

Studium genetické diverzity a struktury populací pomocí mtDNA a jaderných mikrosatelitových markerů u včely medonosné

ISAGREED MENDELU PhD modul-1 CZ

Dobrý den, v této přednášce určené pro studenty doktorského studia se zaměříme na možnosti hodnocení genetické diverzity a struktury populací pomocí mitochondriální DNA a jaderných mikrosatelitních markerů aplikovaných u včely medonosné. Přednáška je součástí modulu 1 – Genetika zvířat. Vytvoření této prezentace bylo podpořeno grantem Erasmus+, KA2 v rámci projektu ISAGREED Inovace obsahu a struktury studijních programů v oblasti managementu živočišných genetických a potravinových zdrojů s využitím digitalizace. Obsahem této přednášky budou témata jako získávání genetických dat, hodnocení genetické variability pomocí sekvencí mitochondriální DNA a hodnocení genetické variability pomocí jaderných STR nebo mikrosatelitních markerů.

Proč je důležitá genetická diverzita? Genetická diverzita je významná pro zdraví populace. Je to zásadní zdroj možnosti vytvářet toleranci nebo rezistenci k současným i budoucím nemocem, patogenům a predátorům. Současný stav populací včel lze přisoudit mimo jiné ke snížení diverzity. Diverzita včel byla hodnocena morfometrickými znaky, jako jsou parametry křídel, pigmentace a podobně.

Nyní je popsáno 31 poddruhů, plemen či ras včely medonosné (*Apis mellifera*). Pomocí analýz DNA, a to hlavně mitochondriální, byly popsány evoluční linie jako západní mediteránní typ M, severní mediteránní typ C, africká linie A, orientální linie O. A jsou poznávány další a další. Genom včely medonosné byl již několikrát celkově osekvenován.

Zde můžeme vidět jednotlivé chromozomy, a zejména předposlední sloupec nám ukazuje velikost jednotlivých chromozomů v délce počtu párů bází. V genomu včely medonosné bylo celkem popsáno nebo se odhaduje, že je celkem na 12 398 genů. Z toho je 9935 genů, které kódují nějaké proteiny, a 2421 genů, které nekódují proteiny, to znamená kódují jiné RNA, jako jsou třeba transferové RNA nebo ostatní malé jaderné RNA.

Mitochondriální DNA je u většiny živočišných druhů v rozsahu kolem 16 až 20 kB. Mitochondriální genom včely medonosné západní má přibližně 16 500 párů bází. V referenčním genomu NC001566 je mitochondriální DNA o velikosti 16 343 párů bází. V kompletním genomu mitochondriální DNA včely karpatské má velikost 16 358 párů bází. Mitochondriální DNA včely medonosné obsahuje 13 genů, které kódují proteiny. Dále obsahuje 22 genů, které kódují tRNA, a 2 geny, které kódují ribozomální RNA.

Pro fylogenetické a fytogeografické analýzy se zejména využívají sekvence barcodingu, což je sekvence cytochromoxidázy 1 a tzv. intergenová oblast, zahrnující část genu pro tRNA leucin a cytochromoxidázu 2. Jakou variabilitu lze pozorovat na úrovni DNA, čili jaké existují molekulární genetické markery? V současné době se nejčastěji využívají bialelické jednonukleotidové polymorfizmy, které se mohou nacházet jak v kódujících, tak i v nekódujících oblastech. Dále se dají využívat informace o tzv. inzercích a delecích, to znamená, kdy chybí nebo přebývá nějaká báze. Můžeme to nazývat jako indely. Na obrázcích vpravo vidíte příklady: nahoře jsou SNP markery a dole delece cytosinu v oblasti sekvence ACA.

Po vyizolování DNA ze vzorku včely a její amplifikaci konkrétního malého úseku, například jednoho z těch dvou námi zmiňovaných genů, dojde k sekvenování v tzv. sekvenátoru na základě kapilární elektroforézy. Výsledek sekvenování, to znamená určení jednotlivých bází v dané sekvenci, vidíme na obrázku.

