Pregledni znanstveni članek ■

Vpliv parametrov sekvenciranja naslednje generacije na zanesljivost rezultatov v metagenomskih študijah

Vasja Progar, Uroš Petrovič

Izvleček. Metagenomika, ki se ukvarja s celostnim proučevanjem genskega materiala vseh organizmov, prisotnih v izbranem okolju, je z razvojem metod sekvenciranja naslednje generacije (NGS) doživela svoj razmah. Med ključnimi parametri platform NGS, ki vplivajo na zanesljivost pridobljenih rezultatov, so dolžina odčitkov, globina sekvenciranja ter natančnost odčitkov. V tem pregledu opisujemo ključne parametre NGS pri metagenomskih študijah in strategije za izboljšanje zanesljivosti le-teh. Parametre in strategije podrobneje obravnavamo na primeru dveh nedavnih študij - določanju sestave človeškega ustnega mikrobioma ter iskanju novih genov za termo-stabilno razgradnjo celuloze.

The Effect of Next Generation Sequencing Parameters on the Reliability of Metagenomic Studies

Abstract. Metagenomics, the integral study of all organisms present in the selected environment, has rapidly advanced with the advent of next generation sequencing (NGS) methods. Among the key parameters of NGS that affect the reliability of the results are read length, sequencing depth and read accuracy. In this review, we depict the essential NGS parameters in metagenomic projects and strategies to improve their reliability. We then further discuss the parameters and strategies along two recent case studies - determination of human oral microbiome and mining of novel genes for thermo-stable degradation of cellulose.

■ **Infor Med Slov**: 2013; 18(1-2): 1-8

Instituciji avtorjev / Authors' institutions: Biotehniška fakulteta, Univerza v Ljubljani (VP); Institut Jožef Stefan, Ljubljana (UP).

Kontaktna oseba / Contact person: Vasja Progar, Biotehniška fakulteta, Univerza v Ljubljani, Jamnikarjeva 101, 1000 Ljubljana. e-pošta / e-mail: vasja.progar@bf.uni-lj.si.

Prejeto / Received: 7.5.2013. Sprejeto / Accepted: 30.7.2013.

Uvod

Metagenomika celostno proučuje genski material vseh organizmov, prisotnih v danem okolju. Izraz metagenom so prvič uporabili leta 1998 Handelsman in sodelavci,[1] ko so skupek genomov organizmov prstne mikroflore poimenovali *metagenom* prsti. Metagenomika je bila kasneje definirana kot »uporaba modernih genomskih tehnik za proučevanje mikroorganizmov neposredno v njihovih naravnih okoljih, brez potrebe po izolaciji in laboratorijskemu gojenju posameznih vrst«.[2] Slednje je hkrati ena glavnih prednosti metagenomskih študij, saj drugi pristopi, ki temeljijo na predhodnem gojenju, praviloma zajamejo le majhen del vrst mikroorganizmov v okoljskem vzorcu - pogosto manj kot 1%, v posebnih primerih pa do 23%.[3] Velika večina vrst mikroorganizmov v združbi torej z metodami, ki vključujejo predhodno gojenje, ni zaznana, metagenomika pa nam omogoča vpogled tudi v to, za gojenje neprimerno skupino mikroorganizmov.

V grobem sta za metagenomske študije značilna dva pristopa - tarčni in naključni. Tarčni pristop se osredotoča predvsem na identifikacijo mikroorganizmov v združbi in sicer s t.i. amplikonskimi študijami, ki z analizo enega oziroma manjšega števila genskih označevalcev omogočajo vpogled v sestavo in raznolikost mikroflore v vzorcu; najpogosteje uporabljan označevalec pri metagenomskih študijah je zaradi evolucijske ohranjenosti 16S rDNA oziroma 18S rDNA.[4] Za razliko od tarčnega pa naključni pristop nukleotidna zaporedja izbira po inherentno naključni metodi hitrega sekvenciranja (angl. *shotgun sequencing*) iz preučevanega okolja izolirane DNA/RNA. Medtem ko tarčni pristop cilja predvsem na identifikacijo mikroorganizmov, predstavlja naključni pristop metagenomiko v bolj celostnem smislu, saj omogoča pridobitev bogatejših podatkov o funkcionalnem potencialu mikrobne združbe; natančnost določitve sestave le-te pa je manjša v primerjavi s tarčnim pristopom.[4]

