1 (*principal component* 1) deformuje samotné znázornění převedení mnohorozměrného prostoru do 2D projekce. Byly ponechány vzorky, které byly získány z jater i 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. 13, 14. Jedná se o PCA pouze několika set genů s největší biologickou variabilitou mezi vzorky - 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, naopak u ooctů očekáváme konervativní expresi. Pro znázornění PCA distancí byly vybrány dvě majoritní komponenty vysvětlující nejvíce variability mezi vzorky.



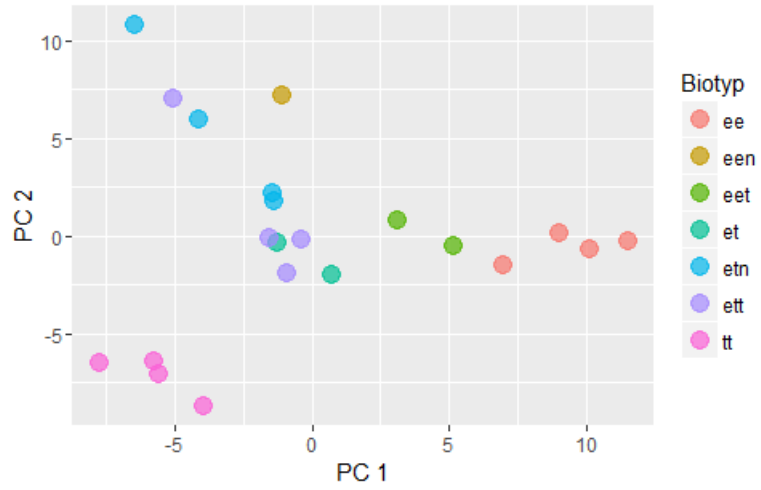
Graf. č. 11: Sloupcový diagram znázorující sekvenační hloubky vzorků oocytů; vzorek s nejnižší sekvenační hloubkou byl z datasetu odstraněn (cab08o).



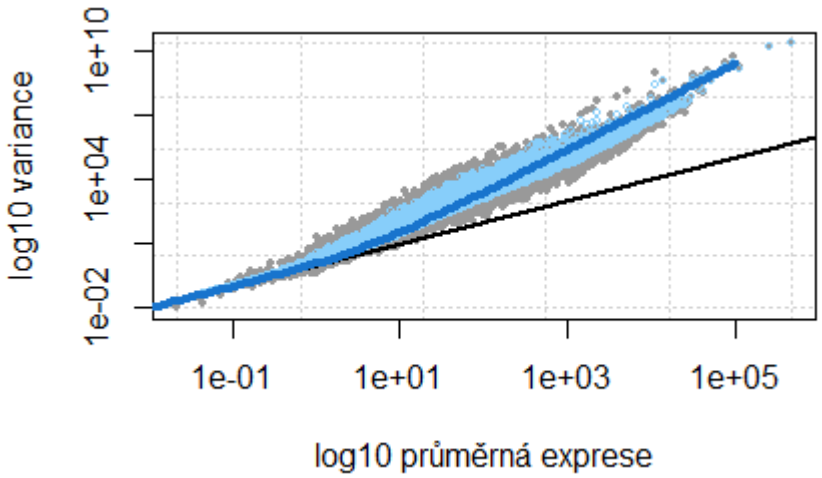
Graf. č. 12: Sloupcový diagram znázorující sekvenační hloubky vzorků oocytů a jater; vzorek označený šipkou byl z datasetu odstraněn pro nedostatečnou sekvenační hloubku



Graf č. 13: PCA plot znázorňující vzájemnou podobnost mezi vzorky **oocytů** při vybrání dvou hlavních komponent. PC 1 vysvětluje: 28 %; PC 2 vysvětluje: 20 % variability



Graf č. 14: PCA plot znázorňující vzájemnou *similaritu* mezi vzorky **jater** při vybrání dvou hlavních komponent. PC 1 vysvětluje: 22 %; PC 2 vysvětluje: 16 % variability



Graf č. 15: 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 detekovaný nové SNPs, byly odstraněny. Vzhledem k metodě získání cDNA, jsou častěji „prosekvenovány“ oblasti 5’ konce cDNA, neboť reverzní transkripce byla provedena aplikací primeru polyT, přepisujíce sekvenci od polyA konce mRNA. Ačkoliv je užitá reverzní transkriptáza schopna přepisu až 20 kb, často od templátu disociuje, a to s patrně lineárně, jak vyplývá z grafu č. 5. Tímto způsobem je do

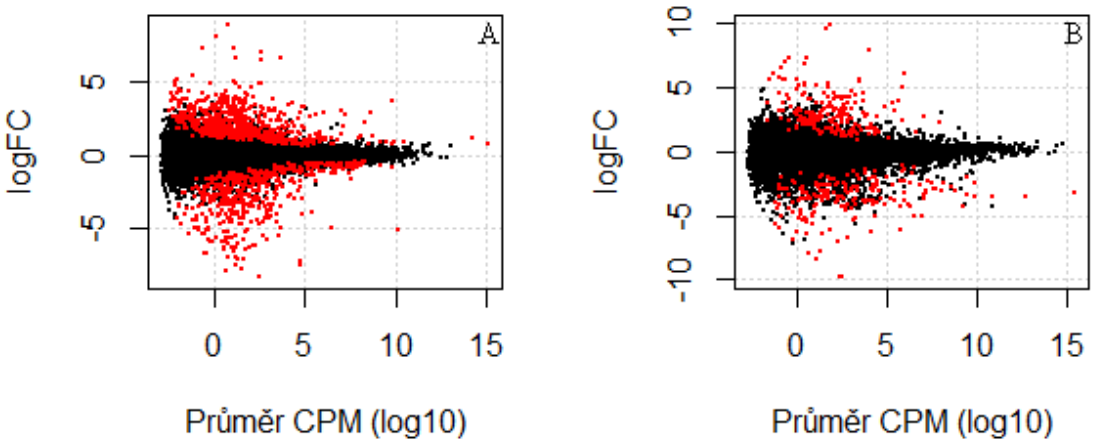
výsledků vzhledem k nadměrně dlouhým sekvencí vnášen další *bias*, jelikož tyto dlouhé cDNA sekvence se mohou jevit jako podexprimované. Naštěstí datasety RNAseq byly získány aplikací identického protokolu včetně téže reverzní transkriptázy, a proto tato chyba nehraje v komparaci dat s cílem identifikace diferenciálně exprimovaných genu žádnou roli (*procesivita* reverzní transkriptázy by měla být sekvenčně nezávislá).

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* byla 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. 7 (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. 7: 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 kontrolní body postupu získání DE genů a samotná vstupní data. Posledním kontrolním bodem je vlastní explorace distibuce výsledků DE genů. Graf č. 16 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 jater a oocytů.



Graf č. 16: 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 stanovené na hladině α 0.05 FDR korigované P hodnoty.

3.10.3 Identifikace transpozabilních elementů

Pro označení transpozabilních elementů (TE) jsme zvolili sekvenční srovnání metodou blastx vůči proteinové databázi TE – repbase (verze 20.05). Nalezené sekvence byly filtrovány způsobem, kdy byl vybrán nejlepší *hit* pro daný *kontigu* s *bitscore* větším než 120. Abychom se vyhli problémům s chimérickými sekvencemi, tázali jsme se specificky na počet *readů* v region *kontigu* s největší shodou vůči TE elementům, pochopitelně jen vůči translatovaným částem. Je tedy zřejmé, že krátké parazitické elementy bez vlastní možnosti transpozice nebereme ve srovnání diferenciální exprese v úvahu.

3.10.4 Detekce nabohacených GO termínů a KEGG drah

K anotaci *kontiguů* referenčního transkriptomu náleží též GO (gene ontology) – kontrolovaný slovník s anotací na třech úrovních – molekulární funkce, biologický proces a buněčný výskyt. Hypoteticky, náhodný výběr GO anotace by se neměl lišit od celkové populace GO v transkriptomu (nelze testovat malý výběr DE genů/GO), cílem je tedy vyvrátit nulovou hypotézu rovnosti výběru s celkovou populací. Toho dosáhneme např. Fisherovým exaktním testem; zvolená hladina signifikance α = 0,05; P hodnota jest korigovaná FDR metodou. Výpočet provedeme programem goatools v0.5.9 (Haibao Tang, 2015, nepublikováno). Podobně můžeme přistupovat i v případě KO (*kegg ortholog*).

KEGG metabolické a signální dráhy (*Kyoto Encyclopedia of Genes and Genomes*) jsou bohužel od roku 2009 zpoplatněny. Naštěstí existují webové aplikace, které umožňují automatickou annotaci *kontiguů* jako např. námi zvolený KAAS server (Moriya et al., 2007), tento server nabízí automatickou anotaci KO identifikátory. Získali jsme takto informaci, kolik DE genů bylo zamapováno do jakých metablických, či signálních drah a taktéž, kolik genů daných drah jsme byli celkově schopni zamapovat. Opět tedy máme počty výběru a celkové populace a je možné se tázat, zda je existjí v našem výběru nenáhodně nabohacené dráhy. Pro statistickou analýzu byl zvolen Fisherův exaktní test. Funkce v jazyce python je uvedena ve schématu č. 8. Korekce P hodnot (FDR) byla provedena funkcí p.adjust v jazyce R.

```
def fisher_t(table):
import stats
    # Výběr KEGG drah, počet genů v celkové populaci a výběru
    cur.execute("select id, a.kegg_name, a.cnt as cnt_bg, b.cnt as cnt_gl
from KEGG_all a left join %s b using (id) where b.cnt is not null;" %
table)
    test = cur.fetchall()
    count_bg = count_gl = 0
    # Suma genů zamapována v celé populaci
    for c in test:
        count_bg = count_bg + c[2]
    # Suma genů zamapována v našem výběru DE genů
    for cc in test:
        count_gl = count_gl + cc[3]
    # Fisherův exaktní test pro každou dráhu, do které byl zamapován
    # alespoň jeden gen
    for t in test:
        oddsratio, pvalue = stats.fisher_exact([[t[3],t[2]],
[count_gl,count_bg]])
        cur.execute("update %s set pVal = %s where id = %s" % (table,
pvalue, t[0]))
        conn.commit()
```

Schéma č. 8: Fisherův exaktní test pro detekci nabohacených drah v podmnožině DE genů

3.10.5 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)? Za druhé: Jsou detekované transkripty diferenciálně exprimované mezi skupinami (biologická variance). Za třetí: 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 (viz tab č. 8) 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 threshold*) 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‘

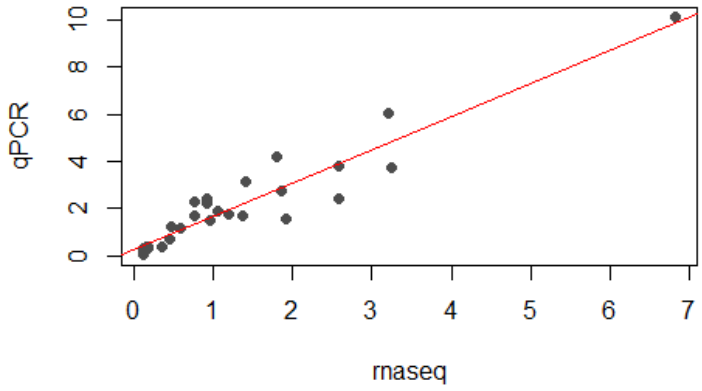
Tab č. 8: Sekvence primerů HS genů „posazených“ blízko 3’ konci sekvence užitých pro RT-qPCR validaci RNAseq relativní exprese

V tab. č. 9 uvedeny primery pro vybrané diferenciálně exprimované geny mezi skupinou sex *vs* asex. jaterní tkáně.

Fatt. Elong. FWD	TGGTGGTTTGTCTTGAACTGG
Fatt. Elong. REV	CAGCAGACAGCCCATAATACG
GSTs FWD	GCTGGAGCTGAGTTTGAGG
GSTs REV	CTGCATTCCATCCATTTCAACC
CH25H FWD	CCAGAACAGAGAAGATGTCTGG
CH25H REV	GAAGAGCACTGGGAAGAAGG
CPA2 FWD	ATGTGGCTCTATCTGCAAGC
CPA2 REV	ATGCCACGCTGGTAAGC
triqt FWD	TCTCTCAGGTGTAGAAGGATGG
triqt REV	CGATCTGTTTACGGTACTGATCC

Tab č. 9: Sekvence primerů DE genů „posazených“ blízko 3’ konci sekvence užitých v RT-qPCR validaci RNAseq relativní exprese

V grafu č. 17 je pozorovatelná těsná korelace - mezi datasety (RNAseq, qPCR) exprese jak pro housekeeping geny, tak geny určené jako diferenciálně exprimované v jaterní tkáni na základě negativně binom. Exprese oocytů nebyla záměrně srovnávána, jelikož HS geny v oocytech často samy vykazují značnou expresní variabilitu a cílem validace je pouze potvrzení repriducibility dat (vyloučení lidské chyby při analýze vzorků). Delta Ct hodnoty qPCR byly srovnány s hodnotami *fold-change* RNAseq vztažené k průměrné expresi vybraných HS genů.



Graf č. 17: Korelace mezi expresemi genů dat qPCR a RNAseq (diferenciálně exprimované geny: CPA2, fatty elongase, triqt, HPRT, 39S L13 protein, non-POU, glutathione S-transferase); spearmanův korelační koeficient činí 0.907.

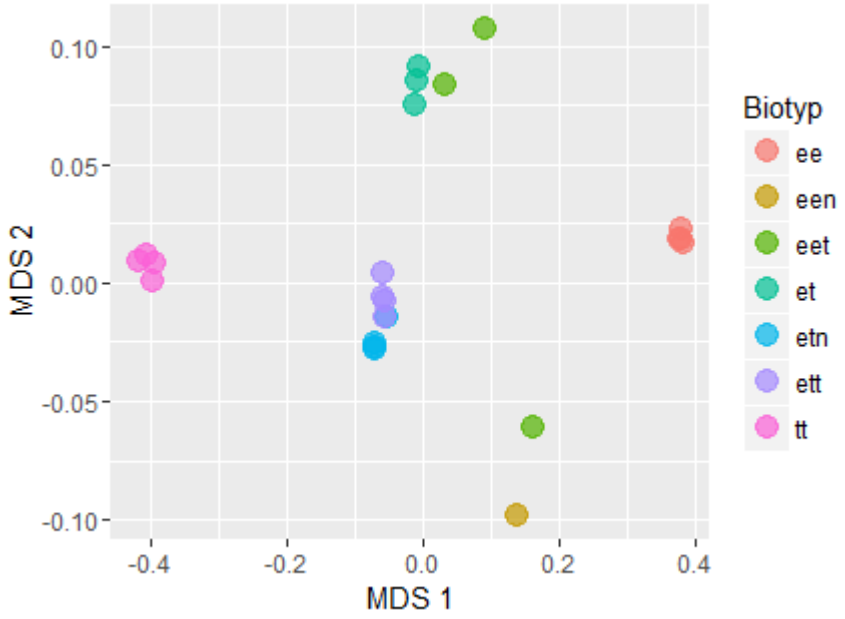
3.11 Analýza alelově specifické exprese (ASE) hybridních jedinců

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.

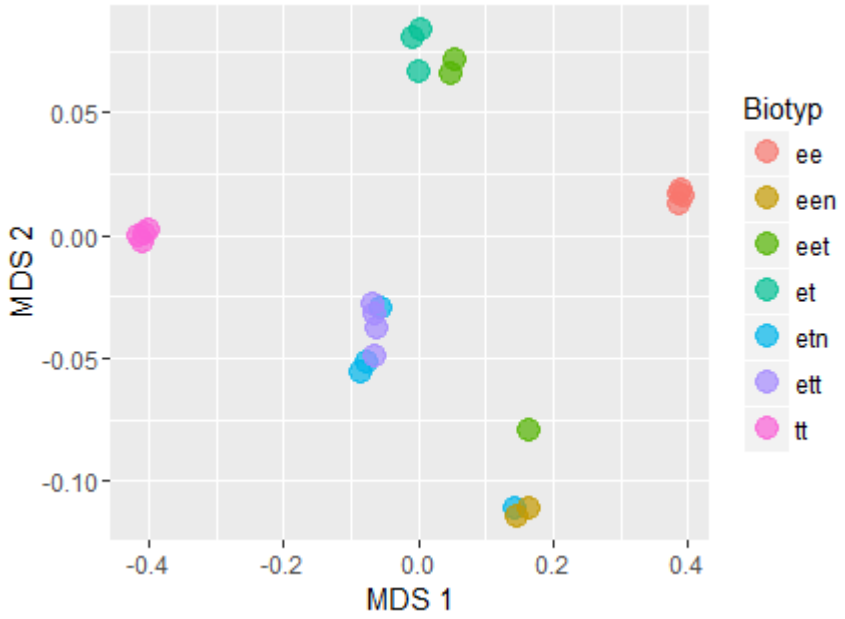
Premisou detekce disbalance exprese RNA v závislosti na původu z rodičovského druhu je následující. V první řádě je nezbytné adekvátním statistickým způsobem testovat, zda alela je, či není disbalancovaná. Nulová hypotéza je uvedena v schématu č. 9 – parametr eallel, jakožto poměr 1:1 dvou hlavních bází 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.

Nejprve ale v grafu č. 18 znázorňuji podobnost mezi SNP datasety jedinců, abychom mohli vyloučit záměnu vzorku, či zásadní chybu v analýze. V grafu č. 19 se bohužel ukazuje, že několik jedinců sekvenčně vybočuje (některé vzorky jsou problematické stejně jako v případě exprese). Nicméně sekvenční divergence mezi vzorky biotypů *eet* nemá biologické vysvětlení, protože by se mělo jednat o klony sdílející stejný

haplotyp dle cytochromu b, navíc pocházejících z blízkých lokalit (Polska Woda). Matice MDS shlukování je vygenerována programem plink v1.07 (Purcell et al., 2007).



Graf č. 18: MDS graf získaný ze SNP oocytů (416748 SNP zredukováno na 58659 s vysokou konfidencí - DP > 10 & MQ > 30 & QD > 20) zobrazeny komponenty s největší vysvětlující variancí.



Graf č. 19: MDS graf získaný ze SNP jaterních vzorků (416748 SNP zredukováno na 58659 s vysokou konfidencí - DP > 10 & MQ > 30 & QD > 20) zobrazeny komponenty s největší vysvětlující variancí.

