Uporaba sekvenciranja naslednje generacije v klinični diagnostiki prirojenih bolezni

Next-generation sequencing in the clinical diagnosis of congenital diseases

Uroš Prešern[1], Jernej Kovač[1,2], Robert Šket[1,2], Tine Tesovnik[1,2], Maruša Debeljak[1,2], Barbara Jenko Bizjan[1,2]
[1]Univerza v Ljubljani, Medicinska fakulteta
[2]Univerzitetni klinični center Ljubljana, Pediatrična klinika, Klinični inštitut za specialno laboratorijsko diagnostiko

Avtor za korespondenco:
Asist. dr. Barbara Jenko Bizjan
Univerzitetni klinični center Ljubljana, Pediatrična klinika, Klinični inštitut za specialno laboratorijsko diagnostiko, Vrazov trg 1, 1000 Ljubljana
e-pošta: barbara.jenko.bizjan@kclj.si

POVZETEK

Sekvenciranje naslednje generacije (NGS) je v zadnjem desetletju postalo osrednja metoda v klinični diagnostiki prirojenih bolezni. Njegova visoka zmogljivost omogoča iskanje vzročnih patogenih sprememb skozi celoten genom, pri čemer pa sam postopek zaradi velike količine generiranih podatkov zahteva uporabo posebnih bioinformatskih orodij. Eden ključnih korakov v analizi je pravilna poravnava pridobljenih zaporedij na zaporedje referenčnega genoma, čemur sledi določitev prisotnih genetskih sprememb. Ozko grlo v analiznem postopku trenutno predstavlja anotacija sprememb, kjer je za odkrite spremembe v genomu treba določiti, če so patogene ali benigne, ter če lahko njihova prisotnost v genomu pojasni klinično sliko preiskovanca. Svojevrstno etično dilemo predstavlja tudi način ravnanja ob odkritju sprememb, ki so sicer patogene, vendar niso neposredno povezane s klinično sliko. V tem preglednem članku je opisan potek uporabe NGS v klinični diagnostiki prirojenih bolezni, ki je dodatno predstavljen na primeru bolnika s prirojeno katarakto.

Ključne besede: sekvenciranje naslednje generacije, diagnostika prirojenih bolezni

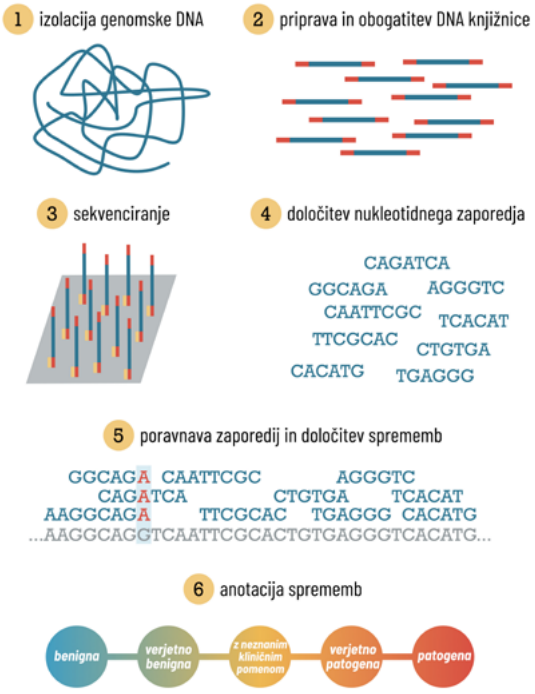
ABSTRACT

Over the past decade, next-generation sequencing (NGS) has become the method of choice in the clinical diagnosis of congenital diseases. Its high throughput enables the identification of causal pathogenic variants throughout the entire genome; however, the process requires specialized bioinformatics tools due to the large amount of generated data. One of the key steps in the analysis is a correct alignment of acquired sequences to the reference genome, followed by variant calling. Currently, the bottleneck in the process is variant annotation, where discovered variants in the genome need to be classified as pathogenic or benign. The causation between detected variants and observed phenotype is also determined. A unique ethical dilemma is encountered when a pathogenic variant not directly linked to an observed phenotype is discovered. This review describes the individual steps of NGS in the clinical diagnosis of congenital diseases. The use of NGS in the case of a patient with a congenital cataract is presented as an example.

Key words: next generation sequencing, diagnosis of congenital diseases

UVOD

Razvoj sekvenciranja naslednje generacije (angl. *next generation sequencing*, NGS) je povzročil revolucijo na področju določanja nukleotidnih zaporedij, saj je le-to postalo dosti hitrejše in cenejše. Zaradi zmožnosti sočasne analize celotnega genoma se je NGS hitro uveljavil v klinični diagnostiki, ki je bila pred tem omejena na določanje zaporedja posameznih genov. V genetski diagnostiki se NGS najbolj pogosto uporablja v diagnostiki prirojenih bolezni in onkologiji. V zadnjem času pa prodira tudi na druga področja, kot sta mikrobna diagnostika (1) in določanje tkivne skladnosti (2). V nadaljevanju se bomo osredotočili na diagnostiko prirojenih bolezni, kjer bodo predstavljeni potek analize ter zmogljivosti in omejitve NGS (Slika 1).