Ena ključnih pridobitev, ki je imela velik vpliv na metagenomiko in njen razmah, je razvoj metod sekvenciranja naslednje generacije (angl. *next generation sequencing*; v nadaljevanju NGS), znanih tudi kot visokozmogljivostne metode sekvenciranja (angl. *high-throughput sequencing*), kot so med drugim v metagenomiki najpogosteje uporabljani platformi Roche 454 (pirosekvenciranje) in Illumina, ter nekoliko manj uporabljana platforma SOLiD.[3-5] Pred nastopom teh metod je bilo določanje zaporedja opravljeno po Sangerjevi kapilarni metodi, ki je s pripravami klonske knjižnice predstavljala ozko grlo v metagenomskih študijah. Z velikimi količinami proizvedenih podatkov pri metodah NGS (stotine milijonov odčitkov na eksperiment) in odsotnostjo potrebe po predhodni pripravi klonskih knjižnic pa se je ozko grlo premaknilo v smeri bioinformatske obdelave pridobljenih podatkov.[3] Poleg velike razlike v količini odčitkov (angl. *reads*) se metode NGS od predhodno prevladujoče Sangerjeve metode sekvenciranja razlikujejo tudi v dolžini odčitkov. Ti so v povprečju dosti krajši in obsegajo od ~50 nukleotidov do ~600 nukleotidov v primerjavi z ~800 nukleotidov dolgimi odčitki pridobljenimi po Sangerjevi metodi. Za metode NGS je značilna tudi nekoliko višja stopnja napak - te predstavljajo tipično približno 1% vseh določenih nukleotidov v primeru NGS, v primerjavi s približno 0,001% v primeru Sangerjeve metode.[4]

Ob tem se platforme NGS po svojih značilnostih močno razlikujejo tudi med seboj, za čim boljšo zanesljivost metagenomskih študij pa je potrebno zagotoviti pravilno ravnovesje med posameznimi značilnimi parametri. Čeprav so zaradi medsebojnih razlik in specifičnosti posamezne platforme bolj ali manj primerne za naslavljanje določenih znanstvenih vprašanj, so nekateri parametri in njihove značilnosti med platformami univerzalni. V tem pregledu se bomo osredotočili na tri izmed njih, in sicer na dolžino odčitkov, natančnost in globino sekvenciranja. Ti so skupni vsem zgoraj omenjenim konkretnim platformam NGS, njihove značilnosti pa je potrebno smiselno upoštevati pri zasnovi poskusov na vsaki od omenjenih platform, kakor tudi morebitnih

prihodnjih platformah. V nadaljevanju bomo opisali značilnosti posameznih parametrov, kako ti parametri vplivajo na metagenomske projekte in kakšne so praktične rešitve za izboljšanje zanesljivosti metagenomskih študij. Nato bomo prikazali pomen teh parametrov v dveh konkretnih primerih metagenomskih študij - določanju sestave človeškega ustnega mikrobioma ter iskanju novih genov za termo-stabilno razgradnjo celuloze.

Parametri NGS

Dolžina odčitkov

Dolžina odčitkov je izražena s številom nukleotidov in označuje dolžino (povprečno ali konkretno) posameznih neprekinjenih zaporedij nukleotidov, določenih z izbrano metodo sekvenciranja. Pri tem je dolžino odčitkov (angl. *read length*) potrebno razlikovati od dolžine vstavkov (angl. *insert size*) - slednja nam pove, kako dolgi (povprečno) so izhodiščni fragmenti DNA oziroma RNA, ki so predmet sekvenciranja.

