Bioinformatika a sekvenování nové generace

Bioinformatics and Next-generation Sequencing

Krejčí A., Müller P., Vojtěšek B.

Regionální centrum aplikované molekulární onkologie, Masarykův onkologický ústav, Brno

Souhrn

Technologie sekvenování DNA nové generace mají v současné době nezastupitelné místo ve výzkumu a postupně nacházejí cestu i do oblasti klinické praxe. Sekvenační přístroje produkují velké množství dat, jejichž analýza metodami bioinformatiky je nezbytná k získání relevantních výsledků. Sekvenování se tak bez pokročilého výpočetního zpracování specializovanými algoritmy naprosto neobejde. V tomto přehledu jsou představeny základní koncepty výpočetního zpracování sekvenačních dat s přihlédnutím ke specifickým aspektům oblasti onkologie. Rovněž jsou uvedeny nejčastější problémy a překážky komplikující zpracování a biologickou interpretaci výsledků.

Klíčová slova

bioinformatika – technologie masivně paralelního sekvenování – mutace – onkologický výzkum – klinická aplikace

Summary

Next-generation sequencing technologies are currently well-established in the research field and progressively find their way towards clinical applications. Sequencers produce vast amounts of data and therefore bioinformatics methods are needed for processing. Without computational methods, sequencing would not be able to produce relevant biological information. In this review, we introduce the basics of common NGS-related bioinformatics methods used in oncological research. We also state some of the common problems complicating data processing and interpretation of the results.

Key words

bioinformatics – high-throughput nucleotide sequencing – mutations – cancer research – clinical application

Práce byla podpořena Evropským fondem pro regionální rozvoj a státním rozpočtem České republiky (RECAMO, CZ.1.05/2.1.00/03.0101), projektem MŠMT – NPU I – LO1413, MZ ČR – RVO (MOÚ, 00209805) a BBMRI_CZ (LM2010004).

This study was supported by the European Regional Development Fund and the State Budget of the Czech Republic (RECAMO, CZ.1.05/2.1.00/03.0101), by the project MEYS – NPS I – LO1413, MH CZ – DRO (MMCI, 00209805) and BBMRI_CZ (LM2010004).

Autoři deklarují, že v souvislosti s předmětem studie nemají žádné komerční zájmy.

The authors declare they have no potential conflicts of interest concerning drugs, products, or services used in the study.

Redakční rada potvrzuje, že rukopis práce splnil ICMJE kritéria pro publikace zasílané do biomedicínských časopisů.

The Editorial Board declares that the manuscript met the ICMJE "uniform requirements" for biomedical papers.

RNDr. Bořivoj Vojtěšek, DrSc.
Regionální centrum aplikované molekulární onkologie
Masarykův onkologický ústav
Žlutý kopec 7
656 53 Brno
e-mail: vojtesek@mou.cz

Obdrženo/Submitted: 21. 4. 2015
Přijato/Accepted: 26. 6. 2015

http://dx.doi.org/10.14735/amko20152S91

Úvod

Objevem funkce DNA v roce 1944 [1] začala snaha o přečtení a rozluštění dědičné informace. Objev struktury DNA v roce 1953 [2] objasnil princip, jakým je informace uchovávána a přenášena do dalších generací. První metody umožňující relativně účinné čtení dědičné informace, neboli sekvenování DNA, se však objevily až koncem 70. let 20. století [3,4]. Vývoj těchto metod umožnil v 90. letech 20. století započít sekvenování lidského genomu [5], v té době extrémně náročný projekt těžko představitelného rozsahu. Po roce 2000 se začaly objevovat první metody masivně paralelního sekvenování, označované jako sekvenování nové generace (next-generation sequencing – NGS) [6]. Ty se vyvíjejí dodnes a v současné době koexistují se zcela novými přístupy ke čtení DNA, někdy označovanými jako sekvenování třetí generace [7].

Bioinformatika jako obor zabývající se zpracováním rozsáhlých molekulárněbiologických dat včetně sekvencí DNA byla nezbytná již od počátků sekvenování. Důležitost výpočetních přístupů dobře ilustrují počáteční neúspěchy vědců při snaze sestavit kompletní genomy i velmi jednoduchých organizmů, způsobené právě neexistencí vhodných počítačových programů. Teprve v roce 1995, pět let po zahájení projektu s cílem sestavit lidský genom, se podařilo sestavit první kompletní genom volně žijícího organizmu, bakterie *Haemofilus influenzae* [8]. Skupina Craiga Ventera tehdy vyvinula program TIGR assembler [9], který dokázal sestavit 24 000 fragmentů sekvence bakteriální DNA ve správném pořadí a zkompletovat tak celý 1,8 milionů bází dlouhý chromozom. Zmíněná čísla ukazují, že i při práci s jednoduchými organizmy ruční zpracování dat nepřipadá v úvahu a že sekvenování DNA se bez bioinformatiky neobejde.

