Kľúčové faktory ovplyvňujúce identifikáciu variantov z celoexómového sekvenovania

RNDr. Katarína Skalická, PhD., MPH
Laboratórium klinickej a molekulovej genetiky Detskej kliniky LF UK a NÚDCH, Bratislava

Implementácia metód masívneho paralelného sekvenovania do rutinnej klinickej praxe priniesla obrovský posun v diagnostike mnohých geneticky podmienených ochorení. Rýchlosť akým sa tieto metódy začlenili do praxe je odrazom nevyčísliteľného benefitu, ktorý prinášajú ako pre samotného pacienta, tak aj pre rodinných príslušníkov. Neustále zlepšovanie technológií masívneho paralelného sekvenovania, ktoré sa prispôsobujú novým znalostiam v oblasti genomiky, prináša aj potrebu systematickej inovácie bioinformatických softvérov určených na analýzu dát. Na druhej strane sa však príchodom nových technológií a postupov objavujú nové výzvy, ktoré treba prekonávať takpovediac za behu, aby sa výsledná diagnostická výťažnosť dostala na čo najvyššiu úroveň. Presné zobrazenie variantov v jednotlivých pozíciách ľudského genómu zo získaných sekvenačných dát je kritickým krokom, na ktorý nadväzujú ďalšie analýzy a interpretácie rovnako čeliace mnohých výzvam. Cieľom prehľadového článku je poukázať na kľúčové faktory ovplyvňujúce identifikáciu variantov, ktoré môžu brániť stanoveniu konečnej diagnózy.

Kľúčové slová: referenčný genóm, sekvenačné pokrytie, pseudogény, minoritná frekvencia alel, interpretácia variantov

Key factors influencing variant identification from whole-exome sequencing

The implementation of massively parallel sequencing methods into routine clinical practice has brought a great shift in the diagnosis of many genetically determined diseases. The speed with which these methods were incorporated into practice is the reflection of the incalculable benefit they bring both to the patient and to family members. The continuous improvement of massively parallel sequencing technologies, which adapts to a new knowledge in the field of genomics, brings the need for systematic innovation of bioinformatics software designed for data analysis. On the other hand, however, with the advent of new technologies and procedures, new challenges appear that must be overcome, so to speak, on the fly, so that the resulting diagnostic yield reaches the highest possible level. The accurate display of variants in individual positions of the human genome from the obtained sequencing data is a critical step, which is followed by further analyzes and interpretations, which also face many challenges. The aim of the review article is to point out the key factors influencing the identification of variants that may prevent the establishment of a final diagnosis.

Key words: reference genome, sequencing coverage, pseudogenes, minor allele frequency, variant interpretation

Lek. genet. diagn., 2024;1(2): 103-108

Úvod

Vznik technológií masívneho paralelného sekvenovania (NGS) výrazne ovplyvnil systém poskytovania zdravotnej starostlivosti v lekárskej genetike a diagnostike genetických chorôb. Na rozdiel od predchádzajúcich sekvenačných metód je možné použitím týchto technológií v jedinom experimente vyšetriť nukleotidové sekvencie miliónov až miliárd DNA fragmentov, čo umožňuje zefektívniť čas stanovenia diagnózy najmä v prípade zriedkavých ochorení. Rýchly vývoj technológií viedol k dramatickému zníženiu nákladov na sekvenovanie a následne sprístupnil genómovú analýzu nielen pre oblasť výskumu, ale aj klinickej praxe (1). V lekárskej genetike sa metódy NGS využívajú na identifikáciu mnohých typov genetických variantov, ktoré vedú k vzniku ochorení akými sú jednonukleotidové varianty, krátke inzercie alebo delécie, ako aj zložitejšie typy variantov zahŕňajúce zmeny v počte kópií a komplexné chromozómové prestavby. Bioinformatická analýza predstavuje komplexnú identifikáciu všetkých variantov prítomných vo vyšetrovanej vzorke. Finálnej interpretácii výsledkov NGS predchádza množstvo ďalších analýz zameraných na filtráciu a klasifikáciu identifikovaných variantov s cieľom nájsť jeden alebo niekoľko variantov kauzálne spojených s fenotypom pacienta (2). Miera stanovenia diagnózy použitím metód celogenómového sekvenovania (WGS) dosahuje úroveň od 40 % do 60 % a môže sa líšiť medzi rôznymi skupinami chorôb. Použitím metódy celoexómového sekvenovania, najmä založených na krátkych čítaniach v rozsahu 100 až 150 bázových párov (bp), sa miera diagnostickej úspešnosti ešte znižuje (3). Účinnosť samotnej diagnostiky závisí od viacerých faktorov a neúplnosť vedomostí o patogenetických mechanizmoch, ktoré stoja za vznikom ochorení, je pravdepodobne najdôležitejším z nich. Veľkou výzvou je bioinformatická analýza a interpretácia získaných údajov, pričom určité pretrvávajúce problémy a chyby pri daných činnostiach môžu negatívne ovplyvniť presnosť a rýchlosť diagnostiky. Súčasné štúdie identifikovali množstvo výziev, ktoré môžu brániť efektívnemu stanoveniu kauzálnych variantov a následnému stanoveniu diagnózy. Ich opis je predmetom predkladaného článku s cieľom osvetliť možné príčiny nediagnostikovaných prípadov, u ktorých je genetická etiológia vysoko pravdepodobná.