3.11.1 Determinace původu alely na základě druhově specifických SNP

Abychom byli schopni říci, která z rodičovských alel dominuje, musíme mít nejen informaci o tom, zda je v dané alele diagnostický SNP, ale také jestli tento SNP mohl vzniknout jako následek mezidruhové hybridizace; tzn., že SNP v hybridním genomu se musí shodovat alespoň s jedním rodičovským druhem. V schématu č. 9 tedy vyznačujeme SNP dle heterozygotnosti a relativní jistoty toho, že se jedná o SNP přítomného v rodičovských populacích v době hybridizace.

```
#!/usr/bin/python
import pymysql
# připojení do lokální databáze MySQL
conn = pymysql.connect(host='localhost', port=3306, user='jenda',
passwd='xxx', db='cobitis_lom300tf')
cur = conn.cursor()
# do fuknce vstupuje list jedinců přes které iterujeme
def update_ok(hybrid):
   for i in hybrid:
      # Vybíráme z databáze SNP se sekvenační hloubkou >= 10
      cur.execute('''SELECT ref_acc, pos, %s_GT
                     FROM SNP_RNAseq_add JOIN SNP_RNAseq_add_INFO USING
(ref_acc, pos) WHERE %s_DP >= 10''' % (i, i))
      animal_GT = cur.fetchall()
      for ii in animal_GT:
         # Výběr dat z "Boole" tabluky, kde máme informaci, jaké
         # polymorfismy se na dané pozici u druhu nacházejí
         cur.execute('''SELECT ref_acc, pos, tt_a, tt_c, tt_g, tt_t,
ee_a, ee_c, ee_g, ee_t FROM SNP_RNAseq_add_bin WHERE ref_acc = '%s' AND
pos = %s''' % (str(ii[0]), int(ii[1])))
         animal_bin = cur.fetchone()
         # Přeskočíme pozici pokud nemáme kompletní data
         if (animal_bin[2] + animal_bin[3] + animal_bin[4] +
animal_bin[5]) < 1 or (animal_bin[6] + animal_bin[7] + animal_bin[8] +
animal_bin[9]) < 1:
            continue
         elif ii[2] is None:
            continue
         # Totéž platí v případě, že polymorfismus není shodný
         # s vnitrodruhovými variantami, tzn., že není děděn po rodičích
         elif (animal_bin[2] + animal_bin[6]) > 1 or (animal_bin[3] +
animal_bin[7]) > 1 or (animal_bin[4] + animal_bin[8]) > 1 or
(animal_bin[5] + animal_bin[9]) > 1:
            continue
         # Níže vybíráme SNP, které podmínky splňují a to pro homozygotní
         # i heterozygotní pozice všech možných genotypů
         elif (ii[2] == "A" and animal_bin[2] == 1) or (ii[2] == "C" and
animal_bin[3] == 1) or (ii[2] == "G" and animal_bin[4] == 1) or (ii[2] ==
"T" and animal_bin[5] == 1) or (ii[2] == "A" and animal_bin[6] == 1) or
(ii[2] == "C" and animal_bin[7] == 1) or (ii[2] == "G" and animal_bin[8]
== 1) or (ii[2] == "T" and animal_bin[9] == 1):
            cur.execute("UPDATE SNP_RNAseq_ok_et SET %s = 1 WHERE ref_acc
= '%s' AND pos = %s" % (i, str(ii[0]), int(ii[1])))
            conn.commit()
         elif ((ii[2] == "W" and animal_bin[2] == 1) or (ii[2] == "W" and
animal_bin[9] == 1)) or ((ii[2] == "S" and animal_bin[3] == 1) or (ii[2]
```

```
== "S" and animal_bin[8] == 1)) or ((ii[2] == "K" and animal_bin[4] == 1)
or (ii[2] == "K" and animal_bin[9] == 1)) or ((ii[2] == "M" and
animal_bin[2] == 1) or (ii[2] == "M" and animal_bin[7] == 1)) or ((ii[2]
== "Y" and animal_bin[3] == 1) or (ii[2] == "Y" and animal_bin[9] == 1))
or ((ii[2] == "R" and animal_bin[2] == 1) or (ii[2] == "R" and
animal_bin[8] == 1)):
            cur.execute("UPDATE SNP_RNAseq_ok_et SET %s = 1000 WHERE
ref_acc = '%s' AND pos = %s" % (i, str(ii[0]), int(ii[1])))
            conn.commit()
```

Schéma č. 9: Stanovení logické hodnoty (1,0) v případě, že platí podmínka původu SNP u hybridních jedinců (MySQL, python)

3.11.2 Stanovení disbalancovaných alel jedinců hybridního původu

Abychom byli schopni snadno sledovat disbalanci exprese mezi genomy *C. taenia* a *C. elongatoides*, označujeme diagnostické mezidruhové SNP třemi kategoriemi, dle disbalance SNP směrem k jednomu nebo druhému druhu (*tt*, *ee*, *NULL*). Níže, ve schématu č. 10 uvádím funkci k tomuto učenou.

```
def update_origin(hybrid):
   # Procházíme list hybridních jedinců vybíraje informace o genotypu
   for i in hybrid:
      cur.execute("SELECT ref_acc, pos, %s, %s_GT FROM SNP_RNAseq_ok_et
JOIN SNP_RNAseq_add USING (ref_acc, pos) WHERE %s IS NOT NULL" % (i, i,
i))
      animal_polymorf = cur.fetchall()
      # Dle primárních klíčů doplňujeme informace o druhové variabilitě
      # na pozici
      for ok in animal_polymorf:
         cur.execute("SELECT ref_acc, pos, tt_a, tt_c, tt_g, tt_t, ee_a,
ee_c, ee_g, ee_t FROM SNP_RNAseq_add_bin WHERE ref_acc = '%s' AND pos =
%s" % (ok[0], int(ok[1])))
         animal_bin = cur.fetchone()
         # Nyní je možné dle uvedených podmínek dodat informaci o původu
         # alely, pokud jedna z alel není exprimovaná
         if ok[2] == 1 and ((ok[3] == "A" and animal_bin[2] == 1) or
(ok[3] == "C" and animal_bin[3] == 1) or (ok[3] == "G" and animal_bin[4]
== 1) or (ok[3] == "T" and animal_bin[5] == 1)):
            cur.execute("UPDATE SNP_RNAseq_ok_et_INFO SET %s = 'tt' WHERE
ref_acc = '%s' AND pos = %s" % (i, ok[0], int(ok[1])))
            conn.commit()
         if ok[2] == 1 and ((ok[3] == "A" and animal_bin[6] == 1) or
(ok[3] == "C" and animal_bin[7] == 1) or (ok[3] == "G" and animal_bin[8]
== 1) or (ok[3] == "T" and animal_bin[9] == 1)):
            cur.execute("UPDATE SNP_RNAseq_ok_et_INFO SET %s = 'ee' WHERE
ref_acc = '%s' AND pos = %s" % (i, ok[0], int(ok[1])))
            conn.commit()
         if ok[2] == 1000 and ((ok[3] == "W" and animal_bin[2] == 1 and
animal_bin[5] == 1) or (ok[3] == "S" and animal_bin[3] == 1 and
animal_bin[4] == 1) or (ok[3] == "K" and animal_bin[4] == 1 and
animal_bin[5] == 1) or (ok[3] == "M" and animal_bin[2] == 1 and
animal_bin[3] == 1) or (ok[3] == "Y" and animal_bin[3] == 1 and
animal_bin[5] == 1) or (ok[3] == "R" and animal_bin[2] == 1 and
animal_bin[4] == 1)):
```

```
            cur.execute("UPDATE SNP_RNAseq_ok_et_INFO SET %s = 'tt' WHERE
ref_acc = '%s' AND pos = %s" % (i, ok[0], int(ok[1])))
            conn.commit()
         if ok[2] == 1000 and ((ok[3] == "W" and animal_bin[6] == 1 and
animal_bin[9] == 1) or (ok[3] == "S" and animal_bin[7] == 1 and
animal_bin[8] == 1) or (ok[3] == "K" and animal_bin[8] == 1 and
animal_bin[9] == 1) or (ok[3] == "M" and animal_bin[6] == 1 and
animal_bin[7] == 1) or (ok[3] == "Y" and animal_bin[7] == 1 and
animal_bin[9] == 1) or (ok[3] == "R" and animal_bin[6] == 1 and
animal_bin[8] == 1)):
            cur.execute("UPDATE SNP_RNAseq_ok_et_INFO SET %s = 'ee' WHERE
ref_acc = '%s' AND pos = %s" % (i, ok[0], int(ok[1])))
            conn.commit()
```

Schéma č. 10: Funkce sloužící k detekci druhově specifických SNP, které zároveň splňují podmínku, že alela pochází od rodičovských druhů.

K určení samotné disbalance exprese vycházíme z počtu podpůrných bází *namapovaných readů* na konkrétní pozici transkriptomu – pouze u pozic z předešlých analýz označených N. Pro statistickou detekci volíme základní binomiální test (Castel et al., 2015), viz schéma č. 12. Níže uvedený výňatek (schéma č. 11) bash a python skriptů nejprve z mpileup výstupu samtools spočítá počet bází vůči N referenci a poté zapisuje do databáze P hodnoty srovnání dvou majoritních bází z *alignmentu* s tím, že u triploidních jedinců je nezbytné testovat modely dva (poměr alel 1/3 a 2/3).

```
# Podle výpisu polymorfních míst extrakce bází mapovaných na daný loci a
# pozici pomocí samools (pominutí indels a bází 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
# Python funkce určená z počítání bazí v alignmentu
def cnt_allele():
    cur.execute("SELECT animal, ref_acc, pos, mpileup FROM
SNP_RNAseq_GT")
    mpileup = cur.fetchall()
    for pozice in mpileup:
        # Malá a velká písmena bazí označují sekvenční orientaci
        baze = str(pozice[3]).lower()
        A_cnt = baze.count("a")
        C_cnt = baze.count("c")
        G_cnt = baze.count("g")
        T_cnt = baze.count("t")
        cur.execute('''UPDATE SNP_RNAseq_GT SET
                        A_cnt = %s, C_cnt = %s, G_cnt = %s, T_cnt = %s
                          WHERE animal = '%s' AND ref_acc = '%s' AND pos
= %s''' % (int(A_cnt), int(C_cnt), int(G_cnt), int(T_cnt),
str(pozice[0]), str(pozice[1]), int(pozice[2])))
        conn.commit()
```

Schéma č. 11: Výňatek postupu získání SNP a alelické sekvenační hloubky z RNAseq dat

```
def binom_t():
    cur.execute("SELECT animal, ref_acc, pos, A_cnt, C_cnt, G_cnt, T_cnt
FROM SNP_RNAseq_ASE2")
    alely = cur.fetchall()
```

```
    alely_poradi = P_vals_di = P_vals_di_FDR = P_vals_pol1 = P_vals_pol2
= P_vals_pol = P_vals_pol_FDR = ls_di = ls_pol = []
    for ii in alely:
        # pokud je jedinec diploid; H0 = 1/2
        if biotyp[ii[0][0:5]] in ("tt", "ee", "et"):
            ii_dict = {'A':ii[3], 'C':ii[4], 'G':ii[5], 'T':ii[6]}
            ii_dict_s = sorted(ii_dict.items(),
key=operator.itemgetter(1), reverse = True)
            P_vals_di = stats.binom_test(x=(ii_dict_s[0][1],
ii_dict_s[1][1]), n=sum(ii[3:len(ii)]), p=1/2)
            cur.execute("UPDATE SNP_RNAseq_ASE2 SET pVal = %s WHERE
animal = '%s' AND ref_acc = '%s' AND pos = %s" % (P_vals_di, ii[0],
ii[1], ii[2]))
            cur.commit()
        # pokud je jedinec triploid; H0 = 1/3, ci 2/3
        else:
            ii_dict = {'A':ii[3], 'C':ii[4], 'G':ii[5], 'T':ii[6]}
            ii_dict_s = sorted(ii_dict.items(),
key=operator.itemgetter(1), reverse = True)
            P_vals_pol1 = stats.binom_test(x=(ii_dict_s[0][1],
ii_dict_s[1][1]), n=sum(ii[3:len(ii)]), p=1/3)
            P_vals_pol2 = stats.binom_test(x=(ii_dict_s[0][1],
ii_dict_s[1][1]), n=sum(ii[3:len(ii)]), p=2/3)
            # Jelikož nevíme, jaký typ poměru alel u hybrida očekávat,
            # volíme vyšší P hodnotu
            if P_vals_pol1 >= P_vals_pol2:
                P_vals_pol = P_vals_pol1
                cur.execute("UPDATE SNP_RNAseq_ASE2 SET pVal = %s WHERE
animal = '%s' AND ref_acc = '%s' AND pos = %s" % (P_vals_pol, ii[0],
ii[1], ii[2]))
            else:
                P_vals_pol = P_vals_pol2
                cur.execute("UPDATE SNP_RNAseq_ASE2 SET pVal = %s WHERE
animal = '%s' AND ref_acc = '%s' AND pos = %s" % (P_vals_pol, ii[0],
ii[1], ii[2]))
```

Schéma č. 12: Binomiální test, srovnání majoritní a minoritní alely s hypotézou „vybalancované“ exprese mezi alelami.

Nezbytností se opět stává korekce pro mnohonásobné testování. V modulu scipy tato fuknce není napsána, pomůžeme si tedy R funkcí - p.adjust – schéma č. 13 R skriptu.

```
#!/usr/bin/Rscript
library(DBI)
library(RMySQL)
m <- dbDriver("MySQL")
con <-
dbConnect(m,user='jenda',password='xxx',host='localhost',dbname='cobitis_
lom300tf')
# Výběr všech řádků primárního klíče a nekorigované P hodnoty
res <- dbSendQuery(con, "select animal, ref_acc, pos, pVal from
SNP_RNAseq_ASE2;")
p_Val <- fetch(res, n = -1)
pVal_adj <- p_Val
pVal_adj <- cbind(pVal_adj, pVal_FDR =
c(rep('NA',length(nrow(pVal_adj)))))
pVal_adj$pVal_FDR <- p.adjust(p_Val$pVal, method = "fdr")
head(pVal_adj)
```

```
# Update FDR korigované P hodnoty dle primárního klíče
for (i in 1:length(rownames(pVal_adj))) {
  sql <- sprintf("update SNP_RNAseq_ASE2
                  set pVal_FDR = %f where animal = '%s' and ref_acc =
                  '%s' and pos = %s;",
                 pVal_adj[i,5], pVal_adj[i,1], pVal_adj[i,2],
                 pVal_adj[i,3])
  dbGetQuery(con, sql)
}
```

Schéma č. 13: R skript popisující korekci P hodnoty binomiálního testu a její zápis do relační databáze

3.11.3 Statistická analýza ASE loci hybridních jedinců

Jelikož je volba binomiální distribuce vhodnější pro naše data nežli např. chi distribuce, jsme si i přesto vědomi, že existují lepší statistické přístupy redukující falešně pozitivní signál. Navíc je otázkou exprese, která z alel je potlačena, ale nikoliv jen na úrovni individuálních SNP, ale na úrovni celého loci nebo genu. Hlavní zdroje problémů analýzy ASE tkví v technické variabilitě sekvenování, mapování a genotypizační chybou. Tedy jak naložit s geny, kde je alela dle SNP disbalancovaná oběma směry. V našem případě se jeví jediná možnost - takový geny vyřadit. Abychom mohli srovnat globální úroveň umlčení bez vnesení velkého množství chyb, je pro nás důležité vyřadit maximum falešně pozitivních signálů. Bohužel i přes veškeré úskalí nejsme bez genomické informace schopni rozlišit genové konverze a *nonsence mediated decay* od cis, trans deregulace. Veškeré analýzy budou zaměřeny spíše na globální komparaci, nežli na funkční popis afektovaných genů, neboť postrádáme podstatnou validaci dat.

3.12 Analýza Müllerovy rohatky

3.12.1 Identifikace otevřených čtecích rámců v genech a zápis SNP jedinců do referenční sekvence

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 „vyřezány“ dle pozic začátku a konce nejdelšího ORF. 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, jak vzorků jater, tak oocytů, a to

identickým způsobem jako vytvoření „konsenzuální“ reference. SNP jedince musely splňovat podmínky kvality genotypu a hloubky: GQ >= 20 & DP >= 10, aby byly zapsány namísto báze referenční sekvence. Soubor byl formátován tak, aby obsahoval modifikované sekvence všech jedinců konkrétního nejdelšího ORF genu. Opět zdůrazňuji, že „pouhých“ ~5000 genů/ORF je dostatečně prosekvenováno na to, aby mohly být odhaleny SNP, ve zbytku ORF byly ponechány referenční báze, tzn. ztrátu informace snížením dN/dS od reálné hodnoty, nikoliv opačně.

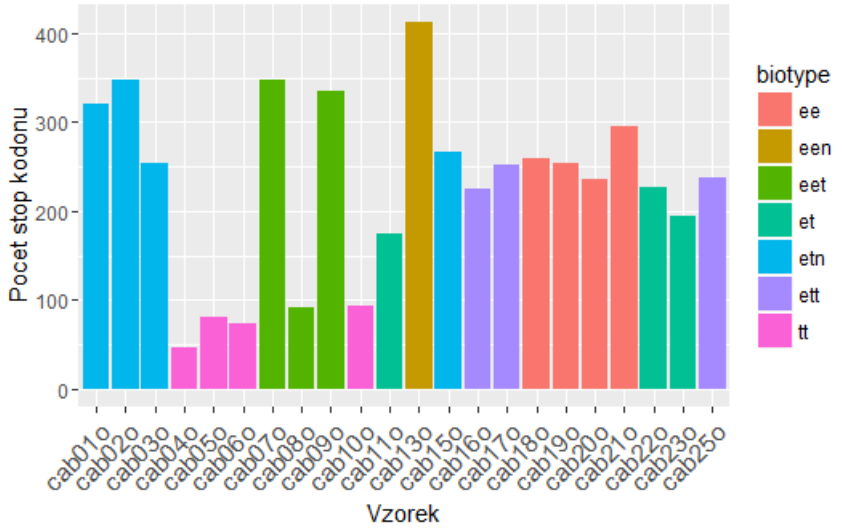
3.12.2 Kontrola ORF pro přítomnost stop kodónů vzniknuvších vnesením SNP

Abychom si byli jisti smysluplností dat, je na místě provést kontrolu, zda nahrazením bází v referenci nedochází ke generování stopkodónů. V určitých případech bychom to očekávat mohli. Pokud některý z loci hybridní linie nebude pod selekcí, nebo naopak bude pod negativní selekcí, můžeme očkávat i frameshift a non-sence mutace, které bychom ale na druhou stranu neměli detekovat díky dráze *nonsence mediated decay*, která detekuje aberantní mRNA se stop kodóny a zajišťuje její rozklad. Pro kvantifikaci stop kodónu v ORF uvádím ve schématu č. 14 jednoduchý skript v jazyce python.