V dnešní době jsou pokročilé metody sekvenování hojně využívány ve výzkumu a své místo si postupně nacházejí i v klinické praxi [10]. Přístroje jsou schopné vyprodukovat enormní množství dat, která jsou však bez pokročilého bioinformatického zpracování bezcenná.

Principy sekvenování nové generace

Základní myšlenkou metod NGS je tzv. masivně paralelní sekvenování. Jde o postup, při kterém je během jednoho experimentu v jednom okamžiku zároveň sekvenováno velké množství samostatných molekul. Vstupní DNA (genomická nebo cDNA získaná reverzní transkripcí) je fragmentována na úseky o délce typicky několika desítek až stovek párů bází a v případě potřeby amplifikována pomocí polymerázové řetězové reakce (polymerase chain reaction – PCR). Tyto krátké fragmenty jsou poté čteny a zpracovávány paralelně. Společným znakem těchto metod je čtení velkého množství (v řádu jednotek milionů až jednotek miliard) fragmentů – „readů" (pojem někdy též překládaný jako „čtení"), které jsou však relativně krátké (typicky v řádu stovek bází). Takto vysoká míra paralelizace je hlavní rozdíl mezi NGS a klasickým Sangerovým sekvenováním, při kterém je zároveň čteno typicky max. 96 molekul DNA.

Princip samotného určování bází DNA dnes spočívá nejčastěji v tzv. sekvenování syntézou. Většina metod využívá DNA polymerázu postupně syntetizující vlákno komplementární k jednovláknovému templátu sekvenovaného fragmentu. Použité chemické sloučeniny zajistí, že těsně po začlenění každé nové báze do řetězce DNA je přímo z místa začlenění uvolněn určitý druh signálu, který je následně zachycen technickým vybavením přístroje. Nejpoužívanější podobou signálu je světlo, existují však i technologie využívající detekci uvolněných protonů H^+ [11]. Zmíněný princip v různých obměnách implementují všechny v praxi používané technologie – přístroje značky Illumina (v dnešní době jednoznačně nejpoužívanější), Ion Proton firmy Life Technologies, Roche 454 (v současnosti se již přestává používat), ABI SOLiD (místo syntézy komplementárního vlákna využívá ligaci krátkých fragmentů) či přístroj RS II firmy Pacific Biosciences. Ve fázi vývoje jsou i přístroje využívající k určení sekvence DNA zcela odlišné přístupy, ovšem k jejich rutinnímu nasazení v praxi zatím stále nedošlo (asi nejslibnější novou technologií disponuje firma Oxford Nanopore – sekvenování je založeno na detekci změn v elektrickém proudu procházejícím přes mikroskopické póry, kterými jsou translokovány molekuly DNA).

Při NGS experimentech je důležité určit požadovanou tzv. hloubku (depth, někdy označováno také jako pokrytí – coverage, či hloubka pokrytí – depth of coverage). Jedná se o průměrný počet kopií vzorku DNA, které sekvenováním získáme. Například při hloubce 30krát by se měl s největší pravděpodobností každý úsek sekvenované DNA vyskytovat v 30 různých readech, tedy informaci o tom, jaká báze se na konkrétním místě nachází, dostaneme v 30 kopiích. Větší pokrytí jednak zajistí eliminaci chyb přístrojů, které se vždy nevyhnutelně vyskytují, jednak zvýší citlivost metody – pokud sekvenujeme DNA ze směsi geneticky odlišných buněk, je možno vyšším pokrytím dosáhnout detekce vzácných variant, které se vyskytují pouze ve zlomku zkoumaných buněk. Na rozdíl od ideální představy ovšem pokrytí v praxi nebývá ani zdaleka rovnoměrné. Při průměrném pokrytí 30krát budou pravděpodobně stále existovat úseky, o nichž nebudeme mít žádné nebo téměř žádné informace, sekvence jiných oblastí bude naopak obsažena ve stovkách readů. Tato nevyváženost může být způsobena mnoha různými faktory, jako jsou rozdílná míra amplifikace pomocí PCR, rozdíly v účinnosti hybridizačních sond při cíleném sekvenování či vliv repetitivních oblastí.