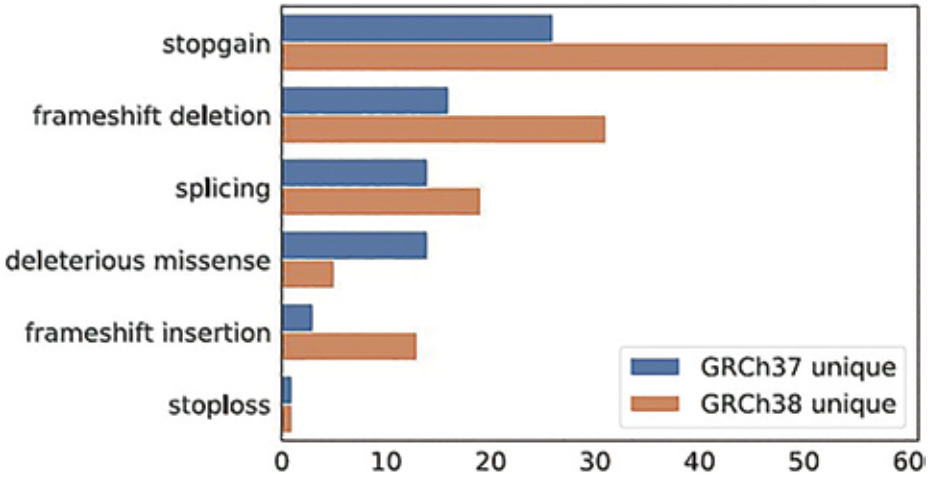
Kvalita referenčného genómu

Prvým krokom analýzy dát získaných z NGS je zrovnanie analyzovanej sekvencie s referenčným genómom.

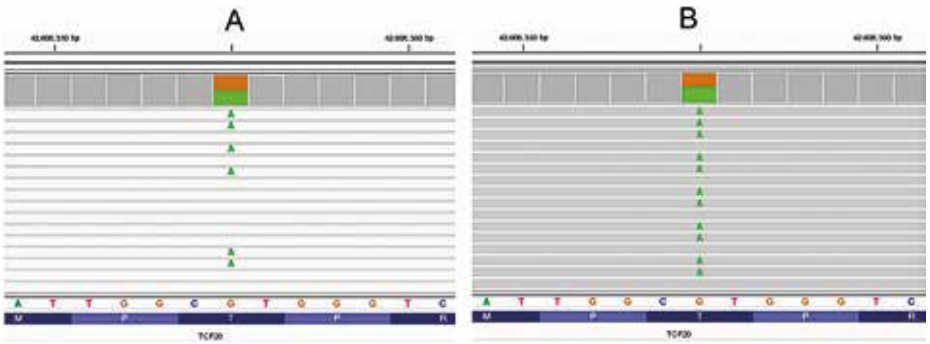
Kvalita sekvencie referenčného genómu je tak nepochybne hlavným faktorom ovplyvňujúcim výsledky sekvenačnej analýzy. V posledných rokoch sa odhalilo niekoľko problémov týkajúcich sa kvality sekvencie referenčného genómu. Bežne používaná sekvencia referenčného genómu pochádza z počiatočného zostavenia definovaného Projektom ľudského genómu (Human Genome Project) z roku 2003, ktorého výsledkom bolo prečítanie 92 percent jeho sekvencií. Predmetom sekvenovania boli euchromatické oblasti jadrového genómu. Avšak zvyšných 8 % genómu rozptýlených v heterochromatických oblastiach centromér a telomér nebolo z dôvodu náročnosti ich sekvenovania a následného zostavenia predmetom pôvodných plánov projektu. V marci 2009 Konzorcium pre referenčný genóm vydalo presnejšiu verziu ľudského genómu (GRCh37, hg19), avšak stále obsahovalo viac ako 300 nesekvenovaných oblastí, tzv. medzier. Najnovšia verzia ľudského genómu, ktorá prišla takmer o 20 rokov neskôr (GRCh38.p12) a v súčasnosti sa využíva pri analýze dát, má stále neprečítaných viac ako 5 % ľudského genómu (>150Mb) (4). Rozdiely v počte identifikovaných variantov pri použití referenčných genómov GRCh37 a GRCh38 stanovených rozsiahlou štúdiou zahŕňajúcou 1 572 vzoriek po celoexómovom sekvenovaní sú zobrazené na obrázku 1 (5).