Jednotlivé píky jsou vyjádřeny barvami, kterými jsou určeny jednotlivé báze. K identifikaci podtypů, mitotypů nebo haplotypů v linii jsme využívali dvě sekvence mitochondriální DNA. A to je ta výše zmiňovaná, tedy intergenový úsek, tRNA leucin-cox2, a oblast cytochromoxidázy 1. Tento úsek se skládá ze dvou mitochondriálních genů – transferové RNA pro leucin a podjednotky cytochromoxidázy 2. Tato sekvence je typická vysokým obsahem mutací a vykazuje významné rozdíly v délce a složení nukleotidů mezi populacemi včely medonosné. Tento ampikon se štěpí restriční endonukleázou *DraI*, která specificky rozpoznává sekvenci TTAA, aby se mohla určit každá linie. Druhou sekvencí je sekvence pro oblast barcodingu a ta je součástí genu cytochromoxidázy 1 (*cox1*). Tato sekvence se porovnává se sekvencemi uloženými v databázích, jako je BOLT systém nebo GenBank. A tento fragment DNA je v rámci taxonů vysoce konzervovaný a často se právě používá pro rozlišení taxonů i jednotlivých druhů.

Podle struktury v sekvenci tRNA-leucin-cox2 lze určit jednotlivé evoluční linie u včely medonosné. Linie C, kam patří právě naše včela kraňská, ale taky včela ligustica, macedonica a další, obsahují pouze jednu kopii tzv. Q sekvence. A linie obsahují jednu až dvě kopie této Q sekvence a navíc ještě tzv. úsek P0. M linie, to je původní černá včela *Apis mellifera mellifera*, která už se v České republice vlastně ani nevyskytuje, může obsahovat jednu, dvě až tři opakování sekvence Q a navíc tzv. P sekvenci. Díky sekvenování a porovnávání jednotlivých sekvencí můžeme identifikovat u každé včely nějakou evoluční linii.

Identifikovat konkrétní linii je možné právě pomocí štěpení restričním enzymem *DraI*, které rozpoznává tu zmíněnou sekvenci TTAA. Jakmile dojde v důsledku nějaké mutace k této změně, tento enzym tuto sekvenci nerozpozná, místo nemůže štěpit a pomocí klasické PCR reakce, pomocí délky jednotlivých fragmentů, lze od sebe odlišit jednotlivé varianty, jako je C a A1, případně A4.

Druhou možností je získat celou sekvenci daného úseku pomocí sekvenování a tu následně analyzovat. Zde vidíte příklad sekvencí tRNA-leucin-cox2 u několika jedinců.

Některé programy, jako například UniPro Eugene, umožňují provést štěpení restriktázou *DraI* in silico, to znamená v počítači. Tady máme příklad jedné sekvence, která patří do C linie a obsahuje právě tato tři štěpná místa, čímž vznikají tři fragmenty o specifických délkách. U jiného vzorku nebylo nalezeno jedno štěpné místo a podle délky fragmentů lze usoudit, že se jedná o včelu, která patří do linie A, čili do africké.

Získané sekvence z větší populace se potom srovnávají pomocí tzv. Multiple Sequential Alignmentu, čili vícečetného sekvenačního srovnání. Mohli bychom to dělat manuálně, ale byly vytvořeny programy s algoritmy, které si s tím snadněji poradí. Některé jsou online, například na serverech Evropského bioinformatického institutu, a pro naše účely byl nejvhodnější Kalign. Existují ale i další nástroje, jako Cluster Omega, MAFFT a podobně.

Existují rovněž programy, které si stáhnete a nainstalujete, a které vám tuto analýzu také mohou provést, například programy MEGA nebo již zmíněný Unipro Eugene.

Pro detekci polymorfismu DNA a haplotypů v obou oblastech mitochondriální DNA s využitím všech sekvencí z vícenásobného zarovnání sekvencí ve formátu FASTA byl použit program DnaSP. Dále jsme nukleotidové substituce a inserce/delece pro každý haplotyp porovnávali s referenčním genomem. Identifikaci specifických haplotypů v tRNA Leucin-cox2 linii C a A byly dále vyhledány referenční sekvence se 100% identitou pomocí nástrojů BLAST lokálního párového zarovnání vůči sekvencím, které se nacházejí v databázi Národního centra pro biotechnologické informace (NCBI) v americkém GenBanku.