Nejčastější aplikace sekvenování

První sekvenační projekty měly za cíl sestavit do té doby neznámé genomy různých organizmů. Takový přístup nazýváme *de novo* sekvenování genomu. V případě, že referenční sekvence genomu zkoumaného organizmu byla již dříve sestavena, hovoříme o resekvenování. Referenční sekvence jsou dnes k dispozici pro mnoho organizmů včetně člověka, což umožňuje široké využití resekvenování. Cílem celogenomového resekvenování je hledání krátkých variant v kódujících oblastech genů i v těch nekódujících, např. pro-

motorových a regulačních oblastech, případně detekce rozsáhlých chromozomových přestaveb a hledání strukturních variací typu CNV (copy number variations).

Další častou podobu experimentu představuje exomové sekvenování, při kterém je sekvenována pouze DNA kódujících úseků – exonů. Sekvenační knihovna je v tomto případě vytvořena separováním fragmentů genomové DNA pomocí hybridizace s fragmenty obsahujícími sekvence komplementární k exonové DNA. Tato metoda tedy slouží primárně k detekci somatických mutací a polymorfizmů v kódujících oblastech genů. Výhodou metody je značná úspora kapacity i ceny oproti sekvenování celého genomu.

V některých případech je třeba určit sekvenci pouze několika málo genů či jejich částí. V takových situacích je použito tzv. cílené (targeted) sekvenování, při kterém je izolována, amplifikována a osekvenována pouze DNA z přesně definovaných oblastí. Cílené sekvenování má důležité klinické aplikace v onkologické diagnostice. Menší počet přesně definovaných úseků je totiž možno za relativně nízkou cenu osekvenovat do velké hloubky, běžná je i hloubka více než 1 000krát. Lze tak detekovat i somatické mutace v komplexním biologickém vzorku, jakým je vzorek nádoru obsahující směs zdravých a nádorových buněk. Citlivost této metody umožňuje detekovat mutace s frekvencí četnosti pod 1 %, což je citlivost, které klasické Sangerovo sekvenování zdaleka nedosahuje.

Velmi rozšířená aplikace NGS je i sekvenování RNA, které se využívá pro kvantitativní i kvalitativní analýzu mRNA, miRNA nebo cílenou sekvenaci variabilních oblastí ribozomální RNA. V tomto experimentu je izolována buď celková RNA, nebo pomocí hybridizačních metod specifická buněčná RNA (např. miRNA nebo mRNA). Pomocí reverzní transkripce je RNA přepsána do cDNA, která je následně osekvenována. Metoda je využívána především k měření intenzity exprese jednotlivých genů, k detekci sestřihových variant a genových fúzí. Teoreticky je možno transkriptomovým sekvenováním detekovat i germinální a somatické mutace. Tato možnost se však příliš nevyužívá, jelikož sekvenování DNA je pro tento účel vždy přesnější [12].

Bioinformatické přístupy v sekvenování

Bioinformatika je nezbytná součást všech fází vyhodnocení sekvenačních dat, od zpracování signálu v přístroji po finální studium výsledků.

Zpracování dat experimentu resekvenování

Ať jde o celogenomové, exomové či cílené resekvenování, základní postup zpracování dat je vždy stejný. Nejdřív je třeba mapovat všechny ready na referenční genom, tedy najít pokaždé tu část referenční sekvence, která odpovídá sekvenci fragmentu. V dalším kroku je potřeba nalézt a spočítat všechny odlišnosti nasekvenované DNA od referenční sekvence či rozdíly mezi jednotlivými ready mapovanými na stejnou oblast genomu. Na základě pokročilých statistických výpočtů je určena pravděpodobnost, s jakou se na pozicích vyskytují skutečné polymorfizmy a s jakou pravděpodobností jde jen o chyby v sekvenaci či v mapování readů. V případě cíleného sekvenování v diagnostice je požadovaným výstupem často pouze informace o přítomnosti či nepřítomnosti známých variant, ve výzkumu jsou data o variantách dále zpracovávána dalšími metodami, např. se snahou odlišit řídící (driver) od následných (passenger) mutací v nádorových buňkách.