Ďalším problémom ovplyvňujúcim identifikáciu variantov je skutočnosť, že 70 % výslednej sekvencie z Projektu ľudského genómu pochádzalo od jedného jedinca afroeurópskeho pôvodu. Referenčný genóm tak zostal nedokonalý z dôvodu kritických 0,2 až 1 % sekvencií, ktoré robia každého jedinca zo 7 miliárd ľudí na svete odlišnými. To viedlo k vzniku inherentnej odchýlky v biomedicínskych údajoch, kde napr. mnohé varianty identifikované v neeurópskych populáciách nie sú v referenčnom genóme vôbec zastúpené. Rozdiely v genóme medzi rôznymi jednotlivcami a populáciami sú v súčasnosti eliminované tzv. alternatívnymi kontigmi, ktoré definujú komplexnú variáciu ľudského genómu vrátane HLA lokusov. Alternatívne kontigy sú však natoľko odlišné, že ich nie je možné použiť ako jednu referenčnú sekvenciu. GRCh38 a GRCh37 (hg19), ktoré predstavujú dve najčastejšie používané zostavy ľudského referenčného genómu sa vzájomne odlišujú počtom extrachromozomálnych sekvencií a spomínaných prídatných alternatívnych sekvencií a kontigov. Avšak mnohé štúdie poukázali na to, že použitie referenčnej sekvencie GRCh38 so zaradením alternatívnych kontigov viedlo k absencii tisícok variantov. Dôvodom je nedostatočná hodnota kvality mapovania (MAPQ), ktorá kvantifikuje pravdepodobnosť nesprávne umiestneného čítania, t. j. predstavuje negatívnu logaritmicky škálovanú pravdepodobnosť, že je čítanie nesprávne zarovnané. Tento problém sa dá riešiť jednoduchým vylúčením alternatívnych kontigov z analýzy alebo použiť zrovnávacie algoritmy, ktoré ich spracujú pomocou zodpovedajúceho indexového súboru (6). Príklad vplyvu alternatívnych kontigov v referenčnom genóme na identifikáciu patogénneho variantu je znázornený na obrázku 2 (10).

Obrázok 1. Rozdiely v počte identifikovaných variantov s potenciálne poškodzujúcim účinkom pri použití referenčných genómov GRCh37 a GRCh38 získané rozsiahlou štúdiou zahŕňajúcou dáta z celoexómového sekvenovania 1 572 vzoriek (5)



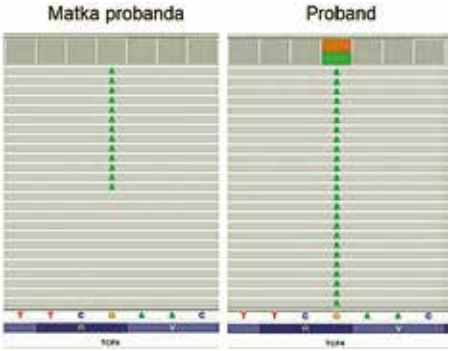
Obrázok 2. Vplyv alternatívnych kontigov v referenčnom genóme na identifikáciu variantu v géne *TCF20*. Snímka obrazovky integratívneho genomického prehliadača (IGV) zobrazujúca zarovnanie sekvencií pre patogénny kódujúci variant v géne *TCF20*, ktorý nebol pôvodne identifikovaný. Dôvodom bolo zahrnutie alternatívnych kontigov, čo viedlo k zarovnaniu sekvenačných čítaní na viacerých miestach referenčného genómu (10).



Prechod z referenčného genómu GRCh37 na GRCh38 v diagnostickej praxi bol značne oneskorený z dôvodu potrebnej aktualizácie bežne používaných databáz genetických variantov ako napr. 1000 genomes, Exome Aggregation Consortium (ExAC) alebo gnomAD, ktoré boli pôvodne skonštruované pomocou hg19 a prechodom na novú referenčnú sekvenciu vykazovali mierne odchýlky.

V roku 2022 vedci predstavili úplne novú kompletnú zostavu ľudského genómu s názvom Telomere-to-Telomere (T2T), ktorá zlepšuje výsledky viacerých analýz vrátane identifikácie variantov. Na rozdiel od predchádzajúcich verzií referenčného genómu, T2T neobsahuje žiadne medzery ani extrachromozomálne sekvencie. Medzinárodnému konzorciu T2T sa tak podarilo skompletizovať chýbajúcich 8 % ľudského genómu. V porovnaní s GRCh38 väčšina pridaných sekvencií zodpovedá segmentálnym duplikáciám a centroméram, čo umožňuje

Obrázok 4. Nesprávne vylúčenie patogénneho variantu typu nonsense v géne *TCF4* z dôvodu nadstavenia filtrácie variantov podľa predpokladaného typu dedičnosti na úrovni de novo. Dôvodom bola prítomnosť patogénneho variantu u matky probanda v mozaike na úrovni 30 % (10).



Obrázok 3. A. Odporúčané hodnoty MAF pre klasifikáciu benígnych variantov v génoch asociovaných s poruchami sluchu (16), **B.** evidované hodnoty MAF ich patogénnych variantov vedúcich k vzniku ochorenia (12)

A

BENIGN CRITERIA	
BA1	MAF of ≥0.005 (0.5%) for autosomal recessive; MAF of ≥0.001 (0.1%) for autosomal dominant
BS1	MAF of ≥0.003 (0.3%) for autosomal recessive; MAF of ≥0.0002 (0.02%) for autosomal dominant. Likely benign, provided there is no conflicting evidence.
BS1_ Supporting	MAF of ≥0.0007 (0.07%) for autosomal recessive. No BS1_Supporting criteria for autosomal dominant.