Pro ověření haplotypů *cox1* byl také použit BLAST, ale sekvence byly ověřeny i pomocí databáze BOLD na základě vícenásobného zarovnání pomocí Kalign s nutností dalšího ručního dořešení. Identifikovali jsme celkem 13 haplotypů, z nichž 3 patřily do A linie a ostatní do C linie. Nejčastější haplotyp byl C1A, o kterém se říká, že je typický pro *Apis mellifera ligustica*, tzv. italskou včelu.

V této tabulce můžeme vidět zařazení jednotlivých haplotypů do C a A linie podle štěpení *DraI* spektra a sekvenování. Jednotlivé haplotypy a jejich sekvence byly vloženy do GenBanku na serveru NCBI. Ve třetím sloupci můžete vidět naše referenční sekvence. Dále tam můžeme vidět počty a délky fragmentů vzniklých štěpením, které také vykazují velkou variabilitu. V posledních dvou sloupcích můžete vidět identifikaci či porovnání s jinými sekvencemi v databázi GenBank, kde jsme vybrali sekvence se stoprocentní shodností s našimi sekvencemi.

Všechny námi určené haplotypy patřící do C-linie byly opravdu popsány u *Apis mellifera carnica*. Tři africké haplotypy byly identifikovány jako *Apis mellifera iberica*. Jedna jediná sekvence s plnou shodou nebyla přisouzena žádnému poddruhu.

V této tabulce můžeme vidět určení nejdůležitějších pozic polymorfních míst, jaké báze se nacházely v jakém haplotypu. Máme zde 10 pozic, kde byly zjištěny především SNP, ale také delece. Všimněte si polymorfismu na pozici 50, kde standardní alela je C. Ostatní haplotypy na této pozici vykazovaly deleci.

Podobně jsme analyzovali i sekvenci *cox1*, kde jsme našli 13 různých haplotypů pro barcoding. Opět zde vidíme jednotlivé SNP mutace. V barcodingové sekvenci nebyly nalezeny žádné inserce či delece, pouze SNP substituce.

V tabulce následně vidíte výsledky frekvencí výskytu haplotypů v genu tRNA Leucin-cox2 a v genu COX1 v České republice. Vidíme, že je zde celá řada haplotypů, které jsou relativně dobře zastoupeny, jako C1A, C2L, C2E, C2C. Poté jsou zde haplotypy nalezené pouze u jednoho nebo několika jedinců.

Podobná situace je v genech *cox1*, kde první čtyři haplotypy byly nejčastěji zastoupeny a ostatní byly v menšině nebo individuálních vzorcích. Z populace včel v České republice z více než 300 vzorků jsme spočítali některé charakteristiky, jako parametry diverzity haplotypů v sekvencích tRNA Leucin-cox2 a *cox1*.

Byly odhadnuty haplotypová diverzita H_D , molekulární diverzita π a Tajimovo D . Tyto indexy genetické diverzity byly vyhodnoceny pomocí balíčku Pegas v programu R, ale mohou být

využity také programy jako DnaSP, MEGA nebo Arlequin. Dále jsme určovali tzv. haplotypové sítě.

Využili jsme metodu RMSAT (Randomized Minimum Spanning Tree), která zohledňuje frekvence a vztahy mezi haplotypy. Tyto sítě haplotypů byly zpracovány pomocí balíčku Pegas v programu R. Můžou být ale použity i jiné programy, jako například PopArt a další. Tady vidíme výsledek analýzy haplotypové sítě u 308 jedinců v sekvenci tRNA Leucin-cox2.

Vidíme, že nejčastější haplotyp je opravdu ten C1A, druhý nejčastější je C2E, mezi nimiž jsou dvě bodové mutace, a třetí nejčastější haplotyp v ČR je C2C. Každá barva v tabulce představuje konkrétní kraj, odkud byla včela získána. Naším cílem bylo pokrýt celou Českou republiku, rovnoměrně odebrat vzorky ze všech krajů.