Referenční genom

Přestože se na referenčním lidském genomu pracuje již od roku 1990, kompletní a přesná sekvence stále není známa. Problém je především s určením správného počtu kopií ve vysoce repetitivních oblastech, např. v okolí centromer chromozomů. Důvodem je nedostatečná délka sekvenačních readů – v případě mnohonásobného opakování krátkého úseku DNA, kdy je výsledná repetice podstatně delší než ready, které umí přístroje vytvořit, je sice možné získat ready začínající levou hranicí repetitivní oblasti a jiné ready končící její pravou hranicí, všechny ready mezi těmito oblastmi však budou vzhledem k repetitivní povaze oblasti téměř nebo zcela totožné. Počet readů lze k určení délky repetice využít jen velmi omezeně, takže jediná informace, kterou takto získáme, bude fakt, že repetitivní oblast je více než dvakrát delší než jeden read. Další problémy nastávají u větších skupin sekvenčně velmi podobných genů a pseudogenů, jako jsou např. geny čichových receptorů. Pokud jsou např. rozdíly v sekvenci genů tak vzácné, že jeden sekvenační read nese pouze jeden (bodový) rozdíl, není možné skládáním readů určit, do kterého z genů který rozdíl patří.

Díky pokrokům v technologii sekvenování se délka readů i jejich přesnost neustále zvětšují, a proto se i sekvence referenčního genomu neustále vyvíjí. První verze referenčního genomu obsahovala 150 000 regionů s neznámou sekvencí. Devatenáctá verze z roku 2009 už obsahovala pouze 357 takových mezer. Nejnovější verze, známá jako hg38 nebo GRCh38, pochází z prosince roku 2013 a mimo jiné jako první obsahuje i přibližné odhady počtu centromerických repetic [13].

Mapování na referenční genom

Hledání pozice na referenční sekvenci, která přísluší sekvenci readu, je obtížné. Nejjednodušší způsob, jak k němu přistoupit, je porovnat read postupně se všemi pozicemi v genomu. Takový přístup by pro genom délky a a read délky b vyžadoval přibližně $a \times b$ porovnání bází. Situace je ovšem ještě složitější, protože jednoduchým porovnáním všech možných pozic nebereme v úvahu možnost existence mezer – pokud by read obsahoval malou deleci či inerci, jeho správnou pozici bychom tímto způsobem vůbec nenašli. Existují ovšem algoritmy, které dokážou najít optimální zarovnání sekvencí včetně mezer taktéž s použitím $a \times b$ kroků. Problém mapování na referenční genom je ovšem tak rozsáhlý, že ani tyto algoritmy nelze jednoduše využít. Pokud bychom např. provedli resekvenování lidského genomu s hloubkou 20krát, získali bychom zhruba 600 milionů readů o délce 100 bp. Pro namapování každého z nich by bylo potřeba asi 3×10^{11} kroků, dohromady, tedy $1{,}8 \times 10^{20}$ porovnání nukleotidů. Sou-

časné procesory pracují s frekvencí okolo 4 GHz, tudíž provedou 4×10^9 tiků za sekundu. Kdyby procesor zvládl v jednom tiku vykonat jedno porovnání nukleotidů, trvalo by mapování genomu $4{,}5 \times 10^{10}$ sekund, tedy 1 427 let.

Proto se při mapování používá několik technik výrazně zkracujících potřebný čas. Hlavní technika je indexování. Nad sekvencí genomu je vytvořen index čili rejstřík. Ten je potom využit při prohledávání genomu velmi podobným způsobem, jakým člověk využívá rejstřík v knihách. Tvorba indexu je náročná, ale stačí jej pro každý genom vytvořit jednou a poté je možno jej použít pro všechna budoucí mapování. V současnosti nejpoužívanější technikou tvorby rejstříku je tzv. Burrowsova-Wheelerova transformace, která je využívána mimo jiné také v kompresi dat. Další urychlení je dosaženo snížením požadavků na přesnost zarovnání genomu a readu – nehledá se vždy optimální zarovnání, což postup velmi urychlí, ovšem za cenu možnosti výskytu malého množství chybně namapovaných readů [14].