B

Gene	cDNA	Protein	Pathogenicity	MAF* % (Population)
GJB2	c.-22-2A>C		Uncertain significance	0.4 % (Ashkenazi Jewish)
GJB2	c.34G>T	p.Gly12Cys	Likely pathogenic	0.3 % (Latino)
GJB2	c.35delG	p.Gly12Valfs*2	Pathogenic	0.9 % (Non-Finnish European)
GJB2	c.71G>A	p.Trp24*	Pathogenic	0.4 % (South Asian)
GJB2	c.101T>C	p.Met34Thr	Pathogenic	2.0 % (Finnish)
GJB2	c.109G>A	p.Val37Ile	Pathogenic	8.0 % (East Asian)
GJB2	c.167delT	p.Leu56Argfs*26	Pathogenic	1.6 % (Ashkenazi Jewish)
GJB2	c.235delC	p.Leu79Cysfs*3	Pathogenic	0.6 % (East Asian)
SLC26A4	c.349C>T	p.Leu117Phe	Pathogenic	0.5 % (Ashkenazi Jewish)
SLC26A4	c.919-2A>G	p.?	Pathogenic	0.5 % (East Asian)

*The highest subpopulation frequency in the Genome Aggregation Database (gnomAD) is shown. *GJB2* (NM_004004.5), *SLC26A4* (NM_000441.1).

dôkladnú analýzu týchto oblastí. Okrem toho bol aktualizovaný počet kódujúcich sekvencií o niekoľko stoviek génov kódujúcich proteín. Vzhľadom na tieto vylepšenia možno očakávať, že sa v budúcnosti T2T stane novým štandardným referenčným genómom (4). Predpokladá sa, že prechod na novú verziu bude trvať oveľa dlhšie ako v prípade prechodu hg19 na GRCh38. Nový koncept ľudského genómu navyše detailnejšie zachytáva genetickú diverzitu ľudstva. Nový pangenóm (celý súbor génov viacerých jedincov) zahŕňa DNA 47 jedincov zo všetkých kontinentov sveta okrem Antarktídy a Oceánie. Vedci dúfajú, že sa im do polovice roku 2024 podarí osekvenovať DNA ďalších 350 ľudí, a získať tak nové informácie. Vďaka novému konceptu poznáme teraz 119 miliónov nových bázových párov, ktoré tvoria ľudský genóm, a tým sa prehlbuje nielen naše chápanie ľudskej genetickej diverzity, ale zvyšuje sa šanca na vyriešenie nediagnostikovaných prípadov, ktorých genetická etiológia je vysoko pravdepodobná (7).

Odchýlky v pokrytí cieľovej sekvencie

Ďalším faktorom, ktorý ovplyvňuje identifikáciu variantov z NGS, je pokrytie sekvencie, t. j. počet čítaní zarovnaných s konkrétnou pozíciou v referenčnom genóme. Okrem strednej hĺbky pokrytia (počet čítaní zarovnaných s priemernou genómovou pozíciou) je tiež dôležité zvážiť šírku pokrytia – podiel báz, ktoré sú pokryté dostatočným množstvom čítaní, aby sa umožnila presná analýza variantov. Šírka pokrytia je určená jednotnosťou pokrytia a existujúcimi odchýlkami pokrytia, ktoré sa medzi jednotlivými platformami líšia. Nerovnomernosť pokrytia zostáva významným problémom vo WES. Hĺbková analýza determinantov nízkeho pokrytia vo WES a WGS ukázala, že obmedzenia mapovateľnosti krátkych čítaní sú jedným z najdôležitejším faktorov ovplyvňujúcich pokrytie kódujúcich sekvencií (CDS) v ľudskom genóme. Oblasti s nízkou mapovateľnosťou siahajú od 400 000 do viac ako 1 000 000 bázových párov a problém s mapovateľnosťou je výraznejší pre WES v porovnaní s WGS bez PCR a to vplyvom nižšej veľkosti inzertu. Obmedzenia mapovateľnosti krátkeho čítania ovplyvňujú mnohé dôležité gény napr. SMN1 a SMN2, ktoré sú úplne reprezentované sekvenciami s nízkou mapovateľnosťou, zatiaľ čo pri iných génoch môžu byť tieto sekvencie prítomné len v určitých oblastiach (napr. v génoch NEB alebo TTN spojené s neuromuskulárnou chorobou). Množstvo báz s nízkou mapovateľnosťou skutočne závisí od použitej zostavy referenčného genómu, pričom segmentové duplikácie sú jedným z hlavných zdrojov problémov s mapovateľnosťou. Referenčný genóm T2T konzistentne ukazuje minimálny počet báz pokrytých nejedinečnými mapovaniami (stredný počet je 430 000 bp), zatiaľ čo GRCh37 zostáva najhorším so stredným rozsahom oblastí s nízkou mapovateľnosťou na úrovni 662 000 bp (8).

Existujúce izoformy, pseudogény a kópie génov

Regulácia génu dlhodobo vychádzala z koncepcie jedného promótora riadiaceho jeho transkripciu, po ktorej nasleduje zostrih pre-messengerovej RNA a delécia všetkých intrónov. V súčasnosti vieme, že génová expresia je riadená spôsobom závislým od času, tkaniva alebo vývojového štádia. Gény môžu mať rôzne miesta iniciácie translácie alebo viacero promótorov, ktoré spôsobujú výskyt rôznych izoforiem. Zostrihové izoformy génu môžu vykazovať absenciu jedného alebo viacerých exónov (prirodzený skipping), prípadne môžu mať ďalšie relevantné exóny (9). Problémom je zvážiť, ktorá izoforma je relevantná pre ochorenie alebo v prípade, že sa čítací rámec medzi izoformami líši, ako zabrániť strate anotácie relevantného variantu.