Slika 1: Shema poteka uporabe NGS v klinični diagnostiki. Posamezni koraki, ki so prikazani na shemi, so podrobneje razloženi v nadaljevanju besedila.

Figure 1: A general outline of the use of NGS in clinical diagnostics. The individual steps mentioned in the figure are explained in more detail in the following parts of the review.

Pojem NGS označuje različne tehnologije sekvenciranja, za katere je značilno, da omogočajo analizo ogromnega števila zaporedij hkrati. Izmed prvotnih tehnologij NGS je danes v širši rabi tehnologija sekvenciranja s sintezo podjetja Illumina (3), zaradi česar izraz NGS pogosto kar enačimo z njo (to velja tudi za nadaljevanje članka, razen če ni navedeno drugače). NGS temelji na sekvenciranju kratkih fragmentov DNA (300 baznih parov), ki so pripeti na pretočno celico, kar omogoča njihovo prostorsko ločbo in sočasno analizo. Sekvenciranje poteka preko sinteze komplementarne verige DNA, kjer se nanjo postopoma dodajajo nukleotidi s fluorescenčno oznako. Zajemanje svetlobnega signala na koncu vsakega dodanega nukleotida omogoča določitev nukleotida, ki se je v tistem krogu vgradil v verigo. V kategorijo NGS uvrščamo tudi novejšo, tretjo generacijo sekvenciranja, katere pomemben predstavnik je sekvenciranje z nanoporami, ki prav tako počasi postaja sestavni del klinične diagnostike. Tretja generacija sekven- »

ciranja omogoča veliko daljša branja (> 10 kb) kot druga (kamor spada Illumina) in služi kot komplementarna metoda v primerih, pri katerih je zmogljivost sekvenciranja s sekvenatorji Illumina slabša, npr. ponavljajoča zaporedja in psevdogeni (4).

Cilj uporabe NGS v diagnostiki prirojenih bolezni je odkritje ene ali več genetskih sprememb, ki bi nam pojasnile klinično sliko preiskovanca. Pod pojmom genetska sprememba razumemo kakršnokoli odstopanje od nukleotidnega zaporedja referenčnega človeškega genoma. Pri tem gre lahko za manjše genetske spremembe, kot v primeru sprememb posameznih nukleotidov (angl. *single nucleotide variant*, SNV) in kratkih insercij/delecij, ali večje strukturne spremembe, kot so duplikacije, translokacije, inverzije ter večje insercije/delecije. Večina genetskih sprememb je benigne narave, nekatere pa lahko vodijo do nastanka patoloških stanj. Pri iskanju vzročnih patogenih genetskih sprememb ločimo tri različne pristope glede na območje, ki ga v genomu analiziramo: analiza panela genov, sekvenciranje celotnega eksoma (angl. *whole exome sequencing*, WES) in sekvenciranje celotnega genoma (angl. *whole genome sequencing*, WGS). Vsak izmed pristopov ima svoje prednosti in slabosti, v praksi se trenutno najpogosteje uporablja sekvenciranje celotnega eksoma (5).

Analiza panela genov

Pri analizi panela genov pripravimo NGS knjižnico za zgolj vnaprej izbrane gene, ki so povezani z opazovanimi kliničnimi znaki (npr. prirojena izguba sluha) (6). Z omejitvijo na manjše število preiskovanih genov se zmanjšajo stroški sekvenciranja, prav tako je enostavnejša tudi nadaljnja bioinformatska analiza rezultatov. Slabost tega pristopa je možnost, da se vzročni gen ne nahaja v izbranem panelu in ga tako v analizi ne bomo zajeli. Prav tako lahko težavo povzročajo neklasične kombinacije kliničnih znakov, kjer se pojavi dilema, panel katerih genov je smiselno analizirati.

Sekvenciranje celotnega eksoma

Eksom obsega vse nukleotidne regije na DNA, ki kodirajo proteine v genomu, in predstavlja med 1 in 2 % celotnega genoma. Pri WES se priprava NGS knjižnice iz panela izbranih genov razširi na vse eksonske regije v genomu. S tem povečamo verjetnost, da bomo v analizo zajeli patogeno genetsko spremembo, hkrati pa je zaradi večje velikosti preiskovanega območja analiza dražja in zahtevnejša (7).