Na dalším slajdu můžeme vidět analýzu haplotypové sítě u sekvence cox1, kde opět vidíme, že pokud byl haplotyp dostatečně četný, vyskytoval se téměř ve všech krajích. Týká se to HpB02, HpB03, HpB01 a HpB04. Ostatní haplotypy se vyskytovaly u několika jedinců nebo pouze u jednoho jedince.

Dále byla provedena fylogenetická analýza na základě těchto sekvencí. To znamená, po provedení MSA, vícenásobného zarovnání sekvencí, jsme využili tvorbu fylogenetických stromů, a to zejména pomocí metody Maximum Likelihood a Tamura-Nei modelu v programu Mega X. Tato metoda je založena na tom, že se vytváří bootstrapový konsenzuální strom na základě 10 000 replikátů. Jednotlivé větve odpovídají oddílům produkovaným v méně než 50 % bootstrapových replikátů a % replikačních stromů, které se související haplotypy shlukovaly v bootstrapovém testu.

Tyto informace jsou potom uvedeny vedle těch větví. Počáteční strom pro heuristické vyhledávání byl získán automaticky použitím algoritmů Neighbor Joining a BioNJ na matici párových vzdáleností odhadnutých pomocí modelu Tamura-Nei a následným výběrem topologie s nejvyšší hodnotou logaritmické pravděpodobnosti. Výsledkem jsme získali tyto fylogenetické stromy.

Vlevo máme fylogenetický strom na základě analýzy sekvence tRNA Leucin-cox2. Vpravo fylogenetický strom na základě COX1 sekvence. Můžeme vidět, že africké včely nebo včely z linie A se shlukují společně na rozdíl od ostatních, to znamená včel linie C, které jsou zde označeny červeně.

Druhým typem markerů, které jsme využívali pro hodnocení genetické variability, byly mikrosatelity. Jedná se o polymorfismy, které se výhradně vyskytují v jaderné DNA. Tyto polymorfismy jsou charakteristické tandemovými opakováními určitého motivu, například GC. Každá alela je označována podle délky tohoto opakování.

Máme například alelu s 8 opakováními, další s 3 opakováními a poslední s 10 opakováními. Alela se označuje podle oblasti, v níž se mikrosatelit nachází, která je ohraničena primery. Pomocí sekvenátoru a fragmentační analýzy jsme schopni určit délku daného úseku obsahujícího mikrosatelit. Dole na obrázku můžeme vidět určení genotypů pro konkrétní

mikrosatelit, který je v tomto případě dán třemi alelami charakterizovanými čísly 156, 152 a 142.

Vidíme, že mikrosatelity jsou velmi polymorfni, nemusí mít pouze tři alely, ale klidně i 20 alel, což v populaci může znamenat velké množství různých kombinací v genotypech. Jsou tedy výhodné pro hodnocení diverzity v populacích. Zde vidíme příklad variability ve dvou populacích.

Populace vlevo je málo variabilní, obsahuje pouze tři typy alel a velké množství homozygotních jedinců. Druhá populace vpravo obsahuje velké množství alel a často se zde mohou vyskytovat heterozygotní genotypy.

Proč se pořád využívají mikrosatelitní markery, když máme již celogenomové sekvence? Mikrosatelitní markery jsou relativně cenově dostupné a jejich identifikace poskytuje multilokusovou genotypovou informaci. Lze je snadno využívat k odhadu genetických diverzit populací a struktur, což je důležité i v konzervační genetice a šlechtění.

Pro určení genotypů se výhradně využívá tzv. fragmentační analýza, která se provádí v sekvenátorech s kapilární elektroforézou. Fragmenty se dělí podle své velikosti, snímač detekuje průchody molekul, jejich barvu a intenzitu signálu v určitém čase a získává informace o délkách fragmentů.

Přístroj, který se využívá, je genetický analyzátor. V našem případě jsme využívali genetický analyzátor ABIPrism 3500. Velikost fragmentů byla přesně stanovena pomocí softwaru GeneScan a genotypy byly určeny pomocí softwaru GeneMapper.