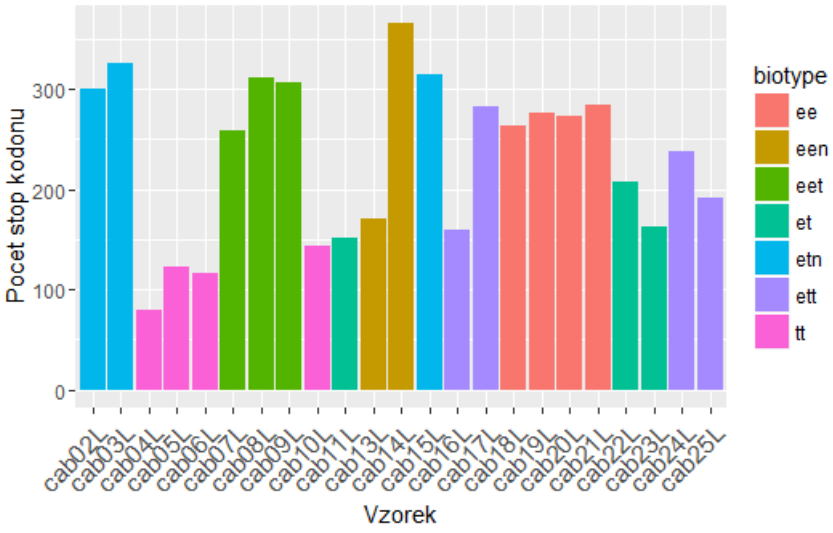
```
# import knihoven z modulu Biopython
import glob
from Bio.Seq import Seq
from Bio import SeqIO
# Definování stopkodónů a slovníku významu ambiguidních bází
stops = ("TAA", "TGA", "TAG")
abbig = ("W", "R", "Y", "S", "K", "M")
ambig_dict = {"W":("A", "T"), "R":("A", "G"), "S":("C","G"), "K":("G",
"T"), "M":("A", "C")}
# funkce pro tisk pozic obsahující stop kodón, a to i v případě, kdy je
zasažna pouze jedna alela, pro kombinace více heterozygotních bází nelze
stop kodón testovat, protože nevíme, zda se jedná o cic, či trans
postavení.
def check_stop_codon(sekvence, nazev, zvire):
    for seq in range(0, len(sekvence), 3):
        # V případě, že nejsme schopni určit jednu bázi tripletu, triplet
        # přeskakujeme
        if "N" in sekvence[seq:seq+3]:
            continue
        # Totéž činíme v případě, že stop kodón nenalézáme
        elif sekvence[seq:seq+3] not in stops:
            continue
# V případě, že kodón nebo alespoň jedna z alel obsahuje stop kodón,
# vypíšeme gen, ve kterém se nachází
        elif sekvence[seq:seq+3] in stops:
            print(zvire,"stopkodon nalezen, nazev genu: %s" % nazev)
        elif sekvence[seq:seq+3][0] in ambig:
            phase = ambig_dict[sekvence[seq:seq+3][0]]
            seq_lis1 = seq_lis2 = list(sekvence[seq:seq+3])
```

```
                seq_lis1[0] = phase[0]
                seq_lis2[0] = phase[1]
                seq_lis1 = ''.join(seq_lis1)
                seq_lis2 = ''.join(seq_lis2)
                if seq_lis1 in stops or seq_lis2 in stops:
                    print(zvire,"stopkodon nalezen, nazev genu: %s" % nazev)
            elif sekvence[seq:seq+3][1] in ambig:
                phase = ambig_dict[sekvence[seq:seq+3][1]]
                seq_lis1 = seq_lis2 = list(sekvence[seq:seq+3])
                seq_lis1[1] = phase[0]
                seq_lis2[1] = phase[1]
                seq_lis1 = ''.join(seq_lis1)
                seq_lis2 = ''.join(seq_lis2)
                if seq_lis1 in stops or seq_lis2 in stops:
                    print(zvire,"stopkodon nalezen, nazev genu: %s" % nazev)
            elif sekvence[seq:seq+3][2] in ambig:
                phase = ambig_dict[sekvence[seq:seq+3][2]]
                seq_lis1 = seq_lis2 = list(sekvence[seq:seq+3])
                seq_lis1[2] = phase[0]
                seq_lis2[2] = phase[1]
                seq_lis1 = ''.join(seq_lis1)
                seq_lis2 = ''.join(seq_lis2)
                if seq_lis1 in stops or seq_lis2 in stops:
                    print(zvire,"stopkodon nalezen, nazev genu: %s" % nazev)
            else:
                print(zvire,"fatal error")
# Iterujeme přes všechny jedince z RNAseq
soubory = glob.glob('lom300tf_*')
for s in soubory:
    handle = open(s, "rU")
    for record in SeqIO.parse(handle, "fasta"):
        zviera = record.id.split("_")
        nazev1 = zviera[0:2]
        nazev2 = '_'.join(nazev1)
        print(check_stop_codon(record.seq, nazev2, zviera[2]))
handle.close()
```

Schéma č. 14: Skript určený pro detekci stop kodónů v ORF tím, že nahradíme referenční bázi druhu *C. taenia* SNP konkrétního jedince; mezi stop kodóny počítáme i heterozygotní stav, kdy v jedné alele vznikne stop kodón.



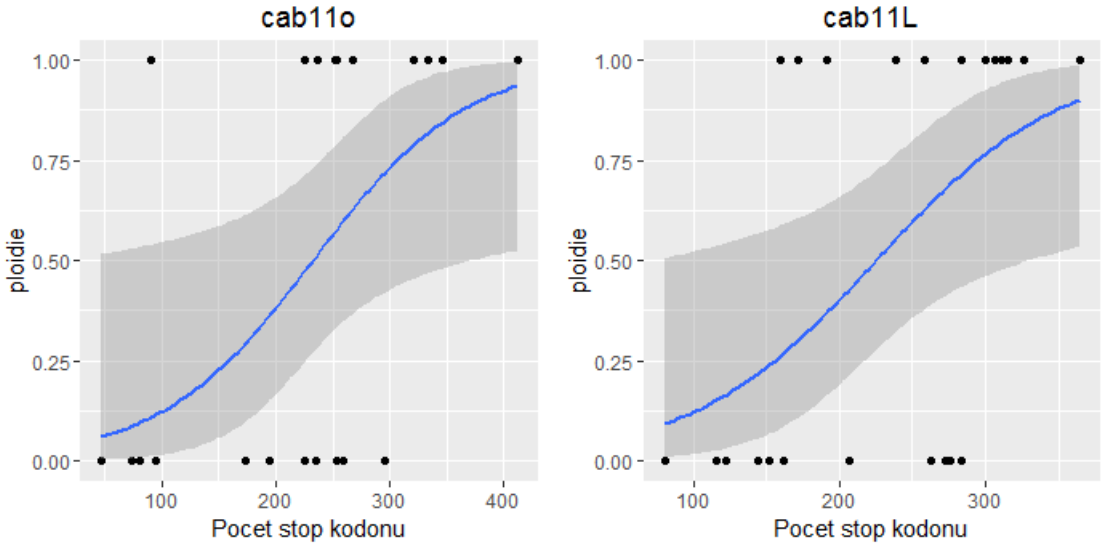
Graf č. 20: Kontrola dat pro Müllerovu rohatku – vznik stop kodónu v ORF vnesením SNP jednotlivých jedinců derivovaných z **oocytů**



Graf č. 21: Kontrola dat pro Müllerovu rohatku – vznik stop kodónu v ORF vnesením SNP jednotlivých jedinců derivovaných z **jater**

Z grafů č 20., 21., 21.1 je patrná závislost počty stop kodónů na ploidii i sekvenační hloubce; závislost na polyploidii a můžeme vyjádřit např. GLM modelem binomiální rodiny, viz graf č. 21.1 (oocyty; ploidie: signifikantní na 0.01, střední chyba 0.0337, s. hloubka: signifikantní na 0.05, std. chyba: 0.0841). S nárůstem počtu *readů* jsme schopni identifikovat více SNP, narůstá tedy i počet chybných určení. Vliv polyploidie bohužel nejsem schopen technicky ani biologicky interpretovat.

Bohužel vysoké počty stop kodónu jsou značně znepokojivé, mohou indikovat problémy v SNP determinaci.



Graf č. 21.1: Logistická regrese (binomiální) počtu stop kodónů na ploidii (diploid – 0, triploid 1)

3.12.3 Výpočet dN/dS poměru z párového srovnání

Pro otestování hypotézy Müllerovy rohatky byl proveden výpočet poměru nesynonymních a synonymních mutací (dN/dS) pro různé biotypy, včetně *et* hybridů. Dále byl stejný výpočet proveden pro uměle vytvořené *et* hybridy, které byly získány náhodným zkombinováním odpovídajících sekvencí rodičovských druhů. Tyto rodičovské sekvence byly nejdříve zduplikovány, poté byla každá variantní pozice (SNP) náhodně (ale unikátně) přiřazena bázi tak, že každý duplikát obsahoval jinou variantu. Takto vzniklé sekvence z rodičovských druhů byly zkombinovány mezi druhy za vzniku umělých F1 hybridů, na kterých bylo opět měřen dN/dS poměr. Ostatní sekvence byly upraveny stejným způsobem a všechny takto upravené sekvence byly následně srovnávány 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 byl proveden 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 opět 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.

Bylo využito 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 nemáme 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.

Byly vytvořeny 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

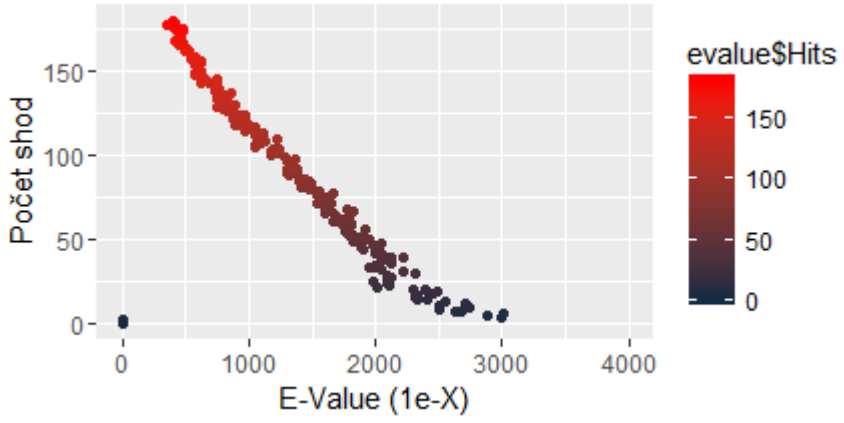
4.1 Evaluace referenční sekvence

Jedním z prvních výsledků, na které navazují veškeré další analýzy této práce, je samotná příprava a vyhodnocení referenční sekvence. Byli jsme schopni sestavit celkem 21508 *kontigů,* z čehož 18342 *kontigů* je složeno na základě normalizovaných 454 dat dlouhého čtení. Zbytek referenčních sekvencí tvoří *assembly* illumina nenormalizovaných dat krátkého čtení doplňující geny, které jsme díky nízkému pokrytí nebyli schopni složit a taktéž geny, které jsou exprimovány pouze u hybridních jedinců. 6789 *kontigů* bylo anotováno. Nízký počet anotovaných genů je zřejmě zapříčiněn značnou distancí od nejbližšího genomu *D. rerio* a také nedokonalou reverzní transkripcí od polyA konce mRNA umocněnou faktorem nízkého sekvenčního pokrytí.

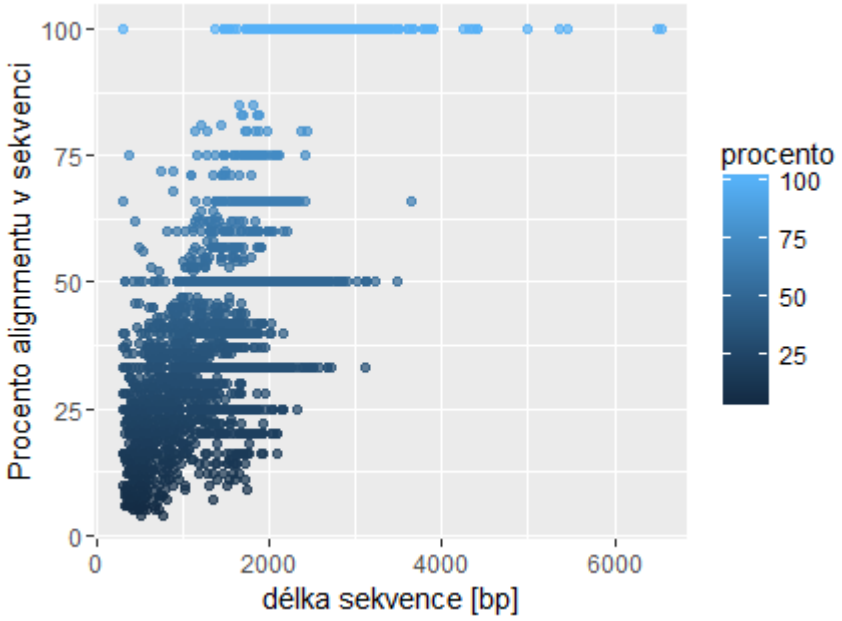
Délka sekvence má na úspěšnost anotace velký význam, protože se snižující se délkou sekvence stoupá pravděpodobnost, že sekvence v databázi může být zcela náhodné kompozice (*e-value* se snižuje exponenciálně s délkou sekvence). Jelikož limit délky cDNA byl nastaven pro sekvence delší než 300, je význam délky cDNA na anotaci již marginální, nicméně u krátkých sekvencí je úspěšnost anotace nižší, viz graf č. 25. V transktriptomu by ale mohly být zastoupeny lncRNA, je tedy zřejmé, že z porovnání obsahu parametrů GC (Niazi and Valadkhan, 2012) (viz graf č. 27), ORF (viz graf č. 26) a délek sekvencí mezi anotovanými a neanotovanými *datasety* (viz graf č. 25) vyplývá, že část těchto sekvencí je zřejmě nekódujících nebo jsou tyto sekvence tvořeny převážně 3´ UTR. Detekce dlouhých nekódujících 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) a Rfam přístupem. Blastn bylo nalezeno 114 ncRNA pod prahovou hodnotou $1x10^{-6}$, z 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 významný exces lncRNA (Kaushik et al., 2013; Pauli et al., 2012). Naopak zřejmě validní premisou může být přítomnost velkého množství UTR sekvencí, především pak 3' UTR, vycházíme-li z toho, že mRNA byla sekvenována přípravou reverzní transkripcí od 3' polyA. Bohužel detekce je opět problematická, zaměřuje se především na přítomnost polyA signálu a dalších regulačních motivů, pro jejichž detekci je nutné aplikovat sofistikovanější algoritmy.

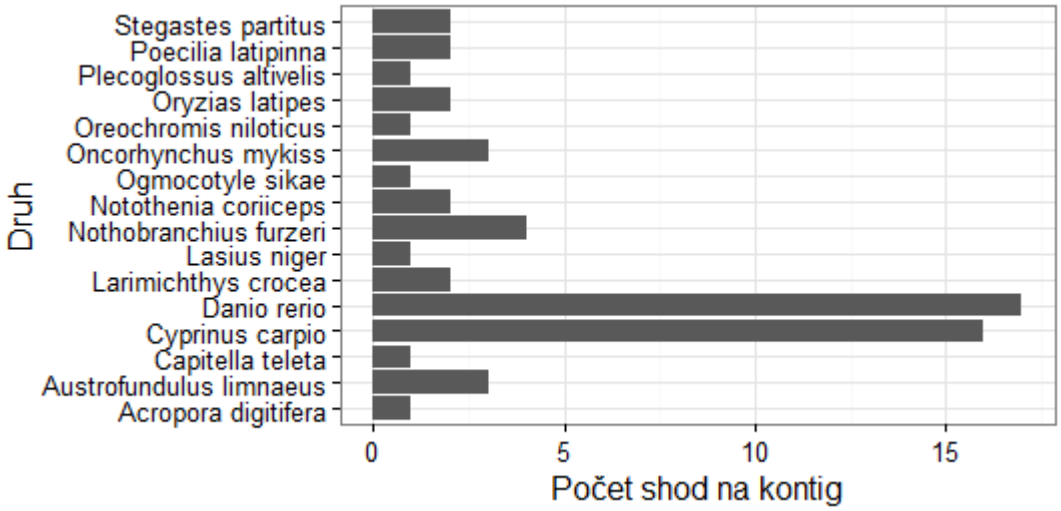
Při přípravě cDNA zřejmě nedošlo k „závažné“ kontaminaci, ať již parazitem, či následkem nesterilní přípravy. Naprostá většina sekvencí totiž náleží kostnatým rybám. V datech se objevilo přibližně 100 sekvencí přiřazených k savčím a bezobratlých živočichům, nejvíce pak lidských (22) a myších (21). Tyto sekvence vysazují velmi nízkou e-*value* - zřejmě se jedná o kontaminace, tudíž byly z následujících analýz vyloučeny. Četnost nejlepších *alignmentů* podle druhů je uvedena v grafu č. 24; distribuce *eValue* je uvedena v grafu č. 22.



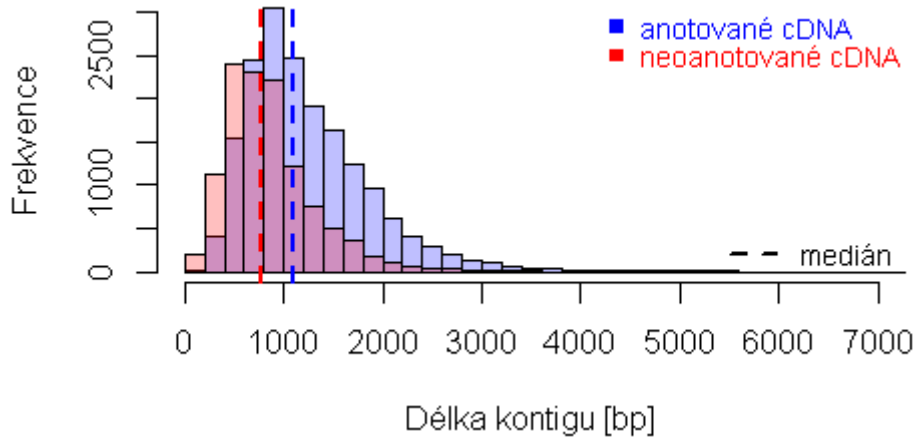
Graf č. 22: Rozdělení eValue hodnot sekvencí vzhledem k počtu *alignmentů* (na ose x je uveden pouze exponent X: 10^{-X}).



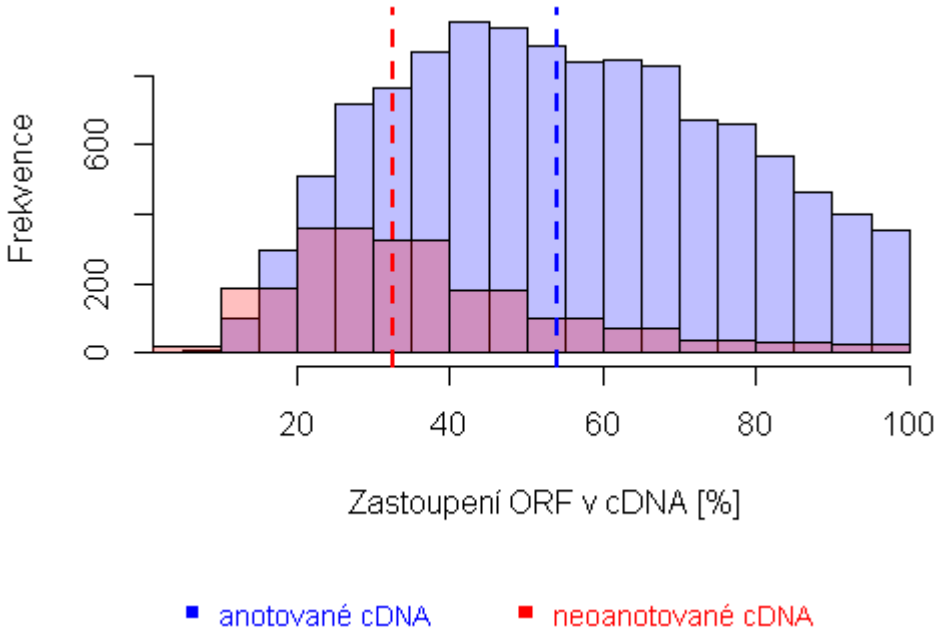
Graf č. 23: Závislost délky anotovaného *kontigu* na procentuálním zastoupení nejdelšího *alignmentu* vůči proteinové databázi.