Sekvenciranje celotnega genoma

WGS dodatno poveča možnost za najdbo vzročne patogene genetske spremembe, hkrati pa se pri WGS izognemo stopnji obogatitve DNA knjižnice, zaradi česar je globina branja skozi celoten genom enakomernejša (8). WGS omogoča tudi lažjo določitev strukturnih genomskih sprememb, opredelitev mitohondrijskega genotipa na mitohondrijskem genomu, opredelitev večjih sprememb v številu kopij odseka genoma in njihovih lomov. Njegova ključna slaba lastnost je dražja in časovno zamudnejša analiza.

POTEK SEKVENCIRANJA S TEHNOLOGIJO NGS

Izolacija DNA in priprava DNA knjižnice

Priprava DNA knjižnice se nanaša na postopek priprave vzorca DNA za izvedbo sekvenciranja. Začetni material je izolirana DNA, pri čemer se za izolacijo najpogosteje uporabljata periferna kri in slina. Ker tehnologija NGS temelji na sekvenciranju kratkih fragmentov DNA, je treba narediti fragmentacijo izolirane DNA, in sicer s fizikalnimi (npr. soniciranje) ali encimskimi metodami, medtem ko se kemične metode pogosteje uporabljajo za fragmentacijo RNA (9). Po fragmentaciji naredimo selekcijo fragmentov dolžine 300–500 bp z metodo čiščenja na paramagnetnih kroglicah. Paramagnetne kroglice glede na ionsko moč selektivno vežejo nukleinske kisline po velikosti ter se tako uporabljajo za visoko učinkovite protokole izolacije in čiščenja fragmentov DNA. Po selekciji velikosti fragmentov se z ligacijo ali verižno reakcijo s polimerazo (PCR) na obeh koncih posameznega fragmenta DNA pritrdita adapterska oligonukleotida s tako imenovano »molekularno črtno kodo«. Ta je sestavljena iz unikatnega nukleotidnega zaporedja, ki v primeru vzporednega sekvenciranja več vzorcev hkrati služi kot identifikacijska regija za določitev, kateremu vzorcu pripadajo prebrana zaporedja DNA. Poleg molekularne črtne kode vsebuje adapter tudi regijo, ki omogoča pritrditev fragmenta DNA na pretočno celico. Pojem DNA knjižnica se nanaša na skupek vseh fragmentov DNA posameznega vzorca s pripetimi adapterji. Po pripravi DNA knjižnice za posameznega preiskovanca je treba ovrednotiti količino, kakovost in dolžino pripravljenih fragmentov DNA. To običajno storimo z avtomatsko elektroforezo. »

Obogatitev DNA knjižnice

Pri analizi izbranega panela genov in WES naredimo tudi obogatitev DNA knjižnice, s katero v njej povečamo delež fragmentov, ki pripadajo preučevanim regijam v genomu. To dosežemo bodisi s pomnoževanjem izbranih fragmentov DNA s PCR ali z njihovo izolacijo preko hibridizacije s komplementarnim zaporedjem (10). V primeru izvedbe WES je priporočljivo, da obogatena DNA knjižnica poleg eksonskih regij vsebuje tudi začetni del nukleotidnega zaporedja sosednjih intronov, saj se tu pogosto nahajajo genetske spremembe, ki spremenijo proces izrezovanja (11). Pomanjkljivost obogatitve DNA knjižnice je neenakomerna pomnožitev oziroma izolacija posameznih fragmentov DNA, zaradi česar tudi globina branja na različnih mestih ni enakomerna. Obogatitev pri WGS ni potrebna, saj določamo zaporedje celotnega genoma.

Sekvenciranje s tehnologijo Illumina

Fragmenti DNA ene ali več pripravljenih DNA knjižnic se s pomnoževanjem preko mostov pritrdijo na pretočno celico, kjer nato poteka sekvenciranje. Ker je dobljena dolžina zaporedja (100–300 bp) običajno krajša od dolžine fragmenta, določimo zgolj zaporedje njegovega 5‘ končnega dela. Sekvenciranje lahko nato ponovimo tudi na drugem koncu komplementarne verige fragmenta DNA, s čimer dobimo zaporedji parnih koncev (angl. *paired-end read*), ki sta ločeni z vmesnim delom, kjer zaporedje ni določeno (3).

BIOINFORMATSKA ANALIZA

Surovi podatki sekvenciranja zahtevajo nadaljnjo analizo z različnimi bioinformatskimi orodji. Proces lahko razdelimo v štiri stopnje: določitev baz, poravnava zaporedij, določitev genetskih sprememb in njihova anotacija.