Protože jsme sledovali 22 lokusů mikrosatelitních, seskupili jsme určité mikrosatelity do jedné multiplexní reakce. Několik mikrosatelitů za stejných podmínek, rozlišeno různými barvami, jsme mohli identifikovat. Výsledkem analýzy je zobrazení, kde vidíme pro konkrétní mikrosatelitní lokus označený barvou jeho píky, což jsou identifikované fragmenty o specifické délce.

Na obrázku můžeme vidět vyhodnocení variability určení genotypů mikrosatelitních lokusů u tří jedinců. Vidíme, že na daných pozicích oblasti se mohou vyskytovat různé alely, což umožňuje snadné rozlišování jedinců.

Po určení genotypů ve všech lokusech a u všech jedinců v populaci následuje analýza diverzity, tedy určení diverzitních parametrů, jako je počet alel, efektivní počet alel, Shannonův informační index, pozorovaná a očekávaná heterozygotnost a nevychýlená očekávaná heterozygotnost, a tzv. fixační index F . Využili jsme program Genalex, který funguje v Microsoft Excelu, ale data a parametry lze vypočítat i pomocí balíčku diversity v programu R.

Vidíme, že očekávaná heterozygotnost je průměrná za všechny lokusy dohromady 0,579 a skutečná pozorovaná heterozygotnost má hodnotu 0,556. Je to relativně vysoká heterozygotnost, což charakterizuje tuto populaci včel jako dostatečně diverzní.

Protože jsme u každého jedince věděli, z jaké oblasti České republiky byl odebrán, rozdělili jsme populaci České republiky do 77 okresů, které charakterizují geografické oblasti. Díky tomu jsme mohli spočítat tzv. Wrightovy F -statistiky, F_{ST} , F_{IS} a F_{IT} , a určit tzv. analýzu

molekulární variance, která určuje podíl variability mezi populacemi, mezi jedinci v populacích a uvnitř jedinců.

Tabulka vlevo charakterizuje jednotlivé lokusy a průměrné hodnoty F-statistik. F_{ST} je nejzajímavější, protože určuje míru rozdílnosti mezi subpopulacemi, tedy okresy. Hodnota 0,086 není příliš vysoká, ale nějakou diversifikaci ukazuje.

V tabulce vpravo vidíme výsledek analýzy molekulární variance, kde v posledním sloupci variabilita mezi regionálními populacemi (okresy) tvoří pouze 1 % celkové variability. Ale i 1–3 % variability mezi geografickými oblastmi jsou obvyklá čísla podle jiných publikací.

Mezi jedinci uvnitř populací je variabilita vyjádřena 6 % a největší podíl variability v populaci tvoří variabilita uvnitř jedinců. Pomocí programu Genalex byly vypočítány párové Neiovy a párové F_{ST} genetické distance. Tyto párové hodnoty, tyto distance byly využité pro výpočty nebo analýzu hlavních komponent PCA.

K analýze genetické diverzity a míry příměsi populací včely medonosné byla použita bayesovská metoda shlukování STRUCTURE ver, 2.3.4. Bylo provedeno deset nezávislých simulací, z nichž každá zahrnovala 10 000 burn-in kroků a následně 100 000 iterací Markovova řetězce Monte Carlo (MCMC). Následně jsme použili programy Clumpak a Structure Selector, které implementují Evannovu metodu a Puechmailleovu metodu, abychom zjistili optimální počet shluků (K), který nejlépe odpovídá datům (ΔK , MedMeaK, MaxMeaK, MedMedK a MaxMedK). V případě této populace oběma metodami bylo určeno, že v populaci včely medonosné v České republice na základě těchto 22 mikrosatelitních markerů se vyskytují tři geneticky odlišitelné populace a že lze s určitou vysokou mírou pravděpodobností každého jedince přiřadit do jedné z těchto tří skupin.

Pro určení genetické struktury lze využívat i jiné metody, například diskriminační analýzu hlavních komponent DAPC, která je naprogramována v balíčku adgenet v rámci programu R. Touto metodou opět můžeme zjistit strukturu populace a hovořit o klastrech, skupinách, které se mezi sebou geneticky liší a jedince je možné do těchto skupin jednoznačně přiřadit. V této populaci včely medonosné v České republice byly odhadnuty celkem pět skupin, pět klastrů, které se odlišovaly.