Helena Štorchová

Co prozradilo srovnání genomů a transkriptomů bublinatek. Od populační genetiky k populační genomice 1.

Biologické bádání prochází další revolucí. Možnosti analýz celého genomu, donedávna vyhrazené jen několika málo modelům, jsou nyní dostupné pro široké spektrum druhů. Poprvé se otvírá příležitost srovnávat jedince, populace, poddruhy i vyšší taxony na základě informace z celých genomů či transkriptomů, tedy souborů molekul RNA. V následujícím článku jsou objasněny některé základní principy genomických analýz, v příštím čísle Živy pak ukážeme výsledky prvního nahlédnutí za nadzvednutou oponu genomu bublinatek (*Utricularia*). Tyto vodní masožravé rostliny s velmi zajímavým mechanismem pastí (Živa 2015, 6: 286–289) mají jedny z nejmenších genomů v říši rostlin. Překvapivě však nikoli méně genů než známý experimentální model huseníček rolní (*Arabidopsis thaliana*) a ještě překvapivěji prošly více cykly polyploidizace než tato modelová rostlina. Avšak než se vydáme společně prozkoumat genomy vodních rostlin, vysvětlíme, jak se sestavují mapy těchto genomů. Bez map se neobejde žádná expedice, ať již míří do pralesů Amazonie, či mezi chromozomy. Pokud nikdo před námi daný druh nezkoumal, nezbývá, než abychom genomickou mapu sami načrtli podle sekvencí DNA. V následujícím článku se dočtete, jaké možnosti současná molekulární biologie nabízí.

V bývalém Československu byl z politických důvodů v 50. letech 20. stol. násilně přerušen rozvoj genetiky na celou jednu generaci. Lysenkovská biologie vnucená sovětskými poradci operovala s dědičností získaných vlastností, zatímco geny a genetiku neuznávala. Avšak přírodní zákony, na rozdíl od těch lidských, mají obecnou platnost, a tak se genetika za katedry časem vrátila. Nicméně mezera se nezacelila snadno. Na základních a středních školách se o existenci populační a evoluční genetiky studenti zpravidla mnoho nedozvědí. Současná rostlinná biologie je již dlouho vědou integrální, zahrnující rovnocenně práci v terénu, laboratorní postupy i analýzy dat. Díky novým generacím sekvenačních technologií (next generation sequencing, NGS; Štorchová 2011) prožíváme převratné období, spočívající ve srovnávací analýze celých genomů, a to i jedinců nemodelových druhů v přírodních populacích. Vznikl zcela nový obor – populační genomika. Na rozdíl od populační genetiky pracující s jedním nebo několika úseky DNA (zpravidla neutrálními markery nepodléhajícími přírodnímu výběru) umožňuje populační genomika také odhalit geny, které mají adaptivní hodnotu a přispěly ke vzniku druhu nebo úspěšnému šíření dané populace (viz např. článek o adaptivní fylogeografii v Živě 2015, 2: 53–58). Populační genomika neodpovídá tedy jen na otázku: jak? (např. kde se nalézala glaciální refugia a kudy se po skončení doby ledové druhy šířily), ale také proč? (např. které mutace nebo alely korelují se zeměpisnou šířkou a jakou mají funkci, jak ovlivňují rychlost reprodukce, mrazuvzdornost nebo něco jiného).

Vlak rostlinné genomiky se teprve rozjíždí, je vhodná doba do něj naskočit, již nic nepromeškat. Bude proto dobré vysvětlit, jak se genomická data získávají a zpracovávají a zejména, co vše se z nich dá vyčíst (také Živa 2006, 5: 198–200). Zvlášť skryté cesty vedou až k době počátku naší planety, kdy se ve vodách praoceánu objevil LUCA (Last Universal Common Ancestor) – poslední univerzální společný předek, tedy předek všeho živého. Ve sledu (sekvenci) čtyř písmen symbolizujících báze DNA najdeme rozluštění mnoha velkých záhad přírody. Některé z nich, jako např. evoluce masožravých rostlin, zaujaly již Ch. Darwina. Proto jsme ke krátkému přiblížení rostlinného genomu vybrali vodní masožravou bezkořennou bublinatku (o jejím životě viz zmíněný článek v Živě 2015, 6). Nejprve však popíšeme, jak se sekvenční data získávají a jak z nich genomickou krajinu vymodelujeme.