Določitev baz

Določitev baz zaobjema pretvorbo signalov (npr. svetlobnih ali električnih), zajetih med sekvenciranjem v nukleotidno zaporedje. V primeru sočasne analize več vzorcev je hkrati potrebno tudi razvrstiti, katera zaporedja pripadajo kateremu vzorcu. Pri tem so v pomoč »molekularne črtne kode,« ki se nahajajo na začetku vsakega zaporedja (12). Razvrščena zaporedja so shranjena v obliki datoteke formata *FASTQ*, ki poleg nukleotidnih zaporedij vsebuje tudi podatke o zanesljivosti določitve posameznih nukleotidov.

Poravnava zaporedij

Zaporedja posameznih fragmentov DNA služijo kot osnova za pridobitev celotnega zaporedja eksoma, genoma ali panela genov. Sestavljanje zaporedja genoma *de novo* (tj. brez uporabe referenčnega genoma) je računsko zahtevno in zahteva večjo količino generiranih podatkov (13), zaradi česar se običajno za sestavo genomskega zaporedja uporablja prileganje fragmentov DNA na referenčni genom. V ta namen so bila razvita različna orodja za prileganje, ki vključujejo kombinacijo globalne in lokalne poravnave (14). Rezultati poravnave so shranjeni v datoteki formata *BAM* (angl. *binary alignment map*), ki za vsak posamezen fragment DNA vsebuje podatke o mestu poravnave in morebitnih neujemanjih (15).

Določitev genetskih sprememb

Na podlagi poravnanih zaporedij in neujemanj z referenčnim genomom določimo prisotnost genetskih sprememb. Te so zbrane v datoteki formata *VCF* (angl. *variant call format*). V njej sta med drugim za vsako genetsko spremembo določena njeno nahajališče v genomu ter podatek, za kakšno spremembo v zaporedju gre. Podan je tudi podatek o vertikalni pokritosti oziroma globini branja, ki je opredeljena kot število unikatnih fragmentov DNA, ki vsebujejo dani nukleotid na nekem mestu v genomu (16). Večja kot je vertikalna pokritost genetske spremembe, večja je verjetnost, da je ta v genomu dejansko prisotna in ni zgolj artefakt. Za določitev prisotnosti heterozigotnih genetskih sprememb je okvirno potrebna vsaj desetkratna vertikalna pokritost posameznega nukleotida (17). V praksi se izkaže, da je vertikalna pokritost nukleotidov skozi celoten genom oziroma eksom neenakomerna, zaradi tega mora biti povprečna pokritost genoma oziroma eksoma (tj. povprečna vrednost vertikalnih pokritosti posameznih nukleotidov v celotnem genomu oziroma eksomu) višja. S tem dosežemo zadostno globino branj tudi v regijah z nižjo vertikalno pokritostjo nukleotidov. Pri WGS, kjer je vertikalna pokritost razmeroma enakomerna, je priporočljivo, da je povprečna pokritost genoma vsaj 30-kratna, pri WES, kjer pokritost močneje variira, pa vsaj 75–100kratna (11). »

Če želimo zaznati nizke stopnje mozaicizma ali heteroplazmije, mora biti povprečna pokritost genoma ustrezno višja. Poleg podatka o vertikalni pokritosti je informativen tudi podatek o alelnem deležu genetske spremembe, ki je opredeljen kot razmerje med vertikalno pokritostjo določene genetske spremembe in celokupno vertikalno pokritostjo mesta, kjer se ta nahaja. Iz alelnega deleža je možno sklepati, ali je neka sprememba homozigotna, heterozigotna ali je prisoten mozaicizem (18).