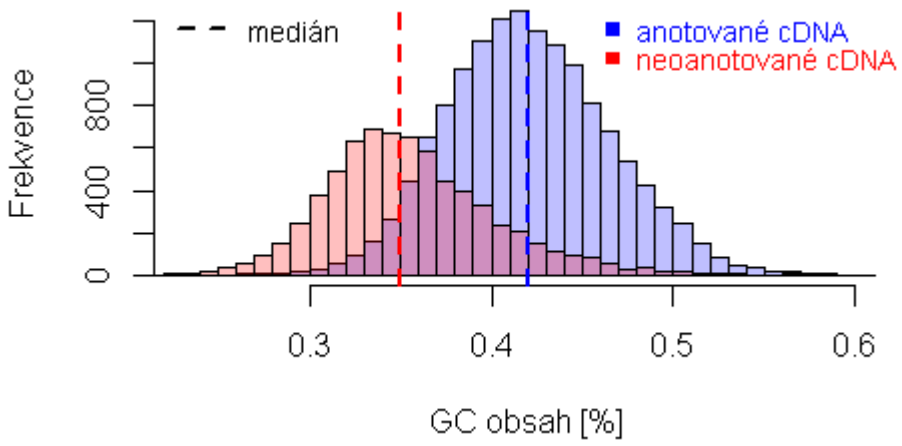
Graf č. 24: Sloupcový diagram znázorňující průměrné druhové zastoupení na *kontigu* („gen“)



Graf č. 25: Histogram, komparace rozdělení délek *kontiguů* anotovaných a bez anotace

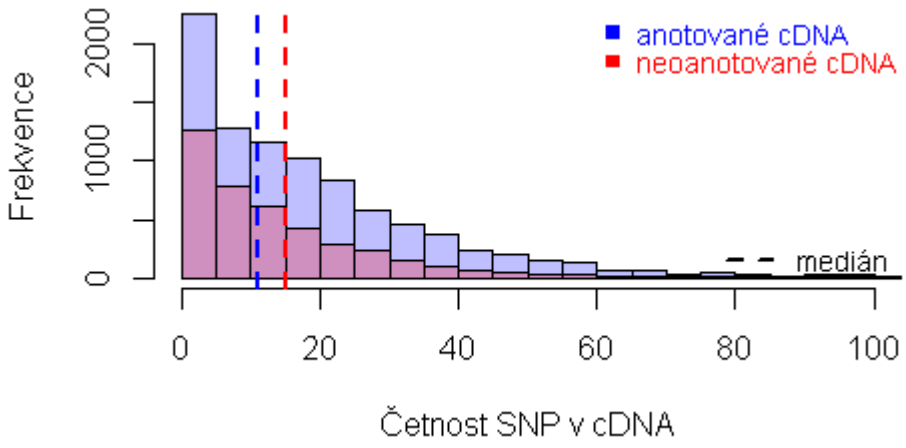


Graf č. 26: Porovnání souborů rozdělení procentuálního zastoupení otevřeného čtecího rámce (ORF) vzhledem k celkové délce kontigů cDNA mezi anotovanými a neanotovanými soubory sekvencí.



Graf č. 27: Histogram porovnání procentuálního obsahu GC v setu anotovaných a neoanotovaných cDNA.

V neanotovaných sekvencích byl také identifikován vyšší počet polymorfních pozic, viz graf č. 28, 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. 28: Histogram distribuce frekvence polymorfismů na cDNA (četnost SNP není normalizována na délku sekvence, neboť neanotovaný soubor sekvencí má nižší medián délky)

4.2 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, kterým je také věnována největší pozornost. Sekundárně se výsledky zaměřují na prezentaci DE genů, které vznikly následkem polyploidizace a DE genů vycházejících z mezidruhových rozdílů. Veškeré

nalezené DE geny následujících srovnání jsou uvedeny v elektronické příloze pro svou velikost.

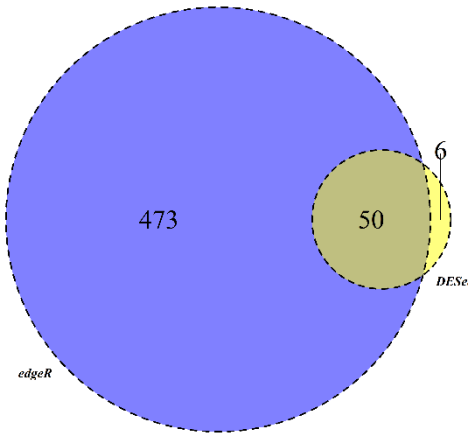
Cíle detekce diferenciální exprese byly tyto:

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). Odpovědět na otázku, zda došlo k aktivaci transpozonů a konkrétně jakých tříd, pojímáme spéše marginálně, protože systematicky opomíjíme celé třídy transpozonů, díky tomu, že jsme redukovali transkriptom na základě sekvenční podobnosti a definovali pouze protein kódující sekvence transpozonů.

Je nezbytné si uvědomit, že oocyty uvedeného stádia jsou téměř transkripčně neaktivní a většina mRNA je pouze maternálním pozůstatkem, který však definuje vývoj embrya. Naše pozorování diferenciální exprese týkající se oocytů je pouze následek již proběhnuvší determinace na gynogeneticky se replikující embryo. Problematickým fenoménem vnášejícím další zdroj variability do dat diferenciální exprese mezi oocyty hybridů a sexuálně se reprodukujícími druhy je polyploidie, která může nebo nemusí, a

často i zcela nepredikovatelně, změnit transkripční profil mnoha drah spojených i s vývojem orgnismu (Anatskaya et al., 2016). Ve třetím případě se u hybridů, a to tkáně jakéhokoliv původu, mohou projevit expresní mezidruhové rozdíly (změny v cis- trans regulaci exprese RNA), které mají větší význam nežli mutace v kóujcích oblastech genu (Wray, 2007). 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. 28 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 v podstatě podmnožinou DE genů získaných přístupem edgeR, ale výběrem na stejné hodnotě alfa (FDR korekce) byly získány velmi rozdílné počty DE genů. Bohužel edgeR je náchylný vůči odlehlým 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 na druhou stranu smířit s rizikem falešně pozitivních dat.



Graf č. 28: Vennův diagram množin DE genů nalezených dvěma přístupy selekcí na hladině α = 0.05, *FDR* (edgeR a DESeq) srovnáním skupin asexuálních a sexuálně se reprodukujících skupin tkáně **oocytů** bez jedinců obsahující *n* genom.

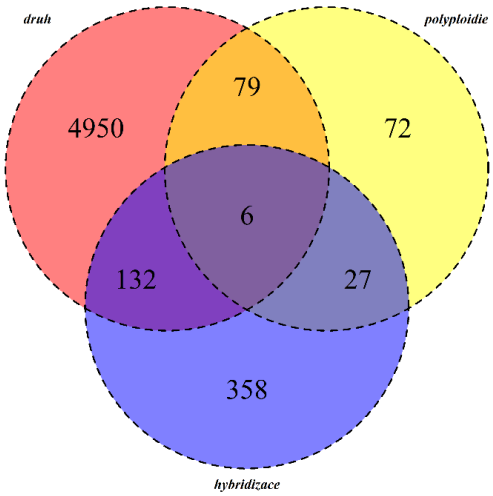
Z celkového počtu celkového počtu 20470 cDNA reference jsme nalezli 473 všech DE genů v oocytech při srovnání sexuálních a hybridních samic bez jedinců s *n* genomem, z toho bylo 227 podexprimováno a 296 nadexprimováno u hybridů. U jater jsem nalezl 821 genů (267 nadexprimováno u hybridů, 554 podexprimováno u hybridů; bez *n* genomu). Veškeré tyto geny ze srovnání oocytů, ale nemůžeme označit jako DE geny s kýženým

biologickým významem, tj. geny, které jsou zodpovědné za gynogenetickou formu rozmnožování, potažmo klonální vývoj embrya.

Významným faktorem generujícím rapidní změny exprese u hybridních forem rodičovských druhů může být i smotná polyploidie, bohužel vysvětlit podíl její variability v komplexních hybridních modelech je obtížné (Wu et al., 2016). Je ale známo, že i autopolyploidizace u rostlin může být spojená s deregulací transkripce skrze epigenetické mechanizmy (Zhang et al., 2014). Významné změny ale nepozorujeme. Nalezli jsme 184 všech DE genů v oocytech při srovnání diploidních a polyploidních samic, z toho bylo podexprimováno 110 genů a 74 nadexprimováno u diploidů (bez *n* genomu). U jater jsem nalezl 646 DE genů z čehož je 599 podexprimováno a 47 nadexprimováno u diploidů.

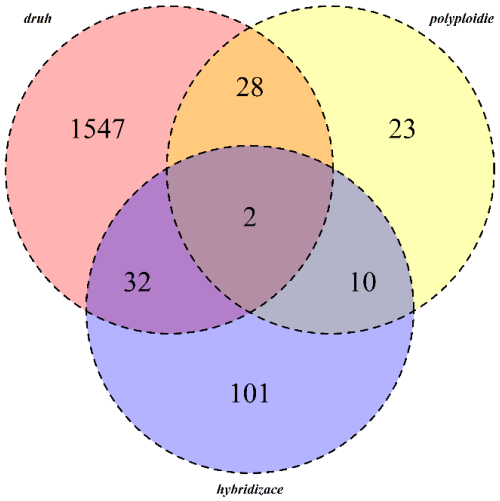
Zdaleka nejvýznamnějším činitelem, co do rozdílnosti genové exprese se v našem případě DE genů ukázalo mezidruhové srovnání *Cobitis taenia* vůči *C. elongatoides*. V jaterní tkáni nalezeno 3848 DE genů z čehož je 1359 *nadexprimováno* u druhu *tt* (viz graf č. 31). V oocytech je situace obdobná, bylo identifikováno 5167 DE genů; 2026 *nadexprimovaných* DE genů u *tt* a 3141 DE *podexprimovaných* u druhu *ee* (viz graf. č. 29).

V grafu Vennova diagramu č. 29 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 genom *n (pocházející od C. tanaitica)*, protože k jeho hybridním formám nemáme jejich rodičovské druhy.



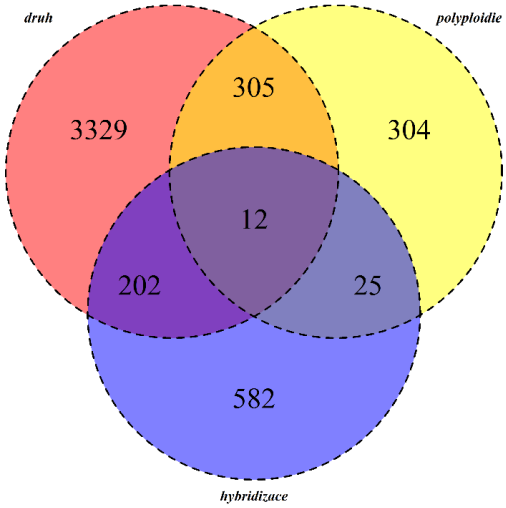
Graf č. 29: Vennův diagram množiny párového srovnání nalezených DE genů (včetně neoanotovaných) tkáně **oocytů** mezi druhy *ee*, *tt* = kategorie **druh**; množiny DE genů získaných srovnáním diploidních a polyploidních jedinců = kategorie **polyploidie** a množiny DE párového srovnání gynogeneticky a sexuálně se reprodukujících jedinců = kategorie **hybridizace**

Když jsem provedl srovnání množin DE genů na datech, ze kterých byly vyřazeny neanotované geny, zůstaly poměry průniků množin DE genů velmi podobné, viz graf. č. 30.



Graf č. 30: 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 reprodukujících jedinců = kategorie **hybridizace**

V jaterní tkáň je celkový počet DE genů všech srovnávaných skupin naopak nižší nežli v oocytech, došlo ale k znatelnému navýšení DE genů mezi skupinami di- a polyploidů, viz graf č. 31.

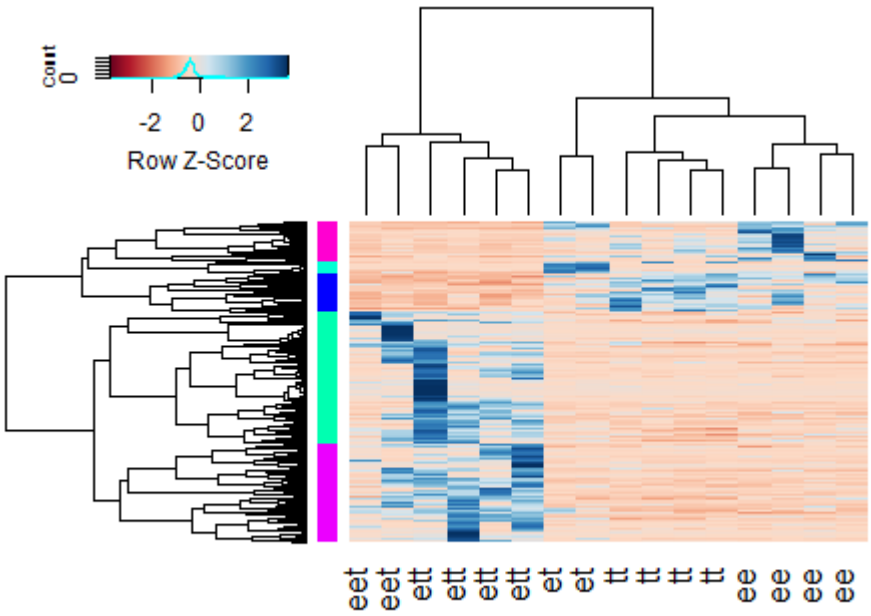


Graf č. 31: Vennův diagram množiny párového srovnání nalezených DE genů tkáně **jater** mezi druhy *ee*, *tt* = kategorie **druh**; množiny DE genů získaných srovnáním diploidních a polyploidních jedinců = kategorie **polyploidie** a množiny DE párového srovnání gynogeneticky a sexuálně se reprodukujících jedinců = kategorie **hybridizace**

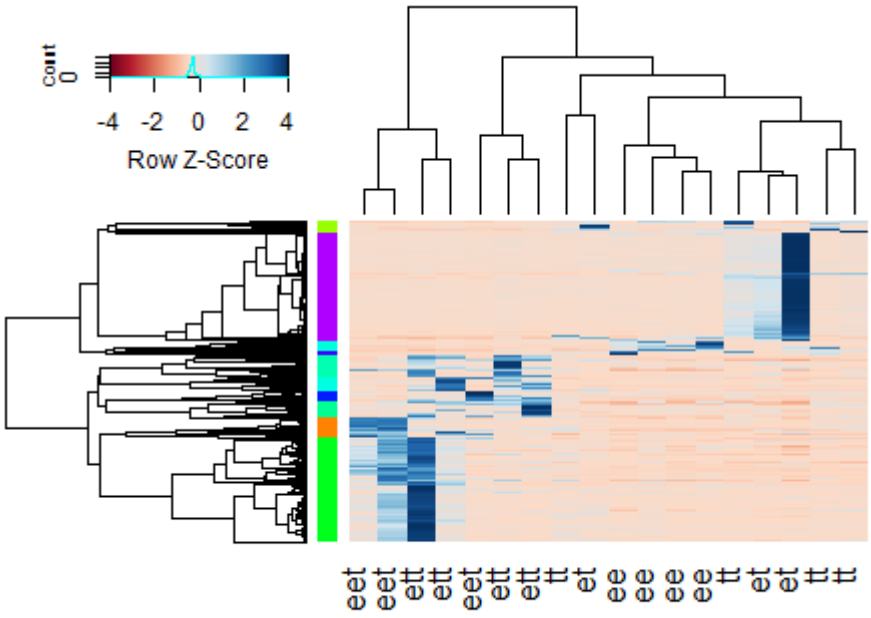
Tzv. heatplots názorně ukazují ne/podobnosti a heterogeneity porovnávaných skupin pro určení DE genů. V řádcích *heatmap* jsou prezentovány TMM normalizované počty *readů* na cDNA, intenzita modré barvy indikduje míru exprese vyjářené formou z-

score. Dendrogramy ve sloupcích jsou vyjádřením hierarchického shlukování – 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. Heatplots tedy slouží jako kontrola experimentu – konkrétně vyjádření podobnosti analyzovaných skupin. Dále toto zobrazení podává informaci o divergenci genů mezi vzorky a genových skupinách sloučených na základě podobnosti v expresi. Podobnost 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.

Jak se ukazuje, ve všech srovnáních kromě mezidruhováho jsou důvody k obavám z falešně pozitivně detekovaných DE genů, které mohou být „taženy“ např. jen jedním vzorkem ze skupiny i přes vysokou varianci ve skupině. Nejvíce jsou variancí ve skupinách zatíženy jaterní vzorky. Bližší pohled na podobnosti DE genů v expresi mezi skupinami di- versus polyploidní jedinci oocytů a jater se naskýtá v grafech č. 32 a 33. Jak je z grafů při srovnání ploidie jasné, naprostá většina DE genů souvisejících s ploidií může být falešně pozitivních, zejména v jaterní tkáni, kde nalézáme mnoho nabohacených genů odpovědných za reakci na stres a imunitní odpovědi. Je zde však vidět, že například skupina sexuálních jedinců není homogenní, ale jasně v ní lze odlišit oba rodičovské druhy, což naznačuje pervazivní efekt druhově specifické exprese (viz. níže). Ze srovnání diploidů a polyploidů jsme nenašli nabohacené žádné KEGG dráhy ani GO. Při pohledu na jednotlivé geny jsme však nalezli mezi DE geny ty, které jsou zodpovědné pro stresovou reakci a imunitní odpověď; konkrétně interleukiny, gama-interfero a heat-shock kódující mRNA. Mezi DE geny obsaženými pouze ve srovnání ploidie u hybridů (jak v jaterní tkáni, tak oocytech) bylo nalazeno několik proliferačních faktorů nabohacených u hybridů. V oocytech nacházíme u polyploidních jedinců podexprimovaný gen *Msx1* důležitý pro vývoj embrya.



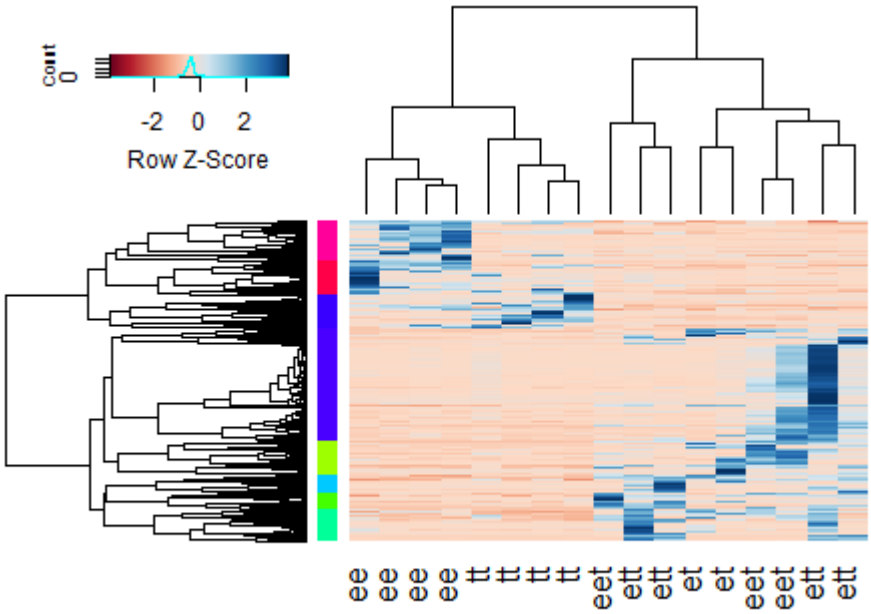
Graf č. 32: Heatplot – zobrazení míry exprese jednotlivých DE genů vzniknuvších komparací **diploidních** a **polyploidních** jedinců vzorků **oocytů** (**s intenzitou modré barvy stoupá míra exprese – normalizovaná data, červená - modrá**). Dendogramy na horizontální ose vyjadřují výsledek shlukovací analýzy korelační koeficientů mezi geny, na vertikální ose vyjadřuje podobnost mezi vzorky.