Určování variant

Ve chvíli, kdy jsou ready namapovány na referenční genom, je možno určit přesnou sekvenci DNA celého vzorku. V nejjednodušším případě bodového polymorfizmu v podobě substituce by (u heterozygotního organizmu) měly v místě homozygotního genotypu obsahovat všechny ready namapované k danému úseku genomu stejný nukleotid, v místě heterozygotního by polovina readů měla obsahovat jednu variantu a polovina druhou. Situaci však komplikuje fakt, že sekvenační přístroje produkují nezanedbatelné množství chyb a jak již bylo zmíněno, některé ready mohou být chybně zarovnány. V důsledku těchto faktorů tak mohou být určeny falešně pozitivní varianty v místech, kde se ve skutečnosti nevyskytují. Aby byla pravděpodobnost chyby snížena, využívá se pokročilé statistiky (v případě bodových mutací jde typicky o bayesovské metody) k určení pravděpodobnosti výskytu skutečné varianty a chyby [15]. Ke zpřesnění výsledků může také posloužit analýza více vzorků zároveň – může např. existovat varianta, která je vzácná nebo se vyskytuje v oblasti s typicky nízkou hloubkou pokrytí, a proto se v každém vzorku jeví spíše jako chyba přístroje. Výskyt ve více nezávislých vzorcích ovšem potvrzuje, že se jedná o skutečnou variantu, jelikož pravděpodobnost opakování stejné chyby je malá [16].

V případě nádorových vzorků je situace komplikovanější, protože DNA často pochází z komplexní směsi geneticky odlišných buněk. Je proto třeba využít sekvenování s velkou hloubkou a upravit parametry statistických metod, protože i bodové rozdíly v malé části readů mohou znamenat skutečnou mutaci, která by jinak mohla být nesprávně považována za chybu sekvenačního přístroje. Pro odlišení somatických mutací od zárodečných variant je vhodné sekvenovat kromě nádorové tkáně i zdravou, která poslouží jako negativní kontrola.

Určování krátkých delecí a inzercí je náročnější problém než určování bodových substitucí, protože v místech, kde se vzorek liší délkou od referenčního genomu, je větší pravděpodobnost chybně namapovaných readů. Situace se často řeší lokálním provedením časově náročnějšího a přesnějšího zarovnání readů pouze v místech možných delecí a inzercí. U nádorových vzorků je situace opět komplikovanější, ze stejných důvodů jako v případě bodových substitucí.

Nejnáročnější úkol je detekce rozsáhlejších strukturních variací, jako jsou delece, změny počtu kopií (CNV) či translokace. Izolovanou událost většího rozsahu je možno v dobře definovaném vzorku detekovat na základě porovnání teoretické a skutečné hloubky pokrytí v postižené oblasti. Pokud je k některé části referenčního genomu namapováno nezvykle mnoho readů, mohl být tento úsek ve vzorku amplifikován. Je-li naopak např. hloubka pokrytí některého chromozomu poloviční než u ostatních, může se jednat o monoploidii. Další možnost je využít informace získané speciálním protokolem sekvenování (techniky nazývané paired-end a mate-pair sequencing). Tento postup využívá technických možností některých sekvenačních přístrojů, které jsou schopny každý fragment DNA číst z obou stran. Pokud např. fragmentujeme sekvenovanou DNA na úseky dlouhé 1 000 bází, přístroj bude schopen přečíst 100 bází z každého konce fragmentu. Ready pocházející z jednoho fragmentu potom vytvoří pár. Jelikož délka fragmentů DNA byla 1 000 bází, můžeme předpokládat, že mezi koncem prvního a začátkem druhého readu z jednoho páru by vždy mělo ležet 800 bází. Najdeme-li po namapování na referenční genom ready z jednoho páru v neobvyklé vzdálenosti, můžeme určit, že v oblasti 800 bází mezi konci readů došlo ke strukturní změně [17].

V některých situacích mohla určitá část sekvenovaného genomu projít tolika strukturními změnami, že ready, které z ní pocházejí, není vůbec možno namapovat na referenční genom, protože rozdíl skutečné sekvence proti referenční je příliš velký. V těchto případech je často jedinou šancí pokusit se takové značně degenerované oblasti vzorku sestavit z readů *de novo*, tedy použít metody, které se běžně používají k prvotnímu určení sekvence genomu neznámého organizmu.