Publikovaným príkladom je variant v géne CACNA1A, ktorý bol identifikovaný u pacienta s epizodickou ataxiou, kde v jednej z piatich izoforiem daného génu predstavoval nonsense variant,

NM_001127221.1:c.5569C>T p.(Arg1857*), zatiaľ čo v ostatných štyroch intrónový variant. Expanzná dráha polyQ zapojená do spinocerebelárnej ataxie typu 6 (OMIM #183086) je kódovaná dvoma ďalšími izoformami CACNA1A (NM_001127222.2 a NM_023035.3), čo naznačuje, že tieto dve izoformy sú nevyhnutné pre správnu funkciu mozočka. Skutočnosť, že nonsense variant je prítomný iba v izoforme, ktorá nekóduje polyQ oblasť, spočiatku viedla k vylúčeniu jeho patogenicity. Avšak vedci ukázali, že táto izoforma používa alternatívny exón 37A namiesto pôvodného exónu 37B a že nonsense varianty v tejto izoforme spôsobujú epizodickú ataxiu (OMIM #108500).

Schopnosť správne identifikovať a klasifikovať varianty do značnej miery závisí aj od schopnosti bioinformatických softvérov automaticky zahrnúť do analýzy aj iné izoformy referenčných génov, pri ktorých sa môže vyskytovať patogénny variant (10).

Varianty prítomné v oblastiach s nízkou výkonnosťou identifikátorov

Presnosť identifikácie variantov môže byť ovplyvnená aj náhodnou alebo systematickou variabilitou. Pre malé varianty typu jednonukleotidových zmien a krátkych delécií (indely) má náhodná variabilita minimálny vplyv na výkonnosť identifikácie variantov vo väčšine genómu, ktorý je sekvenovaný s pokrytím ≥ 30x. Avšak genómové oblasti, ktoré sú systematicky ovplyvňované nižšou kvalitou – ako je zvýšená chybovosť, nízka kvalita mapovania alebo hĺbkové anomálie – nemusia poskytovať konzistentne výsledky pre SNV a indely. Tieto poznatky umožnili klasifikovať v sekvencii genómu oblasti s vysokou a nízkou spoľahlivosťou na detekciu variantov.

Výkonnosť identifikátorov používaných v bioinformatických softvéroch je pomerne vysoká (nad 99 %), avšak treba si uvedomiť, že ich hodnota sa stanovuje analýzou presnosti identifikácie variantov v oblastiach s vysokou spoľahlivosťou. Referenčné štandardy pre stanovenie výkonnosti identifikátorov vyvíja Konzorcium Genome in a Bottle (GIAB), ktorého cieľom je vytvoriť technickú infraštruktúru (referenčné štandardy, referenčné metódy a referenčné údaje) za účelom využitia prekladu sekvenovania celého ľudského genómu do klinickej praxe a overovania vytvorených inovácií v technológiách. Vysoko spoľahlivé regióny sa časom rozširujú vďaka neustálemu úsiliu GIAB zlepšovať súbor údajov pomocou rôznych technológií sekvenovania, akými sú metódy dlhého čítania. Celkovo však oblasti s vysokou spoľahlivosťou zaberajú len 2,37 Gbp (75,2 % sekvencie ľudského genómu) a 30,4 Mbp kódujúcej sekvencie (86,6 %) (11).

Zostávajúce regióny zahŕňajúce až 4,7 Mbp kódujúcich sekvencií sú náročnejšie na identifikáciu variantov. Tieto oblasti zahŕňajú miesta štrukturálnych variácií, segmentálnych duplikácií, ako aj niektoré oblasti s nízkou mapovateľnosťou, akým je napr. lokus hlavného histokompatibilného komplexu (MHC). V rámci kódujúceho genómu sa väčšina takýchto intervalov nachádza na chromozóme X, ktorý nie je zahrnutý do intervalov vysokej spoľahlivosti. Chromozóm X predstavuje 1,3 Mbp z celkovej veľkosti 4,7 Mbp ťažko dostupných kódujúcich oblastí. Identifikácia variantov lokalizovaných na oboch pohlavných chromozómoch sa ukazuje ako veľmi náročná. Tieto informácie sú veľmi dôležité vzhľadom na veľký počet klinicky relevantných génov umiestnených na chromozómoch X a Y (napr. *DMD*, *SRY*). Celkovo až 10 604 známych patogénnych variantov sa nachádza na týchto chromozómoch v ClinVar (v. 2023-03-26), čo je 7,8 % všetkých patogénnych variantov. Tieto čísla naznačujú, že je potrebné venovať väčšiu pozornosť identifikácii variantov a genotypizácii pohlavných chromozómov. Spomedzi autozómov sa najväčšie množstvo (0,4 Mbp) ťažko identifikovaných oblastí nachádza na chromozóme 19. Súbor génov zodpovedajúcich náročným autozomálnym oblastiam sa prekrýva s kódujúcou sekvenciou približne 1 116 génov. Tento zoznam génov je obohatený o gény imunitného systému a hemostázy, ako aj komponenty signálnych dráh (8).

Je dôležité poznamenať, že v rámci jedného exómu existuje viac ako 4 000 takýchto variantov (odhadovaných pomocou súboru údajov HG001 GIAB) a až 17 607 (12,9 %) známych patogénnych variantov spadá mimo oblasti s vysokou spoľahlivosťou. Tieto údaje sú vyššie, ako sa očakávalo, a preto je potrebné k identifikácii variantov lokalizovaných mimo oblastí s vysokou spoľahlivosťou pristupovať individuálne a s väčšou opatrnosťou (8).