První přečtené rostlinné genomy

Genomy krytosemenných rostlin se navzájem podstatně liší svou velikostí o více než čtyři řády (v rozpětí 10^7 až 10^{11} bp – párů bází, tedy písmen genetického kódu, podrobněji viz Živa 2015, 1: 4–5). Kupodivu počet genů, tedy úseků DNA kódujících funkční protein nebo RNA, je přes tyto obrovské rozdíly ve velikosti genomu zhruba stejný, kolem 30 tisíc. Výjimku tvoří čerství polyploidi, kteří ještě nestačili snížit znásobený genový počet. Tento paradox (odborně se mu říká paradox C-hodnoty) naznačuje, že obrovské velikostní rozdíly mezi rostlinnými genomy nezpůsobuje odlišný počet genů, ale změny v rozsahu nekódující DNA (junk DNA, tedy doslova přeloženo harampádí či veteš, byť se hromadí důkazy, že přece jen nějaký význam má – podle současných převládajících představ jde o pozůstatky evoluce DNA, které, ať mají, nebo nemají nějakou konkrétní funkci, mohou být v evoluci opět použity). Velkou část DNA tvoří repetitivní

1 Lokalita bublinatky *Utricularia macrorhiza*, blízké příbuzné i u nás vzácně rostoucí b. obecné (*U. vulgaris*), na Aljašce. Smithovo jezero nedaleko Fairbanks je obklobeno boreálním lesem se smrkem sivým (*Picea glauca*) a smrkem černým (*P. mariana*). V zimě ho pokrývá led o tloušťce až 3 m, který roztaje teprve v polovině května. Vegetační sezona zde trvá jen necelé čtyři měsíce. Foto H. Štorchová

(opakující se) sekvence – transpozony a retropozony, které mají schopnost se autonomně replikovat a „skákat“ po genomu.

V r. 2000 byla publikována první v podstatě úplná sekvence rostlinného genomu o délce 125 Mb (milionů párů bází) u modelového druhu huseníčku rolního (Arabidopsis Genome Initiative 2000; viz např. Živa 2007, 1: 5–7). Patří mezi velmi malé rostlinné genomy, přesto jeho sekvenování trvalo více než 10 let, stálo miliony dolarů a účastnily se ho desítky laboratoří. Postupovalo se klasickou cestou tzv. Sangerova sekvenování (podle jeho objevitele britského biochemika Fredericka Sangera). DNA se rozdělila na malé části, ty se naklonovaly v bakteriálních vektorech a ještě po menších úsecích se postupně sekvenovaly. Jedna sekvenační reakce přečetla jen asi 1 000 bp, takže pro celý genom bylo třeba provést nejméně 125 tisíc takových reakcí. Ve skutečnosti jich muselo být mnohem více, protože tisíce kousků bylo nutno správně seřadit za sebou, k čemuž napomohly sekvence vzájemně se překrývajících úseků. Velké problémy při sekvenování způsobuje repetitivní DNA, která se vyskytuje na stovkách až tisíci místech genomu, a my netušíme, kam danou repetitivní sekvenci do genomické mapy správně zasadit. Není divu, že další úplná sekvence rostlinného genomu dala na sebe čekat – v r. 2005 byl publikován genom rýže o velikosti 466 Mb, tedy podstatně větší než u huseníčku rolního, avšak stále patřící mezi malé rostlinné genomy (International Rice Genome Sequencing Project 2005).

Praktický zájem velel zaměřit pozornost na zemědělské plodiny. Ty jsou ale většinou polyploidní a na rozdíl od rýže, která představuje ve velikosti genomu mezi plodinami výjimku, mají velké genomy. Protože s rostoucí délkou genomu se výrazně zvyšuje podíl repetitivních sekvencí, postupovalo sekvenování jen pomalu. V případě hexaploidního genomu pšenice pomohlo třídění chromozomů, které vyvinula laboratoř Jaroslava Doležela v Ústavu experimentální botaniky AV ČR, v. v. i. (viz Živa 2012, 4: 155–158). To umožnilo rozdělit sekvenování celého genomu na sekvenování jednotlivých chromozomů, čímž se snížila složitost mapování.