Obecně lze říci, že určování variant v genomech, které se příliš neliší od referenční sekvence, je v současné době poměrně dobře zvládnutý problém. Situace je ovšem odlišná v případě rakovinných buněk, které jsou často geneticky nestabilní a jejich genomy podléhají velkému množství rozsáhlých změn. Určení správné sekvence u takových vzorků je obecně velmi obtížné a cesta k jeho řešení zatím není u konce. Aplikace v onkologické praxi se omezují především na hledání předem známých anotovaných mutací v úzké skupině genů pomocí cíleného sekvenování s velkou hloubkou. Existují komerčně dostupné tzv. panely pro testování nejběžnějších mutací v několika genech, z nichž některé jsou dodávány i s bioinformatickým řešením šitým na míru konkrétnímu problému. Začínají tak být k dispozici metodiky založené na NGS, které jsou přesně definované od sběru vzorku po zpracování dat, a mají tak šanci konkurovat jiným, v klinické praxi už dobře ukotveným metodám.

Zpracování experimentu sekvenování transkriptomu

Při zpracování dat ze sekvenování mRNA je opět první nutný krok mapování readů na referenční genom, které je ovšem složitější než v případě sekvenování DNA. Důvodem je sestřih transkriptu. RNA je izolována a sekvenována až po sestřihu, tím pádem se v jednom readu mohou vyskytnout vedle sebe oblasti, které jsou v genomu odděleny dlouhým intronem. Mapování mRNA proto umožňuje vložení mezery reprezentující intron do libovolného místa readu a kromě sofistikovaných algoritmů pro hledání správného rozdělení readů mezi exony využívá i anotovaný referenční transkriptom. Situaci nadále komplikuje možnost alternativního sestřihu RNA – v transkriptu může dojít např. k vystřižení jednoho exonu společně se sousedními introny. Části readů sahající přes hranici sestřihu je potom třeba namapovat na dva exony, které spolu v genomu nesousedí [18].

Pokud je cílem experimentu hledat genové fúze, situace se dále komplikuje. V důsledku fúze se totiž do jednoho readu mohly dostat oblasti, které nejsou v genomu odděleny pouze kratšími oblastmi intronů – může se jednat i o velmi vzdálené regiony, nebo dokonce o oblasti z různých chromozomů. Při fúzích navíc může dojít k inverzím, které změní orientaci jedné ze spojených sekvencí. Algoritmy hledající genové fúze proto nejdříve identifikují ready, které není možno namapovat běžným způsobem. Ty potom rozdělí na kratší fragmenty a snaží se namapovat tyto fragmenty nezávisle, bez omezení jejich vzájemných pozic. Opětovným spojením fragmentů se získá informace o možných fúzích. Čím víc readů je namapováno obdobným způsobem do dvou oblastí genomu, tím větší je pravděpodobnost, že zde došlo ke genové fúzi. Na závěr je ještě třeba odfiltrovat všechny potenciálně falešné nálezy fúze sekvenčně podobných oblastí – ready mohly být namapovány z části na jednu, z části na jinou oblast genomu, ale pokud jsou toto např. oblasti dvou sekvenčně velmi podobných genů, je pravděpodobnější, že jde o chybu mapování a ready ve skutečnosti pochází celé buď z jednoho, nebo z druhého genu [19].

Nejčastějším cílem sekvenování transkriptomu je určení míry exprese jednotlivých genů. Po namapování software spočítá, kolik readů se mapovalo na který gen a tuto hodnotu normalizuje na délku genu a celkový počet readů. Výsledkem je hodnota označovaná jako RPKM (reads per kilobase per million), popř. FPKM (fragments per kilobase per million, tato hodnota se používá u paired-end sekvenování). Jedná se o číslo vypočtené nezávisle pro každý gen, které určuje míru jeho exprese. V případě sekvenování buněčných linií nebo laboratorně pěstovaných organizmů to může být dostačující výsledek, ovšem v případě experimentů s klinickými vzorky lidských tkání bývá většinou cílem spíše porovnání hladiny exprese mezi dvěma tkáněmi, např. zdravou a nádorovou [20]. Je proto potřeba provést a zpracovat dva experimenty, každý pro jeden druh tkáně a výsledky porovnat. Situaci komplikuje potenciálně vysoká míra variability mezi klinickými vzorky. Problém je způsoben tím, že hladina exprese některých genů je přirozeně silně variabilní, zatímco u jiných genů je velmi přísně regulována. Velký rozdíl v hladině exprese jednoho genu tak může být biologicky méně významný než řádově menší rozdíl v hladině exprese jiného genu [21]. Řešení spočívá ve sběru více dat a v pokročilé statistice. Experiment je potřeba replikovat – zpracovat několik nezávislých vzorků z obou tkání. Na základě rozdílů mezi hodnotami exprese v rámci vzorků z jednoho druhu tkáně je potom stanovena úroveň variability exprese každého genu a teprve poté jsou porovnány hodnoty mezi tkáněmi a výsledná významnost rozdílů v expresi je určena s přihlédnutím k naměřené variabilitě [22,23]. Přestože replikování experimentu značně zvyšuje jeho cenu, je to nejspíš jediný použitelný přístup, protože na výsledky experimentů bez dostatečné míry replikace se nelze spolehnout [24].