Filtrovanie variantov na základe frekvencie výskytu v populačných databázach

Odstránenie variantov, ktoré sa štandardne vyskytujú v populácii zdravých jedincov, je základným krokom filtrovania variantov získaných z WES. Veľkou pomocou sú verejne dostupné databázy, ako napríklad gnomAD, ktoré poskytujú súhrnné informácie o variantoch z veľkých populačných kohort. Štandardizácia prahov minoritnej frekvencie alel (MAF) pre benígnu alebo pravdepodobne benígnu klasifikáciu je nevyhnutná na presnú a konzistentnú interpretáciu variantov. Prahová hodnota musí byť nastavená tak, aby sa bežné patogénne varianty neodfiltrovali, ale nemôže byť taká vysoká, aby boli výsledky genetických vyšetrení preťažené variantmi neistého významu, ktoré pravdepodobne nespôsobujú ochorenie. Bežne používané prahové hodnoty pre takéto filtrovanie eliminujú všetky údaje s frekvenciou alel > 1 % alebo na základe frekvencie a vzorcov dedičnosti ochorenia. Pri použití takéhoto filtrovania frekvencie alel však môžu byť klinicky relevantné varianty nesprávne vyradené (12).

Vysoká frekvencia alel však môže byť chybne stanovená dôsledkom sekvenačného artefaktu, napr. prítomnosťou homopolymérnych úsekov, ktoré sú náchylné na sklz polymerázy, čo môže viesť k inzercii alebo delécii množstva nukleotidov. Tieto varianty môžu byť prítomné v kontrolných databázach ako artefakty, ale môžu byť tiež skutočnými kauzálnymi variantmi v analyzovaných sekvenčných údajoch (10).

Príkladom patogénnych variantov, ktorých hodnota MAF presahuje prahovú hodnotu sú varianty v génoch *GJB2* a *SLC26A* vedúce k strate sluchu. Hodnoty MAF pre recesívne a dominantné varianty v génoch asociovaných so stratou sluchu vychádzajúce z usmernení Americkej vysokej školy lekárskej gene-

tiky a genomiky (ACMG) a Asociácie pre molekulárnu patológiu (AMP) pre klasifikáciu benígnych variantov a skutočná hodnota MAF patogénnych variantov v uvedených génoch sú zobrazené na obrázku 3.

Pri genetickom vyšetrení vrodených porúch sluchu, ako aj komplexných ochorení prejavujúcich sa touto syndromologickou jednotkou, sa odporúča nefiltrovať identifikované varianty na základe MAF. Prípadne, ak to softvér na vyhodnocovanie dát umožňuje, zadefinovať len konkrétne gény, prípadne varianty, ktoré je potrebné vynechať z automatizovanej klasifikácie využívajúcej MAF (12). Rovnako je potrebné byť obozretný a precízny nielen pri vyšetrení génov asociovaných so stratou sluchu, ale pri všetkých analýzach.

Anotácia a interpretácia variantov

Interpretácia výsledkov je pravdepodobne najnáročnejším krokom v analýze dát získaných z NGS. Hlavným cieľom je identifikovať variant alebo niekoľko variantov, ktoré sú alebo môžu byť kauzálne pre vznik ochorenia. Táto úloha je komplikovaná obrovským počtom variantov identifikovaných v jednotlivej vzorke, ktoré pri WGS dosahujú hodnotu tri až štyri milióny variantov, pričom v CDS sa nachádza približne 25 000 variantov vychádzajúcich z WGS aj WES (11). Identifikácia jedinej kauzálnej mutácie v takom veľkom súbore variantov nie je jednoduchá a na pomoc pri tomto procese bolo navrhnutých niekoľko smerníc a odporúčaní na interpretáciu výsledkov NGS v lekárskej genetike, pričom usmernenia ACMG/AMP sa považujú za zlatý štandard. Tieto usmernenia obsahujú pevný počet kritérií, ktoré hodnotia potenciálnu patogenitu variantu na základe rôznych druhov podporných dôkazov a umožňujú ho klasifikovať do jednej z piatich kategórií: patogénny, pravdepodobne patogénny, pravdepodobne benígny, benígny alebo variant nejasného významu (VUS). Webové platformy ako Franklin (https://franklin.genoox.com/klinické-db/home) alebo Varsome boli vyvinuté na automatickú klasifikáciu daného variantu podľa kritérií ACMG, avšak diskrepancie vo finálnej klasifikácii identifikovaného variantu v rámci jedného softvéru u viacerých používateľov, ako aj nezhody medzi oboma softvérmi, sú pomerne bežné. Navrhli sa preto ďalšie rozšírenia smerníc ACMG, aby sa formalizoval postup klasifikácie založený na numerických bodovacích schémach (13). Plne automatizovaná klasifikácia variantov však zostáva náročná.