Anotacija genetskih sprememb

Postopek anotacije omogoča povezavo genetske spremembe z informacijami o njenem vplivu na delovanje organizma. Ameriško združenje medicinskih genetikov (angl. *American College of Medical Genetics and Genomics*, ACMG) je izdalo smernice, v katerih priporočajo petstopenjsko razvrščanje sprememb glede na njihov klinični pomen: patogena, verjetno patogena, sprememba z neznanim kliničnim pomenom, verjetno benigna in benigna (19). V kategoriji patogenih (povzročajo nastanek določene bolezni) in benignih (ne povzročajo nastanka bolezni) spadajo tiste genetske spremembe, za katere obstaja trdna znanstvena podlaga o njihovem kliničnem pomenu, medtem ko v kategoriji verjetno patogenih oziroma verjetno benignih spadajo tiste genetske spremembe, pri katerih lahko z vsaj 90odstotno gotovostjo trdimo o njihovem kliničnem pomenu. Če ni jasnih povezav s klinično sliko, se genetska sprememba razvršča kot sprememba z neznanim kliničnim pomenom. Ker je gotovost o povezavi težko kvantitativno oceniti, med posameznimi kategorijami ni ostre meje. Prav tako lahko nova spoznanja privedejo do ponovne razvrstitve posameznih genetskih sprememb. Najpogosteje uporabljeni podatki za določitev kategorije vključujejo pojavnost genetske spremembe v splošni populaciji in populaciji obolelih, funkcionalne študije, vrsto genetske spremembe in predviden vpliv, primerjavo z že znanimi genetskimi spremembami in rezultate računalniških modelov (20). Ti podatki se črpajo iz objavljene znanstvene literature, podatkovnih baz ter rezultatov napovednih algoritmov. Populacijske podatkovne baze (npr. dbSNP, dbVAR in gnomAD) vsebujejo podatke o prisotnosti in frekvenci genetskih sprememb v neki populaciji, vendar ni nujno, da vsebujejo podatke o njihovem kliničnem pomenu (21–24). Kljub temu lahko na podlagi podatka o frekvenci genetske spremembe o njeni funkciji deloma sklepamo, saj običajno velja, da splošno prisotne genetske spremembe niso patogene. Drugi tip podatkovnih baz obsega baze patogenih genetskih sprememb (ClinVar, OMIM, HGMD itd.) in tako vsebuje tudi podatke o njihovem kliničnem pomenu (25–27). Pri interpretaciji podatkov iz podatkovnih baz je potrebna pozornost, v kakšni meri so informacije posodobljene in podprte z zanesljivimi viri. Rezultate podatkovnih baz lahko kombiniramo z algoritmi, ki skušajo napovedati vpliv neke genetske spremembe na izražanje genov ter na funkcijo in strukturo izraženih proteinov. Najpogosteje so uporabljeni algoritmi za napoved posledic na aminokislinski ravni (npr. PolyPhen-2, SIFT in CADD), kjer algoritem preveri, če se prisotnost genetske spremembe izrazi kot nesmiselna oziroma drugače smiselna aminokislinska sprememba ali če pride do spremembe bralnega okvirja (28,29). Algoritem oceni vpliv aminokislinske spremembe na podlagi njenega mesta znotraj proteina ter biokemijskih lastnosti in evolucijske ohranjenosti mutiranega aminokislinskega ostanka. Prav tako se uporabljajo algoritmi za napovedovanje mest izrezovanja intronov (npr. GeneSplicer), ki lahko napovejo pojav ali izgubo mest v prisotnosti neke genetske spremembe (30). Novejša orodja poskušajo napovedati tudi vpliv genetskih sprememb v nekodirajočih regijah (31). Občutljivost in specifičnost napovednih algoritmov se razlikujeta od primera do primera, v splošnem pa ti algoritmi še niso dovolj zanesljivi, da bi jih lahko uporabili samostojno kot edini vir informacije za napoved vpliva genetskih sprememb. Glede na velik napredek napovednih algoritmov na sorodnih področjih, kot je AlphaFold za napovedovanje tridimenzionalnih struktur proteinov (32), lahko pričakujemo, da bo v bližnji prihodnosti uporaba strojnega učenja in nevronskih mrež tudi na področju napovedovanja vpliva genetskih sprememb privedla do znatnih izboljšav napovedne moči (33).

Filtriranje genetskih sprememb

Za opredelitev vzročne genetske spremembe oziroma genetskih sprememb, ki vplivajo na klinično sliko preiskovanca, izberemo panel genov, ki so povezani s kliničnimi znaki. Pri osnovanju panela si pomagamo z bazo podatkov Human Phenotype Ontology (https://hpo.jax.org/app/) in Genomics England PanelApp (https://panelapp.genomicsengland.co.uk/). Obe bazi podatkov sta razviti na podlagi medicinske literature ter sorodnih portalov, kot so Orphanet, DECIPHER in OMIM. Izbrani panel genov uporabimo za filtriranje anotiranih genetskih sprememb. V naslednjem koraku genetske spremembe filtriramo še glede na prisotnost v populaciji ter njihov položaj v intronski, eksonski, promotorski ali regulatorni regiji. Osredotočimo se na redke genetske spremembe, ki so prisotne v eksonski, promotorski ali regulatorni regiji, saj je vpliv intronskih sprememb običajno težje razložljiv. Izjema so že znane opredeljene vzročne »

intronske spremembe. Pri genih, ki se dedujejo avtosomno dominantno, klinično sliko pojasnimo z opredelitvijo ene heterozigotne patološke ali verjetno patološke spremembe. Medtem ko v genih, ki se dedujejo recesivno, za pojasnitev kliničnega fenotipa opredelimo eno patološko ali verjetno patološko spremembo v homozigotni obliki oziroma dve heterozigotni spremembi, ki se nahajata na različnih alelih.