Závěr

NGS je v současné době hojně využíváno ve výzkumu a postupně si nachází cestu i do klinické praxe. Aplikace v klinické molekulární onkologii jsou zatím přes slibné prognózy poměrně omezené, mimo jiné z důvodu nedostatku přenositelných a dobře validovaných protokolů a metod zpracování sekvenačních dat. Sestavování a interpretace genetické informace nádorových buněk patří k vůbec nejsložitějším úkolům moderní genomiky a vývoj nových bioinformatických metod v této oblasti tak bude pro jejich uplatnění v klinické praxi naprosto nezbytný.

Literatura

1. Avery OT, Macleod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J Exp Med 1944; 79(2): 137–158.
2. Watson JD, Crick FH. The structure of DNA. Cold Spring Harb Symp Quant Biol 1953; 18: 123–131.
3. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977; 74(12): 5463–5467.
4. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A 1977; 74(2): 560–564.
5. Lander ES, Linton LM, Birren B et al. Initial sequencing and analysis of the human genome. Nature 2001; 409(6822): 860–921.
6. Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem 2013; 6: 287–303. doi: 10.1146/annurev-anchem-062012-092628.
7. Yanhu L, Lu W, Li Y. The principle and application of the single-molecule real-time sequencing technology. Yi Chuan 2015; 37(3): 259–268. doi: 10.16288/j.yczz.14-323.
8. Fleischmann RD, Adams MD, White O et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995; 269(5223): 496–512.
9. Sutton GG, White O, Adams MD et al. TIGR assembler: a new tool for assembling large shotgun sequencing projects. Genome Sci Technol 1995; 1(1): 9–19.
10. Xuan J, Yu Y, Qing T et al. Next-generation sequencing in the clinic: promises and challenges. Cancer Lett 2013; 340(2): 284–295. doi: 10.1016/j.canlet.2012.11.025.
11. Koubková L, Vojtěšek B, Vyzula R. Sekvenování nové generace a možnosti jeho využití v onkologické praxi. Klin Onkol 2014; 27 (Suppl 1): S61–S68. doi: 10.14735/amko2014S61.
12. Chien-Yueh L, Yu-Chiao C, Liang-Bo W. Common applications of next-generation sequencing technologies in genomic research. Transl Cancer Res 2013; 2(1): 33–45.
13. Human Genome Assembly Data [homepage on the Internet]. Genome Reference Consortium, Great Britain; [updated 2014 January 2; cited 2015 March 1]. Available from: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/data.
14. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009; 25(14): 1754–1760. doi: 10.1093/bioinformatics/btp324.
15. Mckenna A, Hanna M, Banks E et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010; 20(9): 1297–1303. doi: 10.1101/gr.107524.110.
16. Nielsen R, Paul JS, Albrechtsen A et al. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 2011; 12(6): 443–451. doi: 10.1038/nrg2986.

17. Chen K, Wallis JW, Mclellan MD et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 2009; 6(9): 677–681. doi: 10.1038/nmeth.1363.
18. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009; 25(9): 1105–1111. doi: 10.1093/bioinformatics/btp120.
19. Kim D, Salzberg SL. TopHat-fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 2011; 12(8): R72. doi: 10.1186/gb-2011-12-8-r72.
20. Trapnell C, Williams BA, Pertea G et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010; 28(5): 511–515. doi: 10.1038/nbt.1621.
21. Hansen KD, Wu Z, Irizarry RA et al. Sequencing technology does not eliminate biological variability. Nat Biotechnol 2011; 29(7): 572–573. doi: 10.1038/nbt.1910.
22. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010; 26(1): 139–140. doi: 10.1093/bioinformatics/btp616.
23. Auer PL, Doerge RW. Statistical design and analysis of RNA sequencing data. Genetics 2010; 185(2): 405–416. doi: 10.1534/genetics.110.114983.
24. Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics 2014; 30(3): 301–304. doi: 10.1093/bioinformatics/btt688.