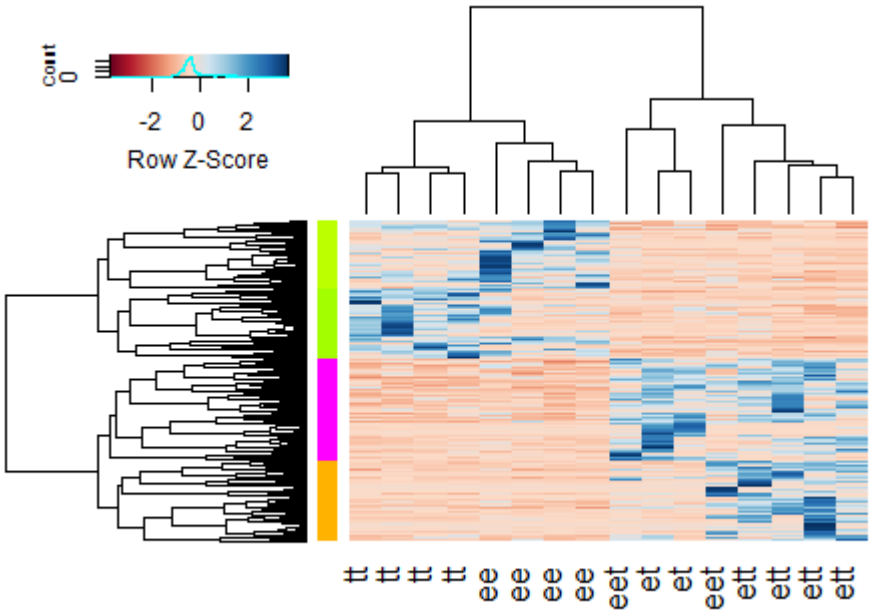
Graf č. 33: Heatplot – zobrazení míry exprese jednotlivých DE genů vzniknuvších komparací **diploidních** a **polyploidních** jedinců vzorků **jater** (**s intenzitou modré barvy stoupá míra exprese – normalizovaná data, červená - modrá**). Dendogramy na horizontální ose vyjadřují výsledek shlukovací analýzy korelační koeficientů mezi geny, na vertikální ose vyjadřuje podobnost mezi vzorky.

V grafech č. 34 a 35 je znázorněna exprese všech signifikantních DE genů vzhledem k typu rozmnožování v oocytech (sex vs asex.) a jater. Opět zde nacházíme jedince, kteří mírně vybočují ve své expresi a mohou deformovat pohled na množství

signifikantní DE genů. Je také opět vidět, že samotná skupina sexuálních jedinců není homogenní, ale jsou v ní vidět polovina mezidruhové rozdíly.



Graf č. 34: 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, červená - modrá). Dendogramy na horizontální ose vyjadřují výsledek shlukovací analýzy korelační koeficientů mezi geny, na vertikální ose vyjadřuje podobnost mezi vzorky.

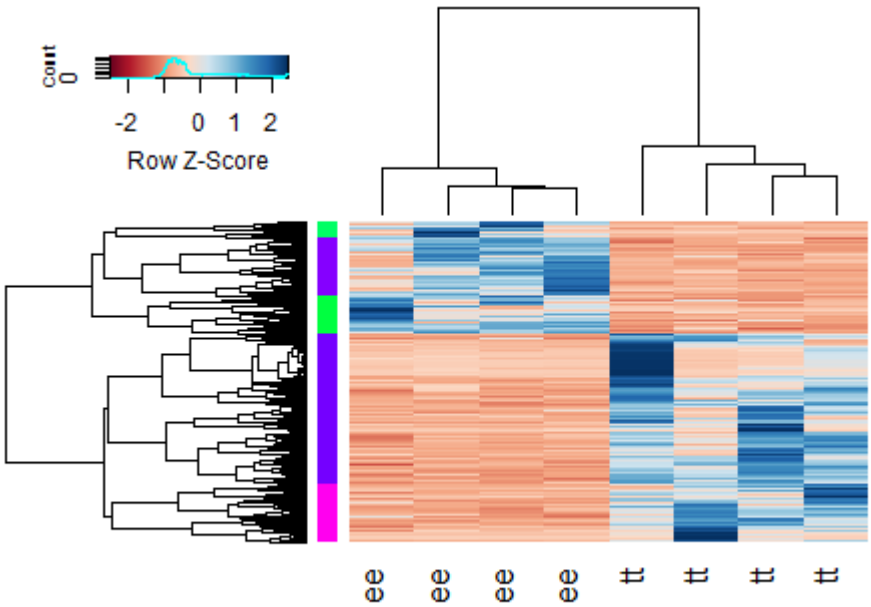


Graf č. 35: Heatplot – zobrazení míry exprese jednotlivých DE genů vzniknuvších komparací **sexuálně** a **asexuálně** se reprodukujících jedinců vzorků **oocytů** (**s intenzitou modré barvy stoupá míra exprese – normalizovaná data, červená - modrá**). Dendogramy na horizontální ose vyjadřují výsledek shlukovací analýzy korelační koeficientů mezi geny, na vertikální ose vyjadřuje podobnost mezi vzorky.

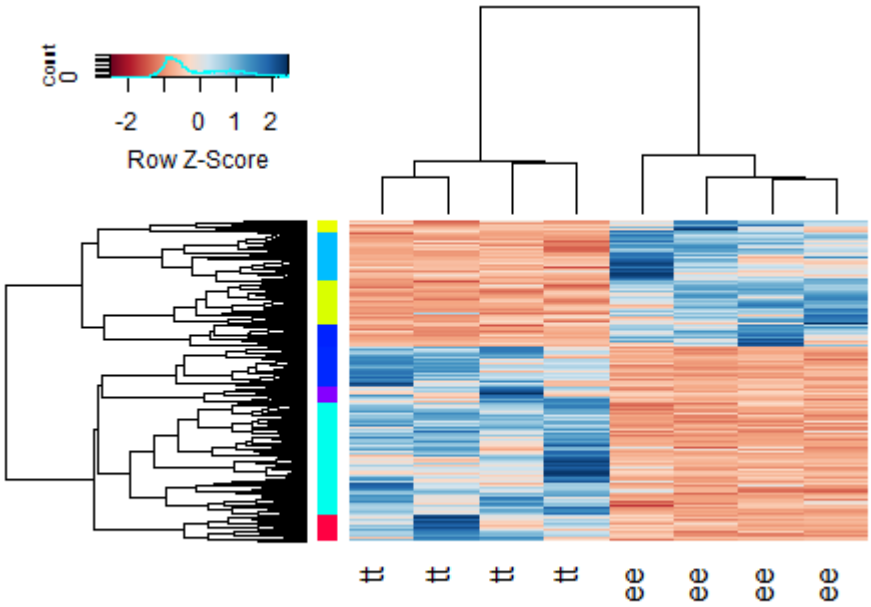
Celkem jsme nalezli 101 anotovaných genů, kde bychom mohli navrhnout funkční koncept z hlediska spojitosti s gynogenetickým rozmnožováním, tj. ty, které jsou DE výhradně při srovnání sexuálních a asexuálních skupin po odfiltrování interferujících kategorií polyploidie a druhově specifických rozdílů. Na jejich základě, nemůžeme implikovat žádné funkční změny, protože jsme neidentifikovali nabohacení konkrétních GO a KEGG drah. Zaměřit se tak můžeme pouze na konkrétní geny. Totéž platí pro jaterní tkáň. Překvapivě jsme z tohoto pohledu individuálních genů pozorovali vliv na dráhu produkující vitelogenin, jež tvoří majoritní část žloutkového vaku (apolipoprotein – zejména transportní funkce, vazba na lipidy, N- doména signálál pro export). Bylo skutečně pozorováno, že hybridní jedinci produkují větší množství žloutku. Připomínám, že mezi analyzovanými jedinci byly pouze samice. V játrech hybridů 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 díky časté incidenci polyploidie.

U posledního typu srovnání jsme hledali DE geny významné na mezidruhové úrovni, konkrétně druhů *tt* a *ee*. Rozdíly genové exprese mezi druhy jsou, jak již výše uvedeno, nejmarkantnější, a hlavně se zde jeví skutečně plně konzistentní mezi skupinami. Heatmap vyobrazení těchto skupin je uveden v grafech č. 36 a 37 a ukazuje celkem jasnou homogeneitu skupin.

DE geny ze srovnání *tt* vůči *ee* jsou zapojeny především v metabolismu - metabolismus lipidů, oxidačně redukční mitochondriální procesy, translaci, transkripci (transkripční faktory), geny asociovány s jaderným prostorem. Tyto geny nacházíme jak ze srovnání oocytů, tak jater. Největší rozdíly byly nalezeny u genů pro cytochrom-oxidázy, ATPázy a vodíkových přenašečů mitochondriálních krist. Opět nacházíme geny imunitní odpovědi genů jak viperin, tool like receptor a komplementy, vyznačující se tím, že jsou nadexprimované u minoritní části jedinců ve skupině. Významné změny exprese jsou zaznamenány také u několika typů lektinů a v poslední řadě cyklinů A2, E1.



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

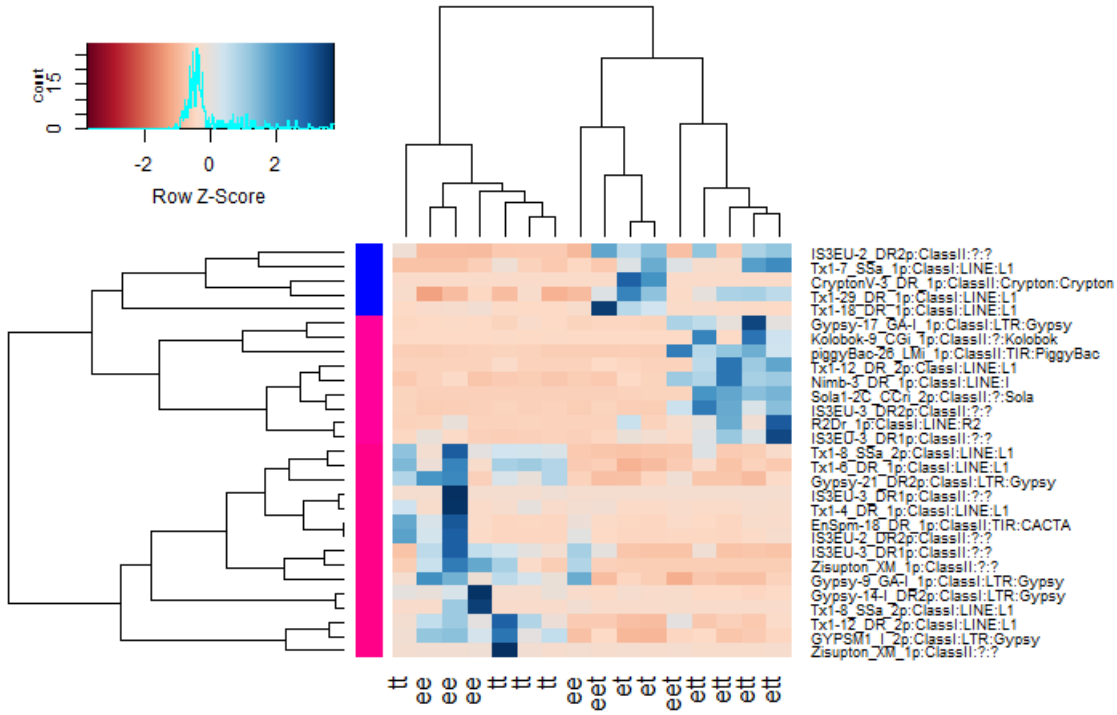


Graf č. 37: Heatplot – zobrazení míry exprese jednotlivých DE genů vzniknuvších komparací **druhů *tt* a *ee*** vzorků **oocytů (s intenzitou modré barvy stoupá míra exprese – normalizovaná data, červená - modrá)**. Dendogramy na horizontální ose vyjadřují výsledek shlukovací analýzy korelační koeficientů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

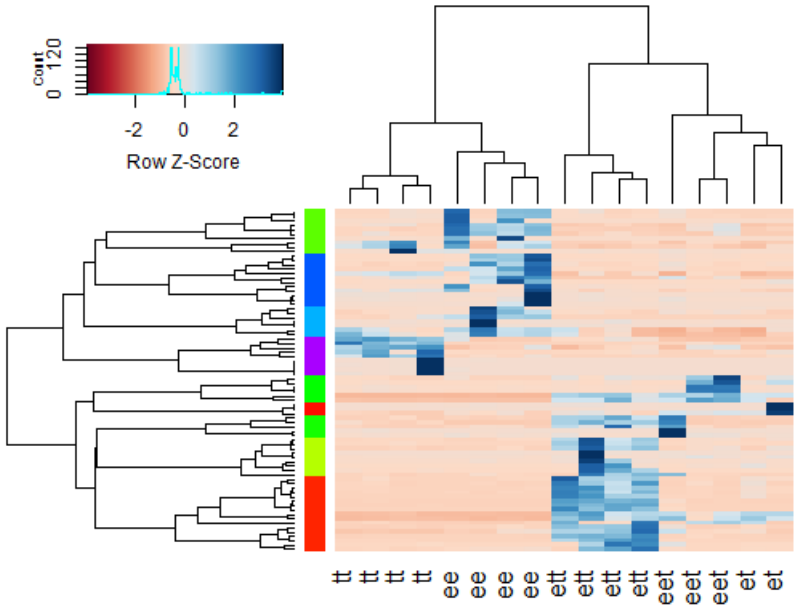
Poslední otázkou, která by si zasloužila konkrétní popis, je exprese transpozabilních elementů. Bohužel nemáme žádné klonální hybridy z F1 generace, tudíž aktivaci TE nelze přímo studovat. Můžeme pozorovat pouze recentní stav klonálních linií.

Srovnáme-li sexuálně a asexuálně se reprodukující jedince, a to včetně hybridů obsahujících genotyp *C. tanaitica*, můžeme pozorovat, že diferenciálně exprimované TE elementy jsou pod- i nadexprimovány rovnocenně u obou skupin (celkem 29 TE v oocytech – 15 nadexprimovaných u hybridů; 78 v játrech, přičemž 38 genů je nadexprimováno u hybridů); nelze tedy tvrdit, že u hybridních linií existuje obecná tendence k výrazně zvýšené expresi TE, viz graf č. 38 Jak u oocytů, tak u jater se jedná zejména o retrotranspozony třídy I a II. Z grafu č. 39 můžeme vypozorovat značné mezidruhové rozdíly v „leaky“ expresi TE. Ač ne nijak výrazně TE jsou nejvíce exprimovány u jedinců biotypu *ett*.

Vzor exprese je podobný taktéž mezi di- a polyploidními jedinci, kde jsme nalezli 18 TE u oocytů z čehož je 15 nadexprimováno u diploidů. V játrech jsme nalezli 59 DE TE, z toho 47 nadexprimovaných u diploidů



Graf č. 38: Heatplot znázornění DE genů **oocytů** a to pouze TE (bez *n* genomu)

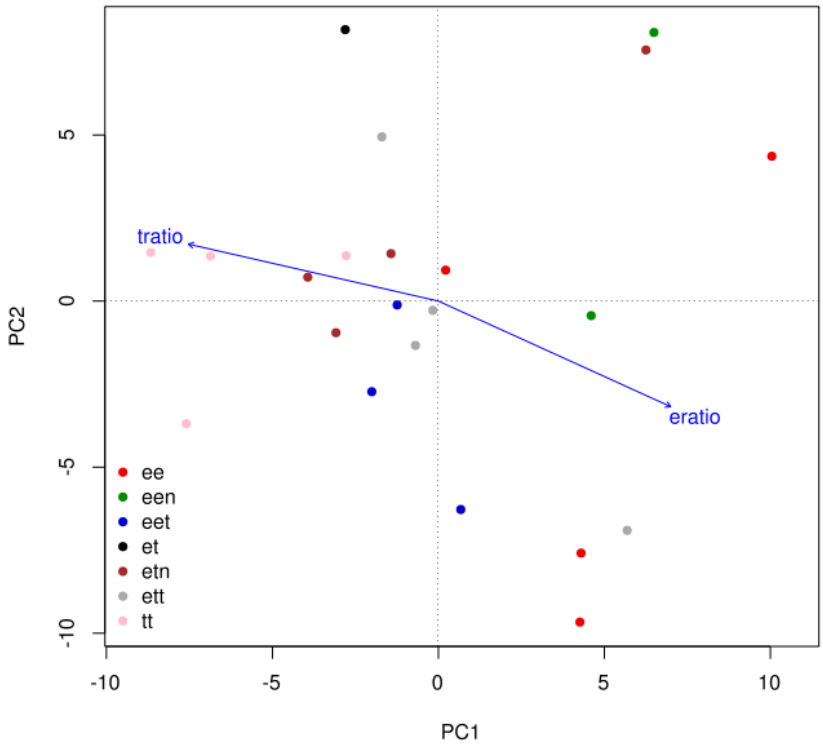


Graf č. 39: Heatplot znázornění DE genů **jater** a to pouze TE (bez *n* genomů)

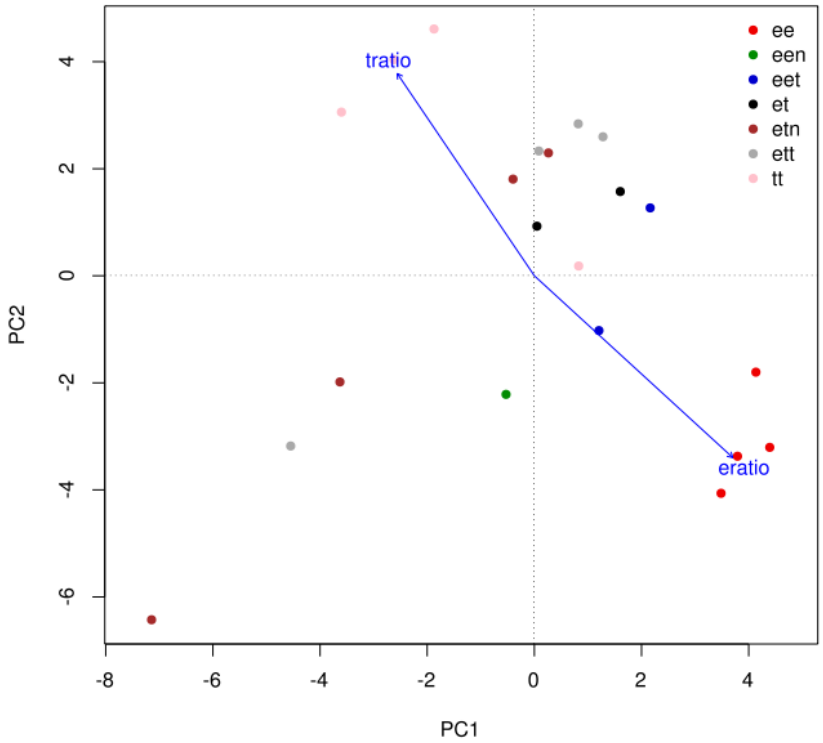
4.3 Imprinting hybridních genomů

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) modelu gradientu největších rozdílů mezi *tt* a *ee*. Signifikance byla provedena 1000 násobnou permutací (přístup permutační annovy). V rda grafech č. 40, 41 je zobrazen každý analyzovaný vzorek na základě exprese všech genů. Koordináta bodů mezi komponentami vyjadřuje podobnost mezi vzorky. Environmentální vektor exprese druhů *ee* a *tt nafitovaný* na data je znázorněn modrou čarou (*e-ratio*, *t-ratio* naznačují směr, ve kterém se nejvíce projevují znaky charakteristické pro druh ee, respektive tt). Jelikož některé vzorky obsahují navíc haplotyp *n*, nesvírá tato přímka (*e-ratio*, *t-ratio*) 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*. Nafitování modelu směru *tt* i *ee* je signifikantní v obou případech na α 0.001.