Interpretáciu variantov komplikujú aj výzvy pri predpovedaní straty funkcie kódujúcich variantov použitím prediktívnych softvérov. Strata funkcie (LoF) je hlavným mechanizmom patogenézy zriedkavých chorôb. Identifikácia genetických variantov kódujúcich proteín s účinkami LoF je preto kľúčová pre diagnostiku zriedkavých chorôb. Tri triedy genetických variantov sa zvyčajne považujú za varianty pLoF: nezmyselné (stop_gained) varianty, inzercie a delécie s posunom čítacích rámcov a kanonické varianty miesta zostrihu. Tieto varianty sa tiež nazývajú varianty skrátenia proteínov (PTV), pretože väčšina takýchto variantov vedie k predčasnému ukončeniu ich syntézy. Napriek očakávaným účinkom PTV na funkciu génu, každý jednotlivec nesie až 100 takýchto variantov, čo naznačuje, že na identifikáciu pravých variantov LoF medzi PTV sú potrebné ďalšie filtre. Bolo vyvinutých niekoľko pravidiel na odfiltrovanie variantov LoF s nízkou spoľahlivosťou a implementovali sa moduly anotátorov variantov, ako napríklad LOFTEE. Avšak preddefinované sady pravidiel nestačia na odfiltrovanie všetkých falošne pozitívnych variantov LoF. Údaje o génovej expresii sa tiež používajú na určenie skutočného potenciálu LoF variantu. Nedávno navrhnutý podiel skóre exprimovaných transkriptov využíva informácie o génovej expresii z genotypovej tkanivovej expresie na identifikáciu a vylúčenie variantov, ktoré ovplyvňujú iba izoformy s nízkou alebo nulovou úrovňou expresie. Boli navrhnuté aj iné prístupy na klasifikáciu variantov pLoF. Jedným z príkladov je algoritmus založený na strojovom učení s názvom MutPred-LoF, ktorý ukázal vysokú účinnosť pri predpovedaní patogenity nezmyselných variantov a variantov posunu čítacieho rámca. Prediktívna sila navrhovaných metód je však stále ďaleko od ideálu (8).

Problém predpovedania efektov LoF je podstatne zložitejší pre varianty typu missense. Na vyriešenie tejto úlohy sa zvyčajne používa niekoľko typov dôkazov, vrátane údajov o evolučnej konzervácii proteínových sekvencií, 3D proteínovej štruktúre a chemických vlastnostiach referenčných a mutantných aminokyselín. Boli navrhnuté desiatky bioinformatických nástrojov na predikciu patogenity missense variantov, akými sú SIFT PolyPhen2, CADD a mnoho iných. Veľký počet nástrojov a vysoká miera nekonzistentnosti však viedli k vzniku algoritmov, ktoré agregujú predpovede z rôznych softvérov (napr. REVEL). Okrem toho boli skonštruované databázy, ktoré uchovávajú vopred vypočítané predpovede, ako napríklad dbNSFP. Stále sa však v rámci ACMG/AMP aktívne vyvíjajú nové algoritmy a stratégie na interpretáciu skóre predikcie patogenity.

Okrem missense a iných variantov s vysokým vplyvom v kódujúcej sekvencii, môžu k strate funkcie proteínu viesť aj iné varianty, ktoré sa zvyčajne považujú za tiché (t. j. neovplyvňujú funkciu génového produktu). Bolo opísaných viacero príkladov neočakávaných funkčných účinkov synonymných variantov, ako aj rôzne mechanizmy ich patogenicity. Medzi najčastejšie patrí zavedenie alebo prerušenie zostrihových miest a negatívne účinky na rýchlosť translácie. Príkladom je synonymný variant NM_001077488.2:c.108C>A [p.Val36Val] v géne *GNAS*, ktorý je lokalizovaný v kryptickom mieste zostrihu, čo vedie k nesprávnemu zostrihu zodpovedajúceho transkriptu a k vzniku pseudohypoparatyreózy. Vývoj výpočtovej techniky prediktorov pre synonymné varianty je komplikovaný, hoci sa v tomto smere vynaložilo veľa úsilia (14).

Opis klinických prejavov ochorenia a predpokladaný typ dedičnosti

Informácie o klinickom priebehu ochorenia či už opisom alebo formou termínov z Human phenotype ontology (HPO) sú kľúčovým nástrojom pre výber vhodných génov zahrnutých do analýzy celoexómového sekvenovania použitím vysoko sofistifikovaných softvérov s umelou inteligenciou. Tento spôsob

analýzy dát je v súčasnosti štandardom aj u nás na Slovensku a vykazuje vysokú mieru úspešnosti. Druhou stranou mince sú však ochorenia s vysoko variabilným fenotypom, ktoré skresľujú typický prejav ochorenia, ako aj ultra zriedkavé ochorenia, ktorých prejav je opísaný na malej skupine jedincov, navyše v ranom detstve, a tak nemusí odrážať celý fenotyp. Rovnako netreba zabúdať aj na množstvo ochorení, ktorých prejav je pravdepodobne ovplyvnený známymi alebo doposiaľ neobjavenými modifikátormi fenotypu. Selekcia vybraných génov na základe HPO termínov tak nemusí viesť k úspešnému stanoveniu diagnózy. V týchto prípadoch je potrebné pokračovať komplexnou analýzou všetkých identifikovaných variantov v rámci celého exómu.

Rovnako stratégia filtrovania variantov na základe predpokladaného typu dedičnosti získaného analýzou rodokmeňa môže naraziť na problémy spojené s variabilnou expresivitou, neúplnou penetranciou alebo s mozaicizmom a tak viesť k vyradeniu kauzálnych variantov nesprávnym nadstavením ich filtrácie (10). Príklad vyradenia patogénneho variantu nadstavením selekcie variantov asociovaných s predpokladaným typom dedičnosti je znázornený na obrázku 4.