Skládání genomu

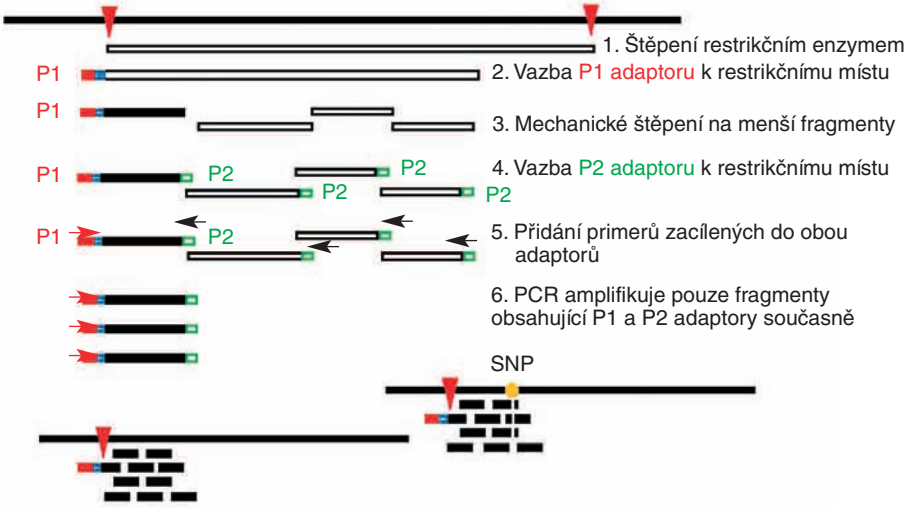
Zásadní převrat přinesl rozvoj metod sekvenování dalších generací. Pro klasické Sangerovo sekvenování je třeba každý fragment DNA připravit zvlášť, sekvenátor čte v rámci jednoho naprogramovaného běhu postupně maximálně 96, nebo 384 úseků DNA (podle počtu jamek v obdélníkových destičkách 8 × 12, resp. 16 × 24). Nové sekvenační postupy umožňují přečíst v jednom běhu současně statisíce až miliony DNA fragmentů. Využívají polymerázovou řetězovou reakci (PCR; např. Živa 1999, 3: 101–104) stejně jako klasické sekvenování, avšak reakční směs pro jednotlivé fragmenty není třeba individuálně připravovat lidským či robotickým pipetováním. Replikace probíhá v mikroskopických kapkách emulze nebo na čipu, kde jsou fragmenty DNA upevněny na obou koncích jako můstky.

Teoreticky tak lze získat sekvenci pokrývající libovolný genom za několik dní. Informací tedy máme ohromné množství a základní význam nabývá jejich zpracování – analýza dat. Není např. snadné hned na počátku rozhodnout, která sekvence je dostatečně spolehlivá a která obsahuje příliš mnoho chyb. Zatímco u klasického postupu hodnotí kvalitu sekvence experimentátor zkušeným okem, prohlédnout miliony sekvencí vychrlených za jediný den již není možné. Musíme tedy vzít zavděk počítačovým zpracováním a nastavit filtr, kterým projdou pouze kvalitní sekvence. Síto však nesmí být příliš husté, abychom zbytečně neodstranili cenná data. Většinu času nad množstvím sekvencí tráví badatel nastavováním a ověřováním parametrů jednotlivých výpočetních kroků.