Poročanje naključnih najdb

Pri iskanju vzročnih patogenih sprememb, ki bi pojasnile klinično sliko, lahko bodisi po naključju ali načrtno odkrijemo tudi druge patogene spremembe, ki pa niso povezane z napotno diagnozo. Takim najdbam pravimo naključne oziroma sekundarne (v primeru načrtnega iskanja) in lahko razkrijejo že obstoječo nediagnosticirano bolezen ali pa povečano verjetnost za njen pojav v prihodnosti. Pri tem se pojavlja dvom, kako v takih primerih ravnati ter kdaj o najdbah preiskovanca obvestiti in kdaj ne. V ta namen je združenje ACMG sestavilo seznam genov, za katere je priporočljivo poročanje patogenih sprememb (34). Seznam ne vsebuje vseh sprememb genov, pri katerih so patogene spremembe povezane z razvojem različnih bolezni, temveč se pri uvrščanju genov na seznam ravnajo po več merilih. Glavno izmed njih je, da mora poročanje o najdbi omogočiti ukrepe, ki bi znatno zmanjšali obolevnost ali smrtnost brez prekomerne obremenitve preiskovanca in zdravstvenega sistema. Trenutni seznam vsebuje 81 genov, od katerih je največ povezanih z razvojem rakavih in kardiovaskularnih obolenj.

POMANJKLJIVOSTI NGS

Kljub temu da se je NGS uveljavila kot glavna metoda v klinični genetski diagnostiki, ima nekatere pomanjkljivosti, ki zahtevajo uporabo dodatnih metod ter pozornost pri interpretaciji rezultatov.

Določanje zaporedja problematičnih regij

Problematično je predvsem določanje prisotnosti genetskih sprememb v homolognih, repetitivnih ali z gvaninom in citozinom bogatih območjih. Homologne regije zaradi svoje medsebojne podobnosti predstavljajo oviro pri poravnavi zaporedij, saj je včasih nemogoče nedvoumno določiti, kateri regiji pripada neko zaporedje. Tako lahko v primeru napačnega prileganja pride do pojava lažno pozitivnih ali lažno negativnih rezultatov (35). Podobna težava nastopi pri repetitivnih regijah, kjer je težko določiti pravo mesto poravnave, če zaporedje fragmenta DNA vsebuje zgolj repetitivni element (36). Hkrati so repetitivne in homopolimerne regije močneje izpostavljene napakam pri sekvenciranju in pomnoževanju s PCR, saj lahko pride do zdrsa polimeraze (37). Slabša kakovost rezultatov je tudi v regijah, bogatih z gvaninom in citozinom, kjer pogosteje pride do napačne določitve baz (38).

Določanje strukturnih sprememb in sprememb v številu kopij

Medtem ko se NGS razmeroma dobro odreže pri določanju SNV in kratkih insercij oz. delecij, je določevanje večjih kompleksnih strukturnih sprememb in sprememb v številu kopij (angl. *copy number variant*, CNV) zahtevnejše (39). Razlog je predvsem kratka dolžina prebranih fragmentov DNA, ki jih je v primeru večjih sprememb težko pravilno prilegati na genom. Opredelitev takih genetskih sprememb lahko olajšajo uporaba podatkov o globini branja prebranih zaporedij, sekvenciranje parnih koncev fragmentov DNA ali pa poravnava zaporedij *de novo* brez uporabe referenčnega genoma (40, 41).

Uspešnost določanja genetskih sprememb v zahtevnejših regijah lahko izboljšamo s sočasno uporabo sorodnih metod. V nekaterih primerih lahko zaporedje problematičnih regij uspešno določimo s sekvenciranjem po Sangerju (42). Prav tako lahko posamezno regijo pred sekvenciranjem specifično pomnožimo s PCR, s čimer povečamo globino branja in se izognemo napačnemu prileganju na homologne regije (35). V zadnjem času se vse pogosteje uporabljajo tehnike tretje generacije sekvenciranja, ki s svojo dolgo dolžino branja olajšajo določanje strukturnih sprememb in zaporedij v problematičnih regijah (4).

KLINIČNI PRIMER

Kot oris uporabe NGS sekvenciranja v klinični diagnostiki prirojenih bolezni je naveden primer bolnika s prirojeno katarakto. Ker je bil bolnik na intenzivni enoti, je bilo naročeno nujno genomsko sekvenciranje. Iz krvi smo izolirali DNA in pripravili DNA knjižnico celotnega genoma. DNA smo z metodo sonikacije fragmentirali ter na oba konca fragmentirane DNA ligirali adapterske oligonukleotide s tako imeno- »