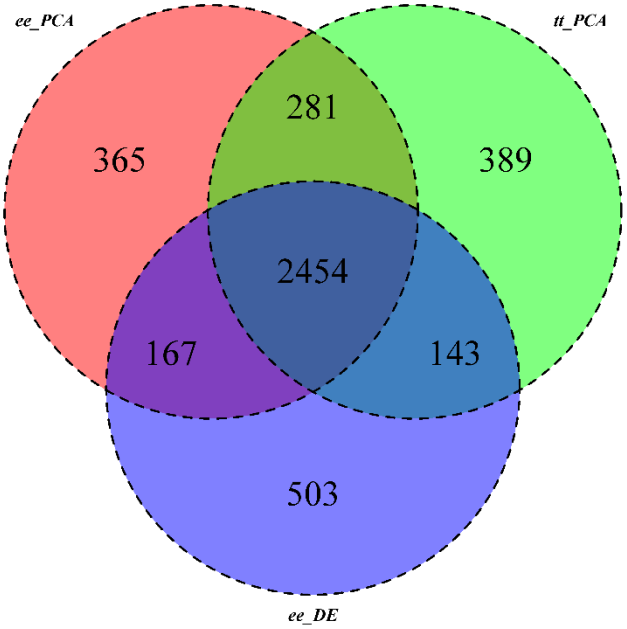


Graf č. 40: *Nafitovaný* environmentální model podle druhů *tt* a *ee* (zde nazváno jako t-ratio, e-ratio) na rda shlukovou analýzu vzorků **jater**. PCA1 vysvětluje 21.13% variability, PCA2 13.88% (Mrg. Ladislav Pekárik, Ph.D., 2015).



Graf č. 41: *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 *e-ratio*, tak i *t-ratio*. Touto rotací jsme získali nové osy rovnoběžné s e-ratio, respektive t-ratio a geny byly seřazeny podle svého významu vůči ose gradientu (absolutní hodnota "přerotované" koordináty genu). Význam tkví především 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. 42 vyplývá, že DE geny mezi definovanými skupinami druhů přispívají taktéž k diferenci mezi druhy osy gradientů.



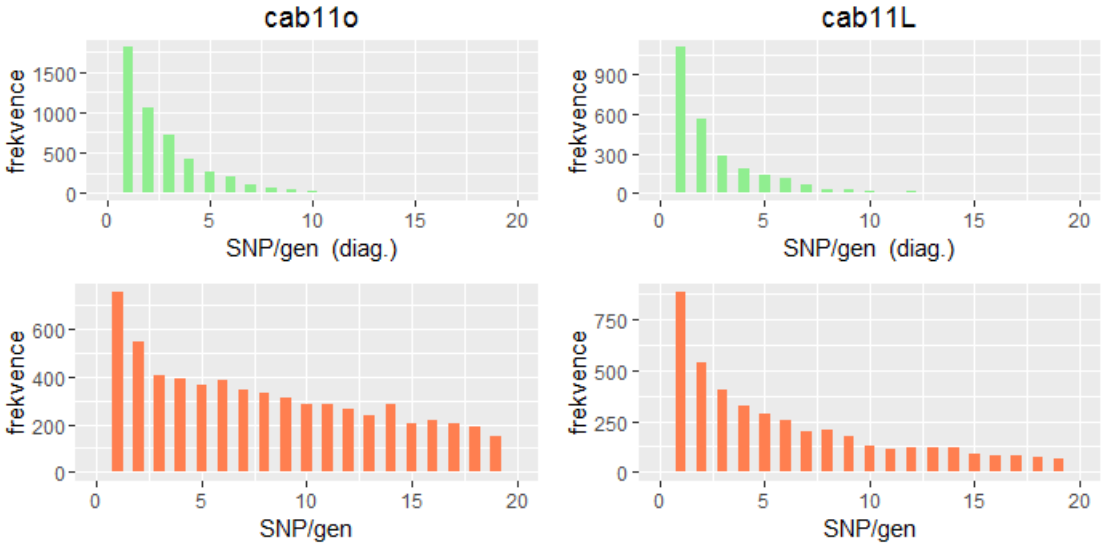
Graf č. 42: Vennův diagram tři množin: geny přispívající ke gradientu osy *e-ratio*, geny přispívající ke gradientu osy *t-ratio* a množina všech nalezených DE genů mezi druhy *tt* a *ee*.

Z grafu č. 42 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 *nafitovanému* gradientu t-ratio/e-ratio na expresní data.

Alelově specifická exprese:

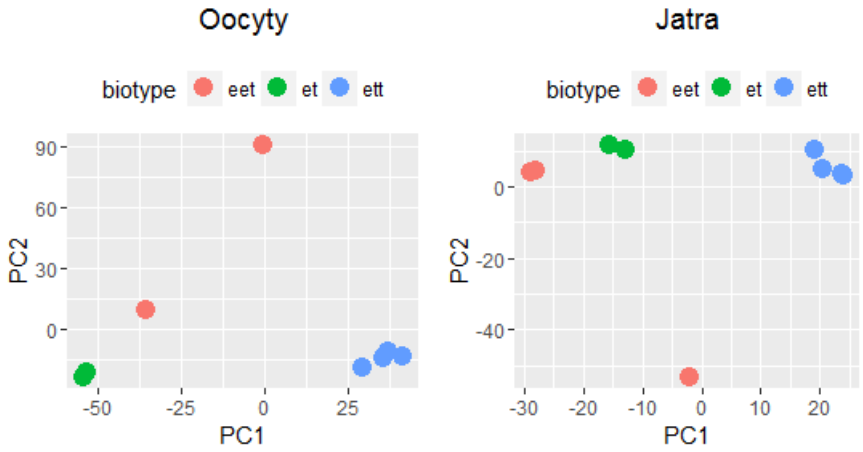
Následující část výsledků se bude upínat k otázce alelové exprese u hybridních jedinců, čili zda je jeden z genomů dominantní ve smyslu regulace exprese, tj. je jeden z genomů hybrida transkripčně umlčen, a pokud takové geny existují, jaký je jejich biologický význam. Připomínám, že nemůžeme analyzovat hybridy biotypu *etn* a *een*, protože jsme neanalyzovali rodičovský druh *C. tanaitica*.

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. V grafu č. 44 znázorňuji četnost genů v kategoriích dle počtu diag. SNP u jedince cab11.



Graf č. 44: Znázornění kategorií - počtu SNP na gen a jejich frekvence, a to pro diagnostické (zelená) a všechny SNP (červená) **u jedince cab11o 7553 genů nebyly nalezeny diag. SNP a 4897 us cab11L**. Ve zbytku genů nebyly nalezeny SNP.

Podobnost mezi vzorky můžeme vyjádřit formou PCA a to na základě P hodnot jednotlivých SNP, přičemž podle směru vychýlení je P hodnotě změněno znaménko (mínus *tt*, plus *ee*), viz graf. č. 47. Nutno zdůraznit, že sekvenační holoubka na SNP pozici je normalizovaná TMM metodou, abychom se opět zbavili faktoru sekvenační hloubky a mohli tak vzájemně porovnat jedince, aniž bychom přiřazovali jedincům s vyšší sekvenační hloubkou signifikantnější P hodnotu. Zaměřujeme se pouze ty jedince, kteří jsou uvedeni v DE analýze – srovnání sex *vs* asex.



Graf č. 47: PCA dvou hlavních komponent srovnání vektorů P hodnot všech diagnostických SNP vzorků jakter a oocytů. Sekvenační hloubka SNP je normalizovaná TMM metodou, abychom P hodnotu nevychylovali sekvenační hloubkou.

V tabulce č. 12, 13 znázorňuji počty diagnostických SNP, jež bylo možné analyzovat. Na základě těchto dat navrhujeme, že atenuace není mezi hybridními genomy rovnoměrná, nýbrž exprese *e* alel je častěji potlačena, jak můžeme vidět na příkladech diploidních hybridů. ASE je zjevně závislá na počtu kopií alel jak ukazuje směr ASE triploidních hybridů.

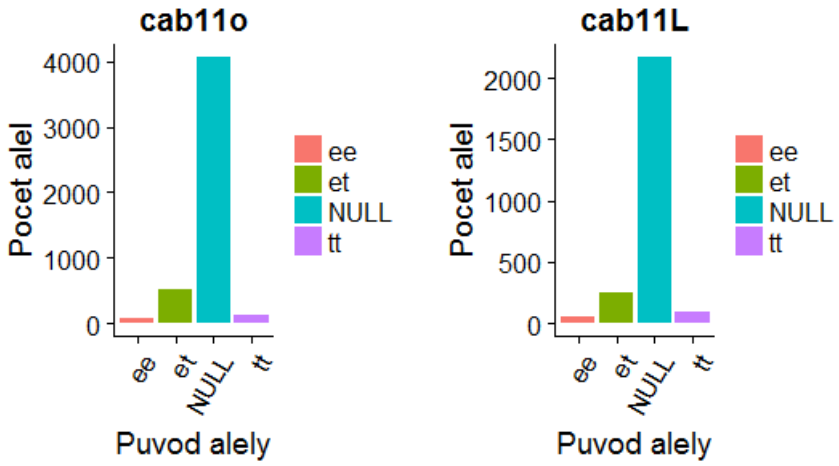
vzorek	biotyp	analyz. SNP	ASE - *ee*	ASE - *tt*
cab07L	*eet*	10827	1060	430
cab08L	*eet*	12027	1186	448
cab09L	*eet*	13036	4987	397
cab11L	*et*	6231	233	650
cab16L	*ett*	6235	96	1337
cab17L	*ett*	15574	184	2552
cab22L	*et*	15647	954	1002
cab23L	*et*	6855	478	648
cab24L	*ett*	11250	168	2031
cab25L	*ett*	9285	149	1713

Tab. č. 12: Srovnání počtu všech analyzovatelných SNP jaterních vzorků, validních SNP a alelově specifických genů exprimovaných buď genomem *ee*, nebo *tt*. Případy, kdy došlo ke konfliktu disbalance v rámci genů, byly označeny jako problematické.

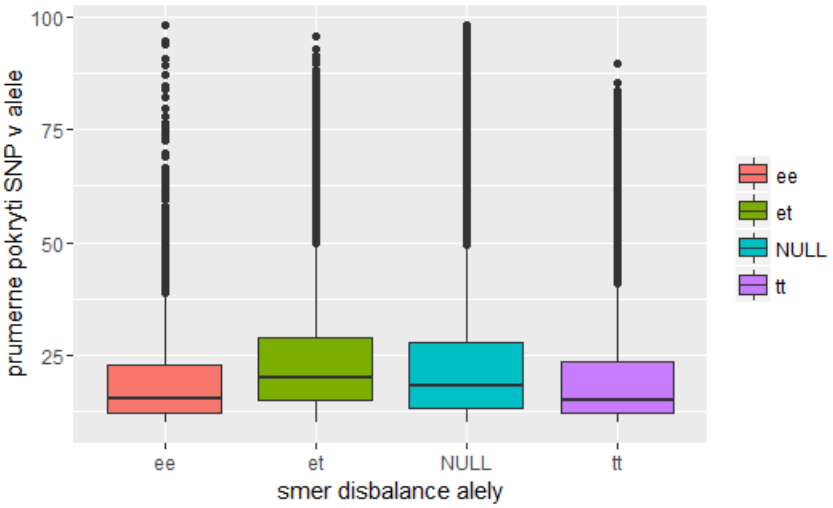
vzorek	biotyp	analyz. SNP	ASE - *ee*	ASE - *tt*
cab07o	*eet*	22306	1844	553
cab08o	*eet*	5048	574	290
cab09o	*eet*	22733	8243	603
cab11o	*et*	13329	404	985
cab16o	*ett*	17530	245	3111
cab17o	*ett*	19747	206	3249
cab22o	*et*	17174	548	1150
cab23o	*et*	7643	258	683
cab24o	*ett*	18634	210	2942
cab25o	*ett*	18147	228	2833

Tab. č. 13: Srovnání počtu všech analyzovatelných SNP vzorků z oocytů, validních SNP a alelově specifických genů exprimovaných buď genomem *ee* nebo *tt*.

Budemeli se zabývat ASE geny bez toho aniž bychom stanovili arbitrárně kolik SNP musí obsahovat gen můžeme např. v oocytech diploidního jednice cab11 nalézt 4729 genů, které obsahují alespoň jeden diagnostický SNP (viz graf č. 43) a 1930 genů obsahující 3 a více diag. SNP. Z těchto genů je jich 83 vychýleno ve směru genomu *ee* a 134 ve směru *tt*. Zbytek - 1832 genů neobsahuje diagnostické pozice;. Z posledně jmenovaných ale mělo 409 genů kontradiktorní stav, kdy alespoň jeden SNP byl vychýlen opačným směrem, než zbytek Disbalance genu ve směrou obou genomů bohužel nelze vysvětlit pouze nedostatečnou sekvenační hloubkou, protože v těchto problematických případech je paradoxně vyšší, viz graf. č. 45.



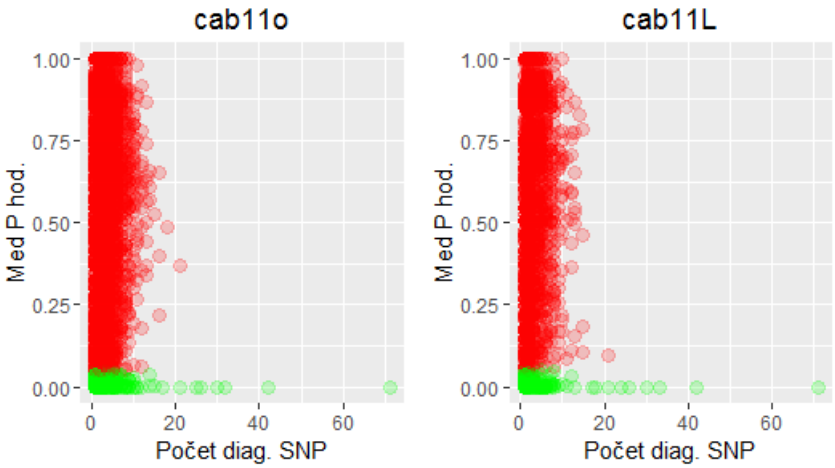
Graf. č. 43: Znázornění počtu genů expresně vychýlených k rodičovským druhům na příkladu jedince genotypu *et*, v grafu zeleně vybarvené geny (**et**) značí geny mající SNP se signifikantním vychýlením k oběma parentálním genomům, kdy alespoň jedna pozice je v kontradikci s ostatními pozicemi v genu. Geny, které nemají ani jedenu signifikantně disblancovanou SNP pozici, jsou označeny jako ***NULL***, nicméně i tyto **SNP jsou druhově diagnostické**!



Graf č. 45: Znázornění rozdělení průměrného pokrytí přes všechny nalezené SNP v genu. Kategorie označují směr ASE genu tak jako v případě grafu č. 43 Označení ***NULL*** se opět týká SNP, u kterých jsme statisticky nezaznamenali disbalanci, ale jsou druhově diagnostické.

Problém nám ale nečinila jen absence diagnostické pozice v genu, ale též směr disbalance. Jak je znázorněno v grafu č. 43, velmi často se totiž stalo, že v genu je alespoň jedna pozice, která se jeví jako disbalancovaná ve směru druhého parentálního genomu, než zbytek signifikantně disbalancovaných SNP, což biologicky nedává smysl, ledaže by došlo k rekombinaci pouze části sekvence alely.

Množství detekovatelných, diagnostických SNP pozic v genech exponenciálně klesá. I přes tento fakt jsme se rozhodli pro redukci dat z důvodu nebetyčného množství falešně pozitivních SNP disbalancí a zaměřili se jen na ty geny, které měly více než 3 detekovatelné, diagnostické SNP. V případě, kdy provádíme globální srovnání z pohledu SNP, neredukujeme data o *kontigy* s nižším počtem diagnostických SNP. Další možný problém způsobený technickou variabilitou znázorněný v grafu č. 46 je v genech s velkým počtem diagnostických SNP – medián vychýlení genu s vyšším počtem diagnostických SNP může být falešně pozitivní. Bohužel tento jev nedokážeme vysvětlit. Srovnáme-li distribuce počtu diagnostických SNP s P hodnototou > 0,05 a P hodnotou <= 0,05, zamítáme hypotézu o rovnosti jejich průměru (P hodnota 0,0001145, jedinec cab11o). To znamená, že s narůstajícím počtem SNP, je alela častěji označena jako disbalancovaná.



Graf č. 46: Srovnání počtu diagnostických pozic v genu a mediánu P hodnot těchto pozic na příkladu diploidního hybrida *et*. V genech s více než 5 diagnostickými pozicemi je medián P hodnot vždy signifikantní (zelená = medián P hodnot < 0.05; červená = medián P hodnot >= 0.05)

Nakonec je nezbytné dodat, že schopnost detekovat disbalanci je spojená se sekvenační hloubkou v konkrétních genech, protože slabší signál represe alely nemůže být identifikován díky nárůstu vlivu technické sekvenační variability. V tomto případě nulovou hypotézu spíše nejsme schopni zamítnout.