Máme tedy několik desítek milionů jednotlivých sekvencí (čtení, anglicky read), pocházejících z krátkých fragmentů rostlinného genomu. Jejich délka činí 100 až 400 bp, podle použité metody. Většinou pocházejí z obou konců DNA fragmentů o známé délce např. 3 000 bp (tzv. párových čtení, paired end reads), nesou zakódovanou informaci, že patří k sobě. Víme, že např. čtení 90KRF páruje s 90KRR, obě vznikla sekvenováním téhož fragmentu 90KR, a v genomu proto musejí být tyto sekvence umístěny zhruba 3 000 bp od sebe. Pak již postačí vložit ty miliony čtení do patřičného programu a počkat několik hodin až dní. Jak počítač krátké fragmenty skládá? Používá některý z jednoduchých algoritmů, jako např. princip překryvu okrajových částí jednotlivých čtení (obr. 2). Řadí pak fragmenty za sebou, jako se k sobě skládají kostky domina, nebo hledá ve čteních krátké definované úseky (tzv. k-mers), jež opět sestavuje za sebou. Delší sekvence vzniklá složením několika čtení se nazývá kontig. V ideálním případě, který ovšem nenastane, by měl počítačový program vyprodukovat tolik kontigů, kolik má daný genom chromozomů. Realitou však je výstup v podobě tisíce kontigů, které můžeme k sobě zase dále přiřazovat do tzv. lešení (scaffolds). Obrovský problém znamenají repetitivní sekvence, např. retropozony, ale také duplikáty genů, velmi časté v rostlinných genomech.

Základní potíž právě popsaného postupu je příliš malá délka jednotlivých čtení, která prakticky nelze správně seřadit do podoby dlouhé nepřerušované sekvence DNA představující jeden chromozom. Řešení přinášejí např. postupy umožňující nepřetržitě číst dlouhé vlákno DNA. V současnosti se šíří metoda SMRT sekvenování (Single Molecule Real Time Sequencing). Spočívá v přímém sledování vkládání jednotlivých nukleotidů v průběhu polymerace, prováděné fixovanou molekulou polymerázy na jednovláknové předloze. Oproti předchozím technikám nevyužívá PCR. Již se uplatňuje metoda čtení písmen vlákna DNA procházejícího nanopórem, aniž by docházelo k syntéze nového řetězce (MinION nanopore sequencer).

Brzy tedy nebude třeba komplikovaně skládat kontigy, dlouhé sekvence se stanou primárním výstupem sekvenování. Dosavadní výsledky ale vykazují vysoké procento chyb, a tak musíme ve skládání krátkých čtení vytrvat a používat doplňkové metody. Jednou z nich je optické mapování. Speciálním postupem se izoluje DNA složená z dlouhých vláken, která se jednotlivě upevní v nanokanálcích. Poté na ni působí restrikční nukleáza (např.



2 Princip metody Restriction site Associated DNA markers sequencing (RAD-seq). DNA zkoumaných jedinců fragmentujeme pomocí restrikčních enzymů (1). Ke koncům vzniklým při štěpení připevníme krátké úseky – DNA adaptory (2). Takto opracované fragmenty mechanicky rozštípeme na menší kousky (3). Pak připevníme jiné adaptory ke koncům všech fragmentů (4). Přidáme primery specifické k prvním (P1) a druhým (P2) adaptorům (5) a provedeme polymerázovou řetězovou reakci (PCR), která pomnoží jen fragmenty obsahující oba adaptory. Tím získáme úseky DNA pocházející z okolí restrikčních míst (6), které se sekvenují např. pomocí metody Illumina. Výsledná čtení (krátké úseky) mapujeme na referenční genom (7). Porovnáme navzájem sekvence z různých jedinců a určíme místa (např. záměny jednoho nukleotidu – Single Nucleotide Polymorphism, SNP), kde se sekvence liší. Takto získáme reprezentativní genetickou informaci z celého genomu, aniž bychom ho museli celý sekvenovat.

SbfI), jež rozpoznává oktamery (8 bp) a DNA v těchto specifických motivech rozštípe. Místa střihu lze vidět pomocí mikroskopu a fluorescenčního barvení. Můžeme tak sestavit mapu vzdáleností mezi restrikčními místy, sloužící jako vodítko pro správné řazení kontigů. Restrikční místo v sekvenci snadno rozpoznáme, a tak lze porovnat mapu definovanou na základě navržené sekvence s mapou určenou přímo optickým mapováním.