vano »molekularno črtno kodo«. Po pripravi DNA knjižnice smo izvedli sekvenciranje s sekvenatorjem Illumina NovaSeq6000. Po končanem sekvenciranju smo naredili bioinformatsko analizo, pri kateri smo s pretvorbo svetlobnih signalov v zaporedje baz določili nukleotidno zaporedje vseh fragmentov DNA ter na podlagi »molekularne črtne kode« določili branja, ki pripadajo preiskovanemu bolniku. S specifičnimi bioinformatskimi orodji smo prilegali izbrana branja na humani referenčni genom. Sledila je določitev genetskih sprememb, ki se razlikujejo od referenčnega genoma ter anotacija le-teh. Pri analizi genetskih sprememb smo pregledali izbrani panel genov, povezanih z izraženim kliničnimi znaki. Po izločitvi intergenskih in intronskih sprememb smo v genu *ATAD3A*, ki kodira mitohondrijsko membransko ATPazo, našli tri heterozigotne spremembe, ki jih v splošni populaciji ni ali pa so zelo redke in prisotne s frekvenco manj kot 0,1 % (gnomAD). Sprememba c.57C>G (NM_001170535) se nahaja 57 nukleotidov pred start kodonom gena *ATAD3A* in lahko potencialno vpliva na učinkovitost promotorja ter tako na izražanje gena, vendar njena biološka funkcija ni znana. Sprememba c.229C>G (rs138594222) povzroči zamenjavo levcinskega ostanka na mestu 77 z valinskim (p.Leu77Val) in je v podatkovni bazi HGMD opisana kot patološka, saj povzroči delno izgubo funkcije proteina ATAD3A (43). Sprememba c.282+484_444+227del predstavlja 1,4 kb veliko delecijo, ki zajema eksona 3 in 4 gena *ATAD3A* (Slika 2). Sprememba je opisana v bazi ClinVar kot patogena delecija. Z analizo sekvenciranja genoma smo opredelili dve patološki spremembi, ki predstavljata verjeten vzrok za opažene klinične znake. Za natančnejšo opredelitev povezave opisanih sprememb z izraženimi kliničnimi znaki bolnika smo opravili tudi analizo družinske segregacije opisanih genetskih sprememb in potrdili, da se patološki spremembi nahajata na različnih alelih.

Slika 2: Slika spremembe c.229C>G (NM_001170535) in 1,4 kb delecije c.282+484_444+227del (NM_001170535) iz interaktivnega genomskega pregledovalnika IGV (angl. *Integrative Genomics Viewer*). Prikazan je izbrani odsek na humanem genomu, kjer se nahajata obe omenjeni genetski spremembi. Zgornje sivo področje prikazuje vertikalno pokritost, spodnje sive črte pa prikazujejo posamezne fragmente DNA. Rdeče črte predstavljajo fragmente DNA, kjer se nahaja delecija c.282+484_444+227del (NM_001170535). Rdeče-modra navpična črta na zgornjem sivem področju predstavlja nukleotidno spremembo c.229C>G (NM_001170535).

Figure 2: Figure of variant c.229C>G (NM_001170535) and 1.4 kb deletion c.282+484_444+227del (NM_001170535) from IGV (Integrative Genomics Viewer). A selected section of the human genome is shown, where both changes are located. The upper grey area shows vertical coverage, while the lower grey lines show individual DNA fragments. The red lines represent the fragments where the deletion c.282+484_444+227del (NM_001170535) is located. The red-blue vertical line in the upper grey area represents the nucleotide change c.229C>G (NM_001170535).

»

ZAKLJUČEK

Prihod NGS je v diagnostiki prirojenih bolezni omogočil premik od zamudnega ciljnega iskanja vzročnega gena do hitre analize celotnega genoma, zaradi česar se je uspešnost iskanja genetskega vzroka bolezni močno izboljšala. Poleg hitrejše in celovitejše analize je NGS močno pospešil tudi pridobivanje informacij o človeškem genomu, zaradi česar poznamo čedalje več genetskih sprememb ter njihov vpliv na razvoj bolezni. Skupaj z razvojem sekvenciranja so se razvila tudi številna bioinformatska orodja, ki omogočajo obdelavo ogromne količine podatkov, ki jih pridobimo z NGS. Kljub uspešnosti sekvenciranja s tehnologijo Illumina pri določanju krajših sprememb, se ta slabše obnese pri daljših strukturnih spremembah. V teh primerih se kot rešitev kaže prihod tretje generacije sekvenciranja z dolgimi branji. V bližnji prihodnosti pa nas verjetno čaka še naslednji preboj, ki ga bo prinesla vpeljava umetne inteligence pri napovedovanju posledic genetskih sprememb na izražanje in delovanje proteinov, s čimer bo močno olajšana anotacija sprememb, ki je pogosto ozko grlo analiz.