V následujícím textu se dále zaměřím na výsledky samotné detekce signifikantně disbalancovaných genů. Za signifikantně disbalancované alely budu považovat pouze ty, který nesly alespoň 3 SNP, u nichž byla pomocí binomiálního testu zamítnuta nulová hypotéza po provedení sekvenční FDR korekce. U většiny genů jsem bohužel nebyl schopen detekovat směr disbalance, protože často byl jeden z diagnostických SNP vychýlen i k druhému rodičovskému genomu. Z následujících tabulek č. 10 a 11 je patrné, kolik SNP bylo detekováno jako průkazně vychýlených a jakým směrem u jednotlivých vzorků a kolik z celkového počtu genů tvoří tzv. problematické SNP.

vzorek	biotyp	>= 3 diag. SNP	ASE - *ee*	ASE - *tt*	probl. alela
cab07L	*eet*	1474	28	5	418
cab08L	*eet*	1681	34	4	478
cab09L	*eet*	1793	199	5	1179
cab11L	*et*	896	10	13	195
cab16L	*ett*	820	1	35	367
cab17L	*ett*	2212	4	53	857
cab22L	*et*	2239	44	19	483
cab23L	*et*	922	30	10	238
cab24L	*ett*	1536	0	46	632
cab25L	*ett*	1234	2	38	510

Tab. č. 10: Srovnání počtu všech genů jaterní tkáně s 3 a více validními diagnostickými SNP, exprimované ve směru genomu *ee* nebo *tt*. Případy, kdy došlo ke konfliktu disbalance v rámci genů, byly označeny jako problematické.

vzorek	biotyp	>= 3 diag. SNP	ASE - *ee*	ASE - *tt*	probl. alela
cab07o	*eet*	3303	65	9	662
cab08o	*eet*	643	12	4	186
cab09o	*eet*	3366	343	3	2208
cab11o	*et*	1930	6	19	409
cab16o	*ett*	2579	1	70	1028
cab17o	*ett*	2896	1	78	1122
cab22o	*et*	2508	16	19	485
cab23o	*et*	995	8	18	211
cab24o	*ett*	2719	1	75	1005
cab25o	*ett*	2653	2	68	981

Tab. č. 11: Srovnání počtu všech genů oocytů s 3 a více validními diagnostickými SNP, exprimované ve směru genomu *ee* nebo *tt*. Případy, kdy došlo ke konfliktu disbalance v rámci genů, byly označeny jako problematické.

Z tabulek je zřejmé, že například diploidní hybridi a triploid typu *ett* měli opět tendenci k častější disbalanci ve prospěch genomu *tt*, zatímco oba triploidi typu *eet* častěji nadexprimovali genomy druhu *ee*. Výběr také zahrnuje různé směry hybridizace, jelikož 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.

4.4 Degenerace hybridních linií - Müllerova rohatka

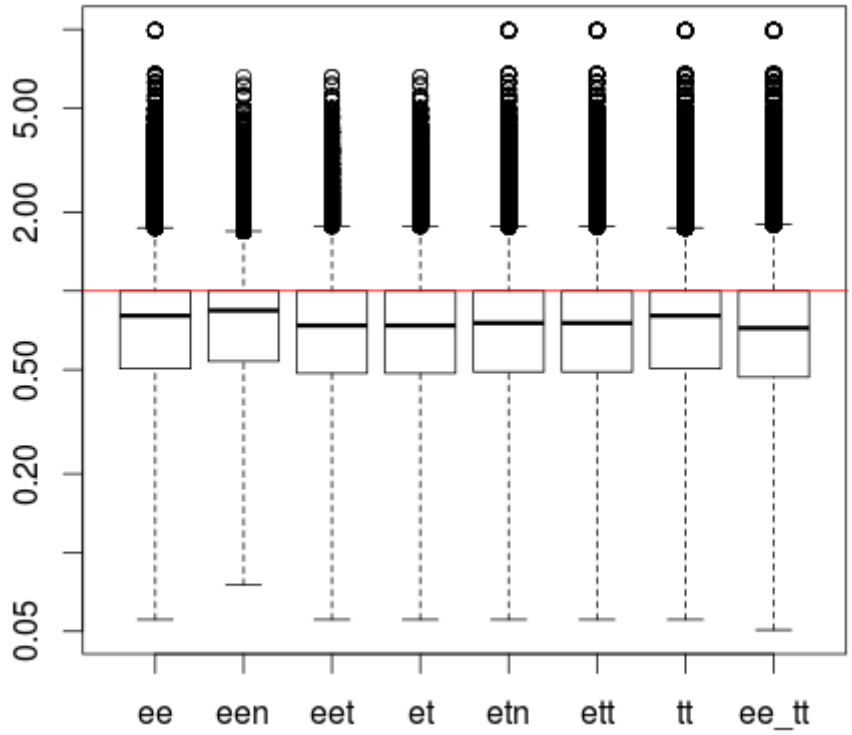
V poslední řadě budou prezentovány ústřední výsledky této práce – odpověď na otázku, jak jsou ovlivněni hybridní jedinci z pohledu evolučního, jakým tempem dochází v genomu hybridů reprodukujících se pouze klonální cestou k akumulaci nesynonymních mutací (měření poměru nesynonymních / synonymních mutací z párové komparace), neboť nedochází k rekombinaci mezi genomy rodičů, a tudíž ani k náhodnému, efektivnějšímu odstraňování selekčně nevýhodných, nesynonymních mutací.

Hybridní jedinci neobsahují v ORF více stop kodónů, což odpovídá představě, že konzervované ochranné mechanizmy jako *nonsence mediated decay* brání expresi nefunkčních peptidů (Chang et al., 2007); naopak ORF se stop kodóny vzniklo více především u jedinců druhu *C. elongatoides*. To můžeme připisovat spíše chybám v mapování, detekci SNP a v poslední řadě několik genů nemusí být pod stejným selekčním tlakem. Absolutní počty stop kodónů na jedince jsou uvedeny v grafu č. 20 a 21.

Z výsledků SNP RNA sekvenování je patrné vychýlení dN/dS poměru *in silico* vytvořených *et* hybridů směrem k nižším hodnotám. Nízký poměr dN/dS pochopitelně vykazují také srovnání mezi rodičovskými druhy *ee* a *tt,* viz graf č. 65 a tabulka č. 11. 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. 14; 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

	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. 14: Párový Wilcoxnův test, neparametrické testování rovnosti průměrů



Graf č. 48: Distribuce poměru dN/dS vypočítaného ze srovnání vůči *in silico* hybridním jedincům a také mezi druhy samými jaderných loci.

5 Diskuse

Data použitá v analýzách jsou technicky různě zpracovaná. Používáme normalizovaná 454 sekvenovanou cDNA data pro tvorbu reference a nenormalizovanou cDNA sekvenovanou technologií illumina. 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 anotací (*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*. 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 ambiguidní 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. Výše uvedené tedy dokazuje maximální úsilí po co nejkvalitnějším assembly referenčního transkriptomu, ale zároveň i fakt, že si případné obtíže uvědomuji a počítám s nimi v další intrepretaci.

SNP byly získány v případě 454 standartním a illumina způsobem, byl užit software vhodný pro mapování cDNA na transkriptom. Bohužel jsou nezbytné další analýzy pro odstranění zjevně chybně identifikovaných SNP, které způsobily vznik velkého množství stop kodonů. 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 diferenciálně exprimované geny pro stanovení korelace mezi relativní expresí zjištěnou qPCR a RNAseq, která se jeví, jako relativně těsná což dokazuje, že naše nextgen sekvenační data jsou v podstatě správně analyzována.

Detekce diferenciálně exprimovaných genů:

Samotnou příčinu vzniku gynogenetického rozmnožování nejsme schopni z dat diferenciální exprese identifikovat, a to, jak bylo řečeno, z triviálního důvodu. Zaměřili jsme se z objektivních důvodů na analýzu rozdílů mezi oocyty 6. vývojového stádia, nikoliv zárodečné tkáně vaječníků, a sledujeme tak pouze výsledek procesu, 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 byl nalezen u skupiny asexuálních jedinců podexprimovaný gen pro rekombinázu 8 a nabohacený cyklin A2 a E1, neboť předpokládáme, že mRNA by měla být po ukončení meiózy I odbourána.

Jak vyplývá z grafu č. 29, 30 překryv s geny z mezidruhového srovnání a skupiny di- vůči polyploidním vzorkům je značný, což znamená, že je často dost těžké odlišit efekt té které vlastnosti studovanýc organismů. Konečně fakt, že efekt ploidie, hybridizace a asexuality na evoluci podobných mezidruhových komplexů se dá obtížně odlišit, je notoricky známý (Kearney, 2005). Bohužel ani jejich odfiltrováním genů, u nichž máme překryv efektu rozdělení na sexuály vs. gynogeny s efektem ploidie, mezidruhových rozdílů, nemůžeme s jistou určit, zda se jedná o funkční podmínku pro umožnění

gynogenetického rozmnožování studovaného hybridního komplexu. V následujících několika odstavcích se zaměříme na diskusi zjištěných DE genů v souvislosti s jinými publikovaným případy.

Zajímavé je, že jsme mezi zjištěnými asex-specifickými DE geny nalezli překryv s DE geny identifikovanými v hybridním komplexu karas-kapr (Li et al., 2014). Ačkoliv existuje shoda mezi geny, nelze ji 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, vezmeme-li v úvahu, že se jedná o zcela nezávislý vznik gynogenetických forem.

V případě srovnání výsledků s rostlinnými apomiktickými modely nacházíme shodu v nabohacení genů u klonálně se reprodukujících jedinců pro regulaci buněčného cyklu, cyklinů, metylačních komplexů, chromodomény a histonových variant (v našem případě H3.3 – podexprimován u hybridů) (Galla et al., 2015). Pozn.: Histon H3.3 je spojen s chromatinem transkripčně aktivních genů.

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 znám 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. Zajímavou otázkou je nárůst exprese u hybridních, klonálních jedinců mRNA SRY, který je přirozeně lokalizován na Y chromozomu, jeho regulační funkce spočívá v aktivaci drah vedoucích k vývoji testes (upozorňuji, že všichni analyzovaní jedinci jsou samice). Recentně nebylo něco takového u asexuálů publikováno, zřejmě se jedná o artefakt.

Mezi geny, které mají logickou přímou spojitost s gynogenetickým rozmnožováním, je glykoprotein zona pelucida bránicí vzniku polyspemického embrya. Biologicky diferenciální regulace tohoto komplexu u asexuálů dává smysl, neboť interakce

vajíčka s okolím je pro gynogena klíčová. Doposud se ví jen velmi málo o tom, jak konkrétně je gynogeneze u různých organismů řízena a otázkou je i to, zda aby došlo k aktivaci vývoje, musí spermie vniknout v případě gynogenetického rozmnožování do oocytu, či jen pouze externě aktivovat rýhování neoplozené zygoty. Empiricky je známo, že celkem často spermie do vajíčka nejen vniká, ale i přímo geneticky přispívá ke vzniku embrya (Copernicus, 2003; Janko et al., 2007; Juchno et al., 2015), což ovšem sebou nese i to, že výsledkem je zvýšení ploidie. Zvýšení z diploidie na triploidii nepředstavuje zásadní problém alespoň z hlediska ontogeneze, ale oplozené triploidní gamety dávají vznik neživotaschopným tetraploidním formám (Janko et al., 2007, 2012; Juchno et al., 2015). Dochází-li k inkorporací genomu spermie k zániku takových embryí, pak je taková situace z hlediska hybridních samic selekčně nevýhodná a případná přestavba zony pelucidy může být žádoucí. Otázkou je, zda k takové přestavbě došlo následkem hybridizační transkripční disregulace nebo pod vlivem pozitivní selekce.

Mezi DE geny, které mohou souviset s přechodem ke gynogenezi, jsou též komponenty cytoskeletu, protože se též významně podílí na fertilizaci. U hybridních jedinců byly 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).

Pozorovali jsme také nabohacenou mRNA pro proteiny vápenatých a draselných kanálů ve vzorcích hybridních oocytů, 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). Nabohacení genů pro vápenaté a draselné kanály, včetně aktivačních drah fertilizace hybridů tak ukazuje na potenciálně důležité funkční aspekty související s gynogenezí, jež by si zasloužily hlubší analýzu.

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. Na základě párové komparace jednotlivých typů hybridů mezi sebou jsme ale došli k závěru, že významnou roli nehrají, neboť nebyl zaznamenán exces DE genů mezi skupinami hybridů; majoritu DE genů můžeme označit opět za mezidruhové diference a to v závislosti na počtu kopií alel parentálních genomů. Nejvíce DE genů pak bylo pochopitelně nalezeno mezi skupinami triploidů *ett* vůči *eet*, mezi kterými je pozorovatelný nejznatelnější nárůst počtu DE genů. Domníváme se, že dochází v hybridních jedincích k "zprůměrování" regulace cis exprese mezi rodičovskými haplotypy.

Jedním z fenoménů intenzívně studovaných u hybridních, klonálně se reprodukujících typů je deregulace epigenetického aparátu, zejména pak demetylace transponovatelných elementů. Především "mladé hybridní linie" zažívají doslova explozi aktivity TE (Dion-Côté et al., 2014; Wu et al., 2016). Postupem času ale dochází pod selekcí k homogenizaci genomu např. vlivem genových konverzí, které mohou způsobit i jejich útlum, aktivace TE skrze hybridizaci je často spojována se speciací (de Boer et al., 2007; Chen and Ni, 2006). 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 (Feschotte, 2008). 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 transpozibilních elementů je spojena s hybridizací, polyploidizací. TE mohou být aktivovány stresem – různé environmentální faktory. Dráhy

spojené s indukcí stresu a aktivací TE se velmi podobají. Tento je byl nazván jako polyploidizační šok (Guerreiro, 2014).

Ačkoliv jsme udělali maximum pro test výše zmíněných hypotéz a k referenci přidali i *kontigy* složené na základě hybridních genomů tak, abychom detekovali případné TEs specifické pro asexuální formy, nenalezli jsme žádný doklad toho, že by hybridi, či polyploidi měli nějakou obecnou tendenci k *upregulaci* či *downregulaci* TE. V expresi TE se spíše liší druhy *C. taenia* a *C. elongatoides* a hybridní linie se vyznačují tím, že zachovávají spíše průměrnou hladinu exprese snad vyjma biotypy *ett,* kde jsme nalezli několik konzistentně exprimovaných TE, které u rodičovských druhů exprimovány nejsou. Aktivace TE při vzniku klonální linie lze ověřit pouze na genomu, a to nejlépe cytogenetickými technikami jako FISH, nebo NGS sekvenování knihoven *mate pair* (k detekcím rekombinace), či celogenomovým sekvenováním a to jak rodičovských druhů, tak hybridů.

Tímto bych rád přešel od diskuse diferenciální exprese jednotlivých zajímavých skupin genů k popisu globální úrovně exprese. Z grafů č. 41, 42 PCA s *nafitovanými* gradienty druhově specifické exprese, vyplývá, že hybridní formy jsou spíše průměrné mezi rodičovskými druhy 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, zatímco triploid je podobnější tomu druhu, od nějž obdržel dvě genomové sady.

To, že hybridní jedinci se projevují expresně průměrně mezi rodičovskými druhy, také implikuje, že ani jeden rodičovských genomů není globálně inaktivován – dědičně metylován, což jinak 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 variability hybridních linií. Epigenetická modifikace jsou přičítány spíše hybridizaci nežli polyploidizaci (Salmon and Ainouche, 2010).

Naše data ukazují spíše jiný vzor – hybridi exprimují geny celkem vzato striktně podle genové dávky, kterou zdědili od rodičů. Velmi zajímavé je, že tato jejich průměrnost, tedy aditivita exprese rodičovských genomů, je patrna i na jiných, vyšších úrovních ontogeneze. Gradient závislosti genotypu byl vysoce průkazně nalezen i z

pohledu morfologické analýzy (Mgr. Miroslav Pertýl Ph.D., Mgr. Janek Kotusz, Ph.D., 2015, nepublikováno). Co je ještě zajímavější, pozorovaná aditivita v závislosti na genové dávce byla patrná i na tak vysoce komplexní úrovni jako jsou interakce organismu s prostředím. Mgr. Ladislav Pekárik, Ph.D. studoval environmentální mikrohabitatové preference sekavčích druhů a typů hybridů a opět se na gradientu genotypů *tt* a *ee* ukázalo, že hybridi okupují spíše intermediární mikrohabitatové niky v závislosti na jejich genotypu (nepublikováno).

Parentální imprinting a druhově specifická exprese genů u hybridů:

V následující části prozkoumáme, zda existují imprintované alely v hybridních jedincích, jakým způsobem změny korelují mezi jednotlivými typy hybridů a jaké funkční kategorie genů byly atenuací 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 limitující 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 také nezanedbatelné, mnohé vycházely ze samotné podstaty dat, jiné byly v podstatě typické pro mnohé nemodelové organismy. Za prvé, kvalitní pročtení druhově diagnostických SNP jsme získali jen u relativně malého procenta studovaných genů. Tento fakt je dán hlavně tím, že mnohé geny měly nízké pokrytí způsobenou jejich relativně nízkou mírou exprese a také tím, že genetická distance mezi druhy není nijak velká. Vzhledem k tomu, že tytéž geny mohly býti dost různě exprimovány mezi zvířaty (viz předchozí kapitola), snížil se ještě více počet genů, u nichž jsme mohli testy alel-specifické exprese porovnat mezi vzorky (u málo exprimovaných genů jsme prostě analýzu ASE nemohli udělat). 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 navazujících studiích bude nezbytné biologicky zajímavé geny validovat tradičními molekulárně biologickými technikami jako alelově specifickou amplifikací. Ze statistického hlediska byla data analyzována zcela adekvátně, ačkoliv mohlo být i přistoupeno k baysiánské statistice a distribuce likelihood pro navržené modely mohly vycházet z vhodnějších distribucí (konjugace) pro naše data (Skelly et al., 2011), nemluvě o zpřesnění prior hodnot iterativním přístupem jako např. EM (*Expectation Maxmization algorithm*) (Munger et al., 2014) pro zajištění vyšší robustnosti testování. Toto jsem již bohužel nestihl a budu se to snažit v navazujícím studiu.

Obzvláště důležité je si uvědomit, že některé SNP či 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. Je totiž nezbytné mít na paměti, že v těchto extrémních případech nejsme schopni 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, ale v současnosti na ní pracujeme za použití technik *Sequence Capture*.

Úroveň alelických disbalancí mezi studovanými hybridními modely obecně se značně liší. Procentuální zastoupení disbalancovaných alel se může pohybovat od 73 % nalezených 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 a změnou genové dávky, či vychýlení disbalancovaných alel není reálné, jelikož studie se zaměřují zejména na F1 generace, zatímco naše hybridní linie se sice obecně jeví jako „F1“, ale značného stáří.

Přes masivní sekvenační úsilí se vlastně ukazuje, že genů vůbec vhodných pro náš test je jen zlomek z celkového počtu. Pokud chceme srovnávat více typů hybridů najednou, toto množství se ještě 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, že z analyzovaných dat, které by měly být náhodně vybrány, je náš soubor výsledků ASE je reprezentovatelný pro stanovení celkové úroveň ASE, ačkoliv většina genů neobsahuje žádný, či SNP, které nejsou druhově diagnostické. Hlavním poselstvím ASE je, že u diploidních hybridů genom druhu *C. taenia* častěji umlčuje alely genomu duhu *C. elongatoides*., tzn., že *C. taenia* umlčuje alely genomu *C. elongatoides*, což bylo prokázáno i na morfologických studiích (Kotusz, nepublikováno).

Z našich výsledků lze také vyvodit, 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 charakteru, 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 jak 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 komplexů 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 srovnat jen oocyty a játra, nenašli jsme mezi nimi těsnou korelaci.

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 nedá jednoduše vysvětlit, že naše pokrytí přirozené variability rodičovských genomů není ideální; frekventovaně mohlo dojít i k situaci, že některý SNP jsme mylně považovali za diagnostický, a on ve skutečnosti nebyl. Rád bych upozornil, že taková to kontrola nebývá zpravidla součástí publikací, které se zaměřují čistě analýzu SNP, ačkoliv výsledky vztahují na úroveň genu (Berletch et al., 2015; Crowley et al., 2015; Gerber et al., 2015).