Potreba reanalýzy

Pravidelné aktualizácie referenčných súborov údajov akými sú populačné frekvencie, informácie v OMIM, klasifikácie variantov v HGMD alebo ClinVar, ako aj nové aktualizácie z RefSeq a Gencode, ktoré menia známe definície génov pribúdaním nových referenčných transkriptov, môžu mať významný vplyv na anotáciu a interpretáciu variantov. Reanalýzy už existujúcich údajov získaných z celoexómového sekvenovania tak môžu viesť k odhaleniu nových diagnóz. Túto skutočnosť môžeme potvrdiť aj na našom pracovisku, kde reanalýza sekvenačných dát použitím nového referenčného genómu, aktualizovaných údajov a v neposlednom rade použitím nových vysoko výkonných softvérov pre bioinformatickú analýzu viedli k odhaleniu kauzálneho variantu a k stanoveniu diagnózy. Potreba pravidelnej opätovnej analýzy údajov pre nediagnostikované prípady môže mať obrovský význam. Nárast počtu stanovených diagnóz z opätovnej analýzy sa líši, ale zvyčajne je podstatný. Výsledky štúdií poukazujú na mieru nových diagnóz z reanalýz na úrovni 15 % (15). Vzhľadom na tempo nového technologického pokroku môže byť opätovná analýza nediagnostikovaných prípadov ešte efektívnejšia. Dúfame však, že nové technológie sekvenovania a bioinformatické prístupy pomôžu výrazne znížiť počet nediagnostikovaných prípadov po genómovej diagnostike.

Záver

V prehľadovom článku sme opísali kľúčové faktory, ktoré môžu brániť identifikácii variantov a následne stanoveniu diagnózy. Verím, že uvedené informácie prinesú nový pohľad na nediagnostikované prípady vychádzajúce z WES, návrhy a riešenia na prekonanie týchto výziev, prípadne zabránia vzniku opakovaných chýb pri hodnotení dát.

Autorka vyhlasuje, že nie je v nijakom potenciálnom konflikte záujmov.

Literatúra

1. Nurchis MC, Riccardi MT, Radio FC, et al. Incremental net benefit of whole genome sequencing for newborns and children with suspected genetic disorders: Systematic review and meta-analysis of cost-effectiveness evidence. Health Policy. 2022;126(4):337-345.
2. Sezerman OU, Ulgen E, Seymen N, et al. Bioinformatics worklows for genomic variant discovery, interpretation and prioritization. 2019, E-book, ISBN 978-1-78923-800-6, 100 pages, doi 10.5772/intechopen.77443.
3. Radio FC, Ruzzeddu M, Bartuli A, et al. An Italian pilot study on undiagnosed patients. New Genet Soc. 2019;38:249-63.
4. Nurk S, Koren S, Rhie A, et al. The compete sequence of human genome. Science. 2022;376:44-53.
5. Li H, Dawood M, Khayat MM, et al. Exome variant discrepancies due to reference-genome differences. The American Journal of Human Genetics. 2021;108:1239-1250.
6. Jia T, Munson B, Largo A, et al. Thousands of missing variants in the UK Biobank are recoverable by genome realignment. Annals of Human Genetics. 2020;84(3):214-220.
7. Liao WW, Asri M, Ebler J, et al. A draft human pangenome reference. Nature. 2023;617:312-324.
8. Barbitoff YA, Ushakov MO, Lazareva TE, et al. Bioinformatics of germilne variant discovery for rare disease diagnostics: current approaches and remaining challenges. Briefings in Bioinformatics. 2024;25(2):1-15.
9. Bodian DL, Kothiyal P, Hauser NS. Pitfalls of clinical exome and gene panel testing: alternative transcripts. Genetics in Medicine. 2019;21:1240-1245.
10. Corominas J, Smeekens SP, Nelen MR, et al. Clinical exome sequencing – mistakes and caveates. Human Mutation. 2022;43:1041-1055.
11. Barbitoff YA, Abasov R, Tvorogova VE, et al. Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy pf coding sequence variant discovery. BMC Genomic. 2022;23(1):1-17.
12. DiStefano MT, Hughes MY, Patel MJ, e al. Expert interpretation of genes and variants in hereditary hearing loss. Medizinische genetik. 2020;32(2):109-115.
13. Nykamp K, Anderson M, Powers M, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. Genet Med. 2017;19(10):1105-17.
14. Apetrei A, Molin A, Gruchy N, et al. A novel synonymous variant in exon 1 of GNAS gene results in a cryptic splice site and cause pseudohypoparathyroidism type 1A nad pseudohypoparathyroidism in a French family. Bone reports. 2021;14:101073.
15. Tan NB, Stapleton R, Stark Z, et al. Evaluation systematic reanalysis of clinical genomic data in rare diseases from single center experience and literature review. Mol Genet Genomic Med. 2020;8(11):1-19.
16. Oza AM, DiStefano MT, Hemphill SE, et al. Expert specification of the ACMG/AMP variant interpretation guidelines for genetic hearing loss. Human Mutat. 2018;39:1593-1613.

RNDr. Katarína Skalická, PhD., MPH
Laboratórium klinickej
a molekulovej genetiky
Detská klinika LF UK a NÚDCH
Limbová 1, 833 40 Bratislava
katarina.skalicka@nudch.eu