LITERATURA

1. Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, García-Cobos S, et al. Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol. 2017;243:16–24.
2. Weimer ET, Montgomery M, Petraroia R, Crawford J, Schmitz JL. Performance characteristics and validation of next-generation sequencing for human leucocyte antigen typing. J Mol Diagn. 2016;18(5):668–75.
3. Modi A, Vai S, Caramelli D, Lari M. The Illumina sequencing protocol and the NovaSeq 6000 system. Methods Mol Biol. 2021;2242:15–42.
4. De Coster W, De Rijk P, De Roeck A, De Pooter T, D'Hert S, Strazisar M, et al. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. Genome Res. 2019;29(7):1178–87.
5. Xue Y, Ankala A, Wilcox WR, Hegde MR. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: Single-gene, gene panel, or exome/genome sequencing. Genet Med. 2015;17(6):444–51.
6. Bean LJH, Funke B, Carlston CM, Gannon JL, Kantarci S, Krock BL, et al. Diagnostic gene sequencing panels: from design to report—a technical standard of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2020;22(3):453–61.
7. Tetreault M, Bareke E, Nadaf J, Alirezaie N, Majewski J. Whole-exome sequencing as a diagnostic tool: current challenges and future opportunities. Expert Rev Mol Diagn. 2015;15(6):749–60.
8. Nisar H, Wajid B, Shahid S, Anwar F, Wajid I, Khatoon A, et al. Whole-genome sequencing as a first-tier diagnostic framework for rare genetic diseases. Exp Biol Med (Maywood). 2021;246(24):2610–7.
9. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: overviews and challenges. Biotechniques. 2014;56(2):61–4.
10. Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, Lyon E, et al. Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. Hum Mutat. 2015;36(9):903–14.
11. Rehder C, Bean LJH, Bick D, Chao E, Chung W, Das S, et al. Next-generation sequencing for constitutional variants in the clinical laboratory, 2021 revision: a technical standard of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2021;23(8):1399–415.
12. Tu J, Ge Q, Wang S, Wang L, Sun B, Yang Q, et al. Pair-barcode high-throughput sequencing for large-scale multiplexed sample analysis. BMC Genomics. 2012;13:43.
13. Sohn JI, Nam JW. The present and future of de novo whole-genome assembly. Brief Bioinform. 2018;19(1):23–40.
14. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010;11(5):473–83.
15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
16. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
17. Mahamdallie S, Ruark E, Yost S, Münz M, Renwick A, Poyastro-Pearson E, et al. The quality sequencing minimum (QSM): providing comprehensive, consistent, transparent next generation sequencing data quality assurance. Wellcome Open Res. 2018;3:37.
18. Wang Z, DiVincenzo C, Elzinga C, Bazinet M, Batish SD, Jaremko M, et al. Zygosity detection by next generation sequencing in a clinical laboratory (P1.329). Neurology. 2014;82(10 Supplement).
19. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405–24.
20. Duzkale H, Shen J, Mclaughlin H, Alfares A, Kelly M, Pugh T, et al. A systematic approach to assessing the clinical significance of genetic variants. Clin Genet. 2013;84(5):453–63.
21. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.
22. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, et al. dbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. 2013;41(Database issue):D936–41.
23. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res. 2020;48(D1):D941–7.
24. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.
25. Stenson PD, Mort M, Ball E V., Chapman M, Evans K, Azevedo L, et al. The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. Hum Genet. 2020;139(10):1197–207.

»

26. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(Database issue):D789–98.
27. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–7.
28. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
29. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4.
30. Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res. 2001;29(5):1185–90.
31. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–5.
32. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
33. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20(7):389–403.
34. Miller DT, Lee K, Abul-Husn NS, Amendola LM, Brothers K, Chung WK, et al. ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2023;25(8):100866.
35. Mandelker D, Amr SS, Pugh T, Gowrisankar S, Shakhbatyan R, Duffy E, et al. Comprehensive diagnostic testing for stereocilin: An approach for analyzing medically important genes with high homology. J Mol Diagn. 2014;16(6):639–47.
36. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2011;13(1):36–46.
37. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 2019;47(21):10994–1006.
38. Shin S, Park J. Characterization of sequence-specific errors in various next-generation sequencing systems. Mol Biosyst. 2016;12(3):914–22.
39. Nord AS, Lee M, King MC, Walsh T. Accurate and exact CNV identification from targeted high-throughput sequence data. BMC Genom. 2011;12:184.
40. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. BMC Bioinform. 2013;14 Suppl 11(Suppl 11):S1.
41. Yang R, Nelson AC, Henzler C, Thyagarajan B, Silverstein KAT. ScanIndel: A hybrid framework for indel detection via gapped alignment, split reads and de novo assembly. Genome Med. 2015;7:127.
42. Li J, Dai H, Feng Y, Tang J, Chen S, Tian X, et al. A comprehensive strategy for accurate mutation detection of the highly homologous PMS2. J Mol Diagn. 2015;17(5):545–53.
43. Yap ZY, Park YH, Wortmann SB, Gunning AC, Ezer S, Lee S, et al. Functional interpretation of ATAD3A variants in neuro-mitochondrial phenotypes. Genome Med. 2021;13(1):55.