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 rychlejší 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. Je známo, že stres vyvolaný polyploidním, hybridizačním šokem může vyvolat metylaci konkrétních genů, podobně jako odpověď na environmentální stres, kteréžto mohou být po mnoho generací děděny. V budoucnu plánujeme též sekvenovat metylomy hybridních linií rozdílného stáří, od F1 generace po linie staré až 350 tis let.

Degenerace klonálních linií?

Vzhledem k tomu, že jsme se celou dobu potýkali s daty pocházejícími z převážně kódujících oblastí, je přirozené 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 loci (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, 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 navíc u asexuálů byly značně heterozygotní díky jejich hybridnímu původu, a tudíž konstrukce těchto 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 naznačuje mírný nárůst aminokyselinových záměn u asexuálních linií s ohledem na stáří hybridní, asexuální linie. Tento fakt zapadá do kontextu

populárních „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) ukazovala na mtDNA lokusu, že k vyššímu tempu akumulace nesynonymních mutací může docházet i u relativně mladých klonů. Naše data naznačují, že sekavčí klony, třeba i 350 tisíc let staré, by mohly mít výraznější problém s fitness, díky nárůstu nesynonymních mutací; nicméně selekční zátěž stále není patrná, alespoň tedy ne ve srovnání s jejich sexuálními protějšky. Především bych rád podotkl, že studované sekavčí klony jsou často dominantami formami sekavců a úspěšně konkurují svým sexuálním protějškům (Janko et al., 2012). Náš pozitivní výsledek částečně není v souladu s terénními daty; je totiž obtížné si představit, že by mutacemi postižený klon byl schopen velmi úspěšně po stovky tisíc let trvající kompetice 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 byly přímo testovány rozdíly ve fitness mladých a starých klonů žab rodu Pelophylax, ale nepotvrdily se, 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 asexuálních komplexů jako sekavec a jiných a ukázal, že data jsou obecně v souladu s očekávanými vzory. Potvrzuje se však, že řada asexuálních taxonů jeví vyšší tempo extinkce ve srovnání se sexuálními druhy (Liu et al., 2012). Naše data tedy zapadají do rostoucího množství evidence, že asexuálové zvláště nesynonymní mutace akumulují. Tuto pozitivní evidenci podpořenou našimi daty pak (Janko et al., 2008) interpretuje tak, že Müllerova rohatka, či jiné podobné procesy existují. V některých analyzovaných modelech se ale může stát, 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 naše ale data nepodporují.

Ačkoliv je trend dat Müllerovy rohatky jasně patrný, uvědomujeme si, že dN/dS je také ukazatelem selekce a že k hybridizaci námi studovaných linií docházelo v době, kdy druhy *C. taenia* a *C. tanaitica* nebyly zcela odděleny. Problémem je, že analyzujeme i ancestrální SNP podléhajících selekčním tlakům, které mohou poměr dN/dS falešně vychýlit.

6 Souhrn

Moje práce měla od svého počátku velice dynamický průběh s řadou změn podle toho, jak se postupně ukazovaly problémy 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 dodatečné 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 obsáhlejší 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 jednotlivými typy hybridů, což dále může přispívat

k pozorovanému lineárnímu gradientu podobností hybridů a di- polyploidů k jejich rodičovským druhům. Nakonec je taky patrné, že genom *C. taenia* častěji umlčuje alely genomu *C. elongatoides*. Tyto výsledky se shodují i s analýzami morfologie hybridů (Kotusz and Kotusz, 2008), říkající, že diploidní hybridi jsou mírně podobnější s rodičovským druhem *C. taenia*.

Za významný nakonec považuji fakt, že jsem v práci potvrdil mírnou tendenci asexuálů k akumulaci nesynonymních záměn, což také potvrdilo obecné očekávání.

Myslím si, že moje práce patří spíše ke komplexnějším geonomickým studiím, které na asexuálních organismech byly podniknuty; tomu také odpovídá její rozsah, v některých část však možná na úkor její kvality. Výsledky, které jsem zde prezentoval, budu dále rozvíjet do formy několika publikací.

7 Seznam užitých zkratek

454	Pyrosekvenování (syntézní sekvenační metoda založená na bioluminiscenci pyrofosfátu vznikajícího při polymeraci)
A	adenin
ASE	alelově specifická exprese
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
cDNA	komplementární DNA - pocházející z mRNA
CpG	„ostrovy“ cytosinu a guaninu – regulační fce, promotory
DAG	diacylglycerol
DE	diferenciálně exprimované geny v rámci definovaných skupin
DNA	deoxy-ribonukleová kyselina
DP	sekvenační hloubka – počet *readů* na pozici reference
DTT	di-thio treitol, redukční činidlo
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. tanaitica*
ett	triploidní hybrid druhů *C. elongatoides* a *C. taenia*
F1	První filiální generace
FDR	*False rate discovery*, korekce P hodnot mnohonásobného testování
FISH	eng. *Fluorescence in situ hybridization*
G	guanin
gDNA	genomická DNA
GI	Identifikátor genu
GO	genová ontologie (kontrolovaný slovník přiřazující genům známé atributy molekulární funkce, buněčné lokalizace a biologických procesů)

GQ	*genotype quality*, phred score
H3, H1	Histony
hprt1	Hypoxanthin fosphoribosyltransferasa 1
HS	*housekeeping gen*, stabilně exprimovaný gen ve všech tkáních
IP3	ionosin-3-fosfát
KEGG	eng - *Kyoto Encyclopedia of Genes AND Genomes*, databáze signálních a metabolických drah
lncRNA	dlouhé nekdující RNA
LOH	eng – *loss of heterozygosity*, ztráta heterozygotnosti
lom300_ex_N	extendovaná verze transkriptomu o hybridní *kontigy*, které nejsou u druhu *C. taenia* exprimovány (N značí zanesení „N“ báze na polymorfní pozice)
lom300tf	játra oocyty 300 bp limit, druhu *C. taenia* (454 normalizovaný transkriptom), finální verze
MDS	mnohorozměrná analýza – multidimenzionální škálování (v případě užití euklieánské distance je výsledek „identický“ s PCA)
MQ	kvalita mapování udávána v Phred skóre
mtDNA	mitochondriální DNA
N	ambiguidní báze (A, T, C, či G)
nn	*Cobitis tanaitica*
non-POU	gen kódující Non-POU oktamer vázající doménu
ORF	eng – *open reading frame*, otevřený čtecí rámec
PCA	mnohorozměrná analýza – analýza hlavních komponent
PcoA	mnohorozměrná analýza – analýza hlavních koordinát
PCR	eng – *polaymerase chain reaction*, polymerázová řetězová reakce
PE	eng - *pair end*, sekvenování fragmentu ssDNA v obou směrech
PGS	Primordiální zárodečné buňky
PGS	*primordial germ cell*, primordiální zárodečné buňky
PLC	fosfolipáza C
Q	Phered skóre: záporný dekadický logaritmus pravděpodobnosti přečtení chybné báze
QD	*quality depth*, konfidence SNP beroucí v potaz sekvenační hloubku
qPCR	kvantitativní polymerizační řetězová reakce

RLE	*relative log expression*, normalizační metoda RNAseq
RNAseq, rna-seq	sekvenování nenormalizované RNA skrze reverzeně tanskribovanou cDNA
RPKM/FPKM	*Reads (fragments) per kilobase of exon per million,* norrmalizační metoda RNAseq
rpl13a	ribozomální protein velké podjednotky 3a
SE	eng - *single end*, sekvenování fragmentu ssDNA v jednom směru
SeqCap	sekvenování vzorku se sníženou komplexitou aplikací hybridizačních sond cílených na gDNA
SH2, SH3	Src Homology; konzervované proteinové domény
SNP	jedno-nukleotidové polymorfismy
SRY	Sex determinující region Y, transkripční faktor
SS	*sum of squares*, součet čtverců
T	tymin
TC	*total count*, normalizační metoda RNAseq
TMM	*trimmed mean of M-values,* normalizační metoda RNAseq
tt	*Cobitis tenia*
UQ	*upper quartile*, normalizační metoda RNAseq
UTR	netranslatované regulační oblasti na 3' a 5' konci mRNA (fce - lokalizace mRNA, účinnost translace a stabilita mRNA)
VCF	*variant call format*

8 Bibliografie

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

Anatskaya, O.V., Erenpreisa, J.A., Nikolsky, N.N., and Vinogradov, A.E. (2016). Pairwise comparison of mammalian transcriptomes associated with the effect of polyploidy on the expression activity of developmental gene modules. Cell Tissue Biol. *10,* 122–132.

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

Andersen, C.L., Jensen, J.L., and Ørntoft, T.F. (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.

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

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

Beaudet, A.L., and Jiang, Y.H. (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., Kane, N.C., Rieseberg, L.H., and Adams, K.L. (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. In Lost Sex, I. Schön, K. Martens, and P. Dijk, eds. (Dordrecht: Springer Netherlands), pp. 1–19.

Berletch, J.B., Ma, W., Yang, F., Shendure, J., Noble, W.S., Disteche, C.M., and Deng, X. (2015). Identification of genes escaping X inactivation by allelic expression analysis in a novel hybrid mouse model. Data Brief *5*, 761–769.

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

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

de Boer, J.G., Yazawa, R., Davidson, W.S., and Koop, B.F. (2007). Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. BMC Genomics *8*, 422.

Bohlen, J., Perdices, A., Doadrio, I., and Economidis, P.S. (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., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma. Oxf. Engl. *30*, 2114–2120.

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

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics *11*, 94.

Burge, S.W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E.P., Eddy, S.R., Gardner, P.P., and Bateman, A. (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.

Castel, S.E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. Genome Biol. *16*.

Chang, Y.-F., Imam, J.S., and Wilkinson, M.F. (2007). The nonsense-mediated decay RNA surveillance pathway. Annu. Rev. Biochem. *76*, 51–74.

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

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

Chen, Z.J., and Ni, Z. (2006). Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. BioEssays News Rev. Mol. Cell. Dev. Biol. *28*, 240–252.

Cho, H., Davis, J., Li, X., Smith, K.S., Battle, A., and Montgomery, S.B. (2014). High-Resolution Transcriptome Analysis with Long-Read RNA Sequencing. PLOS ONE *9*, e108095.

Choleva, L., Janko, K., De Gelas, K., Bohlen, J., Šlechtová, V., Rábová, M., and Ráb, P. (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., Musilova, Z., Kohoutova-Sediva, A., Paces, J., Rab, P., and Janko, K. (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.

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

Copernicus, I. (2003). Folia Biologica (Index Copernicus).

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

Crowley, J.J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I.K., Kim, Y., Wang, J.R., Morgan, A.P., Calaway, J.D., Aylor, D.L., et al. (2015). Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nat. Genet. *47*, 353–360.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Degner, J.F., Marioni, J.C., Pai, A.A., Pickrell, J.K., Nkadori, E., Gilad, Y., and Pritchard, J.K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics *25*, 3207–3212.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (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.

Dion-Côté, A.-M., Renaut, S., Normandeau, E., and Bernatchez, L. (2014). RNA-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young lake whitefish species. Mol. Biol. Evol. *31*, 1188–1199.

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., Tsuchiya, D., Andrews, J., Lynch, M., and Zolan, M.E. (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.

Enright, A.J., Dongen, S.V., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584.

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

Feschotte, C. (2008). The contribution of transposable elements to the evolution of regulatory networks. Nat. Rev. Genet. *9*, 397–405.

Flegr, J. (2007). Úvod do evoluční biologie (Praha: Academia).

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

Galla, G., Vogel, H., Sharbel, T.F., and Barcaccia, G. (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 Roberts, S.B. (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.

Gerber, M.M., Hampel, H., Zhou, X.-P., Schulz, N.P., Suhy, A., Deveci, M., Çatalyürek, Ü.V., and Ewart Toland, A. (2015). Allele-specific imbalance mapping at human orthologs of mouse susceptibility to colon cancer (Scc) loci. Int. J. Cancer *137*, 2323–2331.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011a). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011b). 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., Hotz, H., and Semlitsch, R.D. (2002). Deleterious alleles and differential viability in progeny of natural hemiclonal frogs. Evol. Int. J. Org. Evol. *56*, 1036–1044.

Hansen, K.D., Brenner, S.E., and Dudoit, S. (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., and Eisner, J. (2009). Sperm-dependent parthenogens delay the spatial expansion of their sexual hosts. J. Theor. Biol. *261*, 431–440.

Janko, K., Culling, M.A., Ráb, P., and Kotlík, P. (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., Bohlen, J., Lamatsch, D., Flajshans, M., Epplen, J.T., Ráb, P., Kotlík, P., and Slechtová, V. (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., Drozd, P., Flegr, J., and Pannell, J.R. (2008). Clonal Turnover Versus Clonal Decay: A Null Model for Observed Patterns of Asexual Longevity, Diversity and Distribution. Evolution *62*, 1264–1270.

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

Janko, K., Kotusz, J., De Gelas, K., Šlechtová, V., Opoldusová, Z., Drozd, P., Choleva, L., Popiołek, M., and Baláž, M. (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 Bragg, E. (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 Boroń, A. (2006). Age, reproduction and fecundity of the spined loach Cobitis taenia L. (Pisces, Cobitidae) from Lake Klawój (Poland). Reprod. Biol. *6*, 133–148.

Juchno, D., Jabłońska, O., Boroń, A., Kujawa, R., Leska, A., Grabowska, A., Nynca, A., Świgońska, S., Król, M., Spóz, A., et al. (2015). Erratum to: Ploidy-dependent survival of progeny arising from crosses between natural allotriploid Cobitis females and diploid C. taenia males (Pisces, Cobitidae). Genetica *143*, 127.

Kaushik, K., Leonard, V.E., Kv, S., Lalwani, M.K., Jalali, S., Patowary, A., Joshi, A., Scaria, V., and Sivasubbu, S. (2013). Dynamic Expression of Long Non-Coding RNAs (lncRNAs) in Adult Zebrafish. PLOS ONE *8*, e83616.

Kearney, M. (2005). Hybridization, glaciation and geographical parthenogenesis. Trends Ecol. Evol. *20*, 495–502.

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

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

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., and Kotusz, J. (2008). Morphological relationships between polyploid hybrid spined loaches of the genus Cobitis (Teleostei: Cobitidae) and their parental species (Natura optima dux Foundation).

Kotusz, J., Popiołek, M., Drozd, P., De Gelas, K., Šlechtová, V., and Janko, K. (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., Hansen, K.D., and Leek, J.T. (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol. *11*, R83.

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

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

León-Novelo, L.G., McIntyre, L.M., Fear, J.M., and Graze, R.M. (2014). A flexible Bayesian method for detecting allelic imbalance in RNA-seq data. BMC Genomics *15*, 920.

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

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

Li, H. (2011c). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinforma. Oxf. Engl. *27*, 2987–2993.

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

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

Li, C.-Y., Li, J.-T., Kuang, Y.-Y., Xu, R., Zhao, Z.-X., Hou, G.-Y., Liang, H.-W., and Sun, X.-W. (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.

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

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

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

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

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

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

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

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

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

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. *35*, W182–W185.

Munger, S.C., Raghupathy, N., Choi, K., Simons, A.K., Gatti, D.M., Hinerfeld, D.A., Svenson, K.L., Keller, M.P., Attie, A.D., Hibbs, M.A., et al. (2014). RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. Genetics *198*, 59–73.

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

Nadeau, N.J., Whibley, A., Jones, R.T., Davey, J.W., Dasmahapatra, K.K., Baxter, S.W., Quail, M.A., Joron, M., ffrench-Constant, R.H., Blaxter, M.L., et al. (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., Miya, M., Mabuchi, K., Saitoh, K., and Nishida, M. (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., Hehman, G., Miller, J.T., Logsdon, J.M., and Taylor, D.R. (2010). Accelerated mutation accumulation in asexual lineages of a freshwater snail. Mol. Biol. Evol. *27*, 954–963.

Niazi, F., and Valadkhan, S. (2012). Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3′ UTRs. RNA *18*, 825–843.

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

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

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

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

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

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., et al. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. *22*, 577–591.

Pellino, M., Hojsgaard, D., Schmutzer, T., Scholz, U., Hörandl, E., Vogel, H., and Sharbel, T.F. (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., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Qi, W., Tang, Z., and Yu, H. (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.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

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

Ravi, M., Marimuthu, M.P.A., and Siddiqi, I. (2008). Gamete formation without meiosis in Arabidopsis. Nature *451*, 1121-U10.

Rehrauer, H., Opitz, L., Tan, G., Sieverling, L., and Schlapbach, R. (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 Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

Ruffalo, M., Koyuturk, M., Ray, S., and LaFramboise, T. (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., Kouyos, R.D., and Bonhoeffer, S. (2008). The state of affairs in the kingdom of the Red Queen. Trends Ecol. Evol. *23*, 439–445.

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

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

Seehausen, O., Butlin, R.K., Keller, I., Wagner, C.E., Boughman, J.W., Hohenlohe, P.A., Peichel, C.L., Saetre, G.-P., Bank, C., Brännström, A., et al. (2014). Genomics and the origin of species. Nat. Rev. Genet. *15*, 176–192.

Shen, Y., Catchen, J., Garcia, T., Amores, A., Beldorth, I., Wagner, J., Zhang, Z., Postlethwait, J., Warren, W., Schartl, M., et al. (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., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J.M. (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 DesRosiers, N.J.D. (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.

Teare, M.D., Heighway, J., and Santibáñez Koref, M.F. (2006). An Expectation-Maximization Algorithm for the Analysis of Allelic Expression Imbalance. Am. J. Hum. Genet. *79*, 539–543.

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

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

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

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

Wang, J., Ye, L.H., Liu, Q.Z., Peng, L.Y., Liu, W., Yi, X.G., Wang, Y.D., Xiao, J., Xu, K., Hu, F.Z., et al. (2015). Rapid genomic DNA changes in allotetraploid fish hybrids. Heredity.

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

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

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. *8*, 206–216.

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

Wu, Y., Sun, Y., Wang, X., Lin, X., Sun, S., Shen, K., Wang, J., Jiang, T., Zhong, S., Xu, C., et al. (2016). Transcriptome shock in an interspecific F1 triploid hybrid of Oryza revealed by RNA sequencing. J. Integr. Plant Biol. *58*, 150–164.

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

Zhai, R., Feng, Y., Zhan, X., Shen, X., Wu, W., Yu, P., Zhang, Y., Chen, D., Wang, H., Lin, Z., et al. (2013). Identification of Transcriptome SNPs for Assessing Allele-Specific Gene Expression in a Super-Hybrid Rice Xieyou9308. PLoS ONE *8*, e60668.

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

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

Zhang, X., Deng, M., and Fan, G. (2014). Differential Transcriptome Analysis between Paulownia fortunei and Its Synthesized Autopolyploid. Int. J. Mol. Sci. *15*, 5079–5093.

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

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

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

Zhulidov, P.A., Bogdanova, E.A., Shcheglov, A.S., Shagina, I.A., Wagner, L.L., Khazpekov, G.L., Kozhemyako, V.V., Lukyanov, S.A., and Shagin, D.A. (2005). A method for the preparation of normalized cDNA libraries enriched with full-length sequences. Russ. J. Bioorganic Chem. *31*, 170–177.