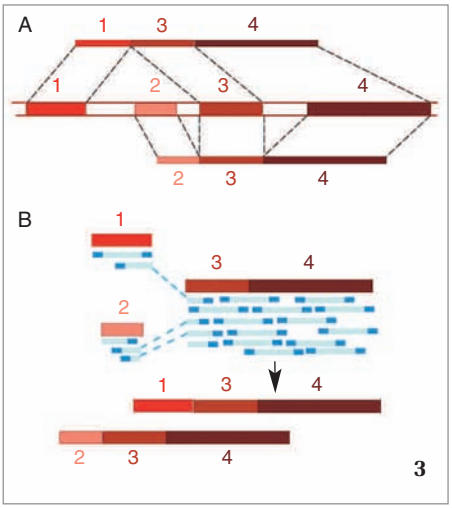
I neúplná informace hodně prozradí

Jakkoli představuje sekvenování dalších generací obrovský pokrok a zrychlení, je stále velmi pracné získat sekvence genomů mnoha jedinců v populacích za účelem vzájemného srovnání. Můžeme však zvolit postup, který přinese informaci jen o vybraných částech genomu, jako bychom se v mapě krajiny zaměřili pouze na určité řeky nebo horské hřebeny. Toto umožňuje metoda RAD-seq (Restriction site Associated DNA markers sequencing), která podává informace o sekvencích DNA v sousedství vybraných restrikčních míst, třeba právě oktamerů rozpoznávaných enzymem *SbfI* (obr. 2). Je důležité, aby sekvenované úseky byly vhodně rozmístěny napříč celým genomem, a pomohly tak např. odhalit oblasti, na něž působí selekční tlak a jsou tedy důležité pro adaptaci. Metodu RAD-seq můžeme použít i u druhů, kde dosud kompletní genomická sekvence chybí a lze ji též vhodně aplikovat na populační vzorky.

Další způsob, jak nahlédnout do důležitých oblastí genomu, poskytuje analýza transkriptomu, tedy souboru transkriptů. V zásadě jde o stanovení sekvence všech molekul RNA vzniklých přepisem (transkripcí) kódujících oblastí genomu – úseků, kde je uložena informace pro bílkovinu či strukturní nebo regulační RNA. Protože velkou většinu rostlinného genomu tvoří nekódující DNA, která se málo, pokud vůbec, přepisuje, transkriptom odpovídá pouze malému dílku genomu. Navíc se v daném vzorku např. listu, květu nebo kořene přepisuje jen část kódujících genů. V kořenu není potřeba chlorofylu, proto zde nejsou aktivní geny kódující enzymy nutné pro jeho syntézu. A v listech se zase nepřepisují geny pro tvorbu kořenových vlásků nebo základů květních orgánů. Transkriptom proto ze své podstaty není úplný. I kdybychom metodicky dokonale zvládli přečíst každou jednotlivou molekulu RNA, což je ideální, ale nedosažitelný stav, obrázek všech genů stejně nezískáme, protože v žádném orgánu nebo pletivu rostliny se neexprimují všechny geny. Smíchání RNA z rozmanitých pletiv pomůže jen zčásti, protože rostlina rezervuje značný díl své genetické výbavy pro odpověď na stres – sucho, horko, mráz i napadení hmyzem. Tím vším bychom na rostlinu museli působit, aby aktivovala také „stresové“ geny a abychom se přiblížili osekvenování všech kódujících úseků genomu.

Nicméně i neúplný transkriptom obsahuje obrovské množství genových sekvencí. Pokud izolujeme vzorek RNA ze stejné části rostliny (např. z mladého listu) a odebereme ho u všech jedinců ve stejnou dobu (kořenářky dobře věděly, proč některé byliny trhat jedině za úplňku, kdy obsahují vysoké hladiny metabolitů – transkripce genů se v průběhu dne a noci dramaticky mění), získáme srovnatelné soubory dat. Lze je využít dvěma způsoby:

- Srovnáváme sekvence homologických (navzájem si odpovídajících) genů u jedinců, populací, poddruhů apod. Počítačové programy naleznou mezi desítkami tisíc genů ty nejvíce proměnlivé, korelující s nějakým faktorem prostředí (např. zeměpisnou šířkou) nebo jinak zajímavé.
- Analýza transkriptomu poskytuje kromě sekvencí transkriptů informaci o jejich početnosti. Vysoká četnost svědčí o vysoké transkripci, a tedy i expresi daného genu. Rozsah genové exprese je však velmi proměnlivý a zatím se používá především pro srovnání rostlin pěstovaných za definovaných podmínek (např. v růstové komoře). V budoucnu jistě najde uplatnění také při studiu rostlin v terénu.



3

3 Transkriptom je soubor všech molekul RNA z daného pletiva nebo jedince. K jeho získání musíme nejprve reverzně přepsat RNA na komplementární DNA (cDNA), která se pak sekvenuje a jednotlivá čtení se skládají obdobně jako u genomu. Jeden kontig odpovídá jednomu transkriptu, tedy jedné určité molekule mediátorové RNA (mRNA). Skládání transkriptomu komplikuje alternativní sestřih. A – Gen tvořený čtyřmi exony je přepsán a alternativně sestřižen do podoby dvou různých transkriptů. Jeden z nich obsahuje exon 1 a druhý má místo něj exon 2. B – Při skládání transkriptomu narazíme na párová čtení, která mají konce v odlišných kontigách. Čtení na obr. nemůžeme seskládat do jediného kontigu. Musíme sestavit dva kontigy, které v našem případě odpovídají dvěma alternativním transkriptům. Pokud není k porovnání referenční genom, vzniká v tomto kroku mnoho chyb, kdy program generuje neexistující transkripty. Rychle se zlepšující metody dlouhých čtení a pokroky bioinformatických postupů umožňují spolehlivě složit transkriptomy i u nemodelových druhů. Polyploidní rostliny však stále představují nejtvrdší oříšek. Všechny orig.: H. Štorchová

4 Lokalita bublinatky obecné v rezervaci La Petite Camargue Alsacienne na hranici mezi Francií a Švýcarskem. Horní tok Rýna zde meandruje a vytváří četná slepá ramena a tůně, zarůstající orobincem (*Typha*) a vodními rostlinami (např. prustka obecná – *Hippuris vulgaris*, voďanka žabí – *Hydrocharis morsus-ranae*). Toulky přítmím lužního lesa jsou stejně dobrodružné jako objevování evoluční minulosti v genomech bublinatek. Foto Z. Ragettli

Sestavování transkriptomu probíhá obdobně jako v případě genomu (obr. 3). Jen počet kontigů je mnohem vyšší, protože jeden kontig odpovídá přibližně jednomu transkriptu, tedy molekule RNA přepsané z jednoho genu. U rostlin bývá navíc z jednoho genu odvozeno více transkriptů, což je dáno alternativním sestřihem. Kódující úseky genu – exony – nemusejí být využity všechny, ale jsou různým způsobem kombinovány, přičemž jeden exon výrazně mění sekvenci transkriptu a strukturu kódovaného proteinu. Skládání transkriptomu je tedy komplikované a méně jednoznačné než skládání genomu. Jednotlivá čtení odvozená z cDNA (komplementární DNA vzniklé zpětným přepisem RNA) se řadí za sebou podle překryvů podobnými algoritmy jako při skládání genomů. Počet čtení tvořících jeden kontig je tím větší, čím více molekul RNA dané sekvence transkriptom obsahuje. Tato tzv. pokryvnost pak odpovídá aktivitě genu, tedy jeho expresi.

Transkriptom je souborem sekvencí přepsaných v drtivé většině z kódujících oblastí genomu. Z jediného vzorku RNA získáme velmi podrobný, byť ne zcela úplný přehled o desítkách tisíc genů kterékoli rostliny. Analýza transkriptomů přináší nepřeberné množství poznatků. Můžeme objevit geny získané, nebo naopak ztracené v průběhu evoluce. Srovnání sekvencí odhalí geny evolučně konzervované, či naopak dramaticky změněné selekčním tlakem na danou populaci, ekotyp nebo druh. V příštím čísle představíme zajímavé poznatky, které přinesla výprava do genomické a transkriptomické krajiny bublinatky *U. gibba* a b. obecné (*U. vulgaris*).

Studium transkriptomu bublinatky obecné bylo podpořeno grantem GA ČR P504/11/0783, za spolupráce Botanického ústavu AV ČR, v. v. i., Ústavu experimentální botaniky AV ČR, v. v. i., a Přírodovědecké fakulty Jihočeské univerzity v Českých Budějovicích.

Použitá literatura uvedena na webu Živy.


4