Glede poimenovanja po dolžini odčitkov ni splošno sprejetega standarda, a pogosto veljajo za kratke odčitki z do 50 nukleotidi, za srednje dolge tisti s 50 do 400 nukleotidi in za dolge oni s 400 do 1000 nukleotidi.[6] Slednji so primerljivi z dolžino odčitkov pridobljenih po Sangerjevi metodi, presegajo pa jih le še 'podaljšane' dolžine odčitkov (angl. *extended reads*), nad 1000 nukleotidov, ki so trenutno dosegljive le pri platformah PacBio in Starlight, vendar ob občutno manjši natančnosti sekvenciranja odčitkov.[6] Potrebno pa se je zavedati, da tehnologija NGS platform napreduje zelo hitro, kar pomeni, da se tudi omenjene okvirne vrednosti lahko že kmalu spremenijo.

Pri določanju nukleotidnih zaporedij je seveda zaželena čim večja dolžina odčitkov, saj ta olajša proces sestavljanja (angl. *assembly*) in zmanjša pogostnost pojavljanja dvoumnih poravnav z referenčnimi genomi ter posledično olajša iskanje homolognih genov po podatkovnih bazah. Po drugi strani se na račun daljših odčitkov povečuje stopnja napak,[5-7] kar ponovno oteži nadaljnjo uporabo podatkov in kaže na tesno prepletenost parametrov NGS.

Metagenomske študije so na dolžino odčitkov posebno občutljive na račun velikega števila prisotnih mikroorganizmov - daljši odčitki lahko pripomorejo k izboljšanju zanesljivosti taksonomske klasifikacije, določanja novih taksonov ali prepoznavanju specifičnih biooznačevalnih (angl. *biomarker*) organizmov.[4] Da kratke dolžine odčitkov, značilne za metode NGS, ne bi imele prevelikega vpliva na zmanjšanje zanesljivosti študij, so raziskovalci ubrali različne strategije, ki med seboj niso izključujoče. Najenostavnejša med njimi je povečanje globine sekvenciranja (podrobneje obravnavana v sledečem razdelku), ki privede do tesnejšega prekrivanja med posameznimi odčitki, kar olajša njihovo sestavljanje.

Drugo pomembno strategijo za izboljšanje zanesljivosti raziskav temelječih na NGS predstavljajo obojestranski odčitki (angl. *paired-end reads*) - pri teh je zaporedje vsakega fragmenta DNA določeno z obeh koncev molekule. Običajno se za metodo obojestranskih odčitkov uporabijo izhodiščni fragmenti dolžine med 100 in 300 nukleotidi in sicer z dolžino odčitkov med 76 in 125 nukleotidov - tako se nekateri pari odčitkov medsebojno prekrivajo, medtem ko ostane pri drugih sredinski del nukleotidnega zaporedja nedoločen.[4] Prekrivajoči se pari odčitkov tako pravzaprav efektivno predstavljajo odčitke z večjo dolžino, bogatejšo informacijo pa vsebujejo tudi neprekrivajoči se obojestranski odčitki, saj lahko z njihovo pomočjo sklepamo, da oba odčitka prihajata iz istega izhodiščnega fragmenta in posledično istega organizma. Obstajajo pa tudi izjeme, kot so npr. himerna zaporedja, ki so lahko posledica napak pri pomnoževanju in povzročijo napačno sklepanje o povezanosti določenih zaporedij.

Poleg omenjenih dveh strategij za kompenzacijo kratke dolžine odčitkov obstajajo še druge, kot je npr. izbor krajšega tarčnega zaporedja pri tarčnem

metagenomskem pristopu (npr. primerjava zgolj krajšega odseka 16S rDNA namesto celotnega zaporedja[9]) ali pa komplementacija z daljšimi odčitki, pridobljenimi na drugih platformah NGS.

Globina sekvenciranja

Globina sekvenciranja (angl. *sequencing depth*) je skupno število odčitkov ali nukleotidov, pridobljeno pri enem sekvenciranju ali seriji sekvenciranj.[10] V tesni povezavi z globino sekvenciranja je pokritost (angl. *coverage*), ki odraža delež izhodiščnih zaporedij, ki je bil sekvenciran.[10] Pokritost je torej poleg globine sekvenciranja odvisna tudi od raznolikosti izhodiščnih nukleotidnih zaporedij, kar se pri kompleksnih metagenomskih vzorcih prevede v zahtevo po zelo veliki globini sekvenciranja, če želimo doseči dobro pokritost, na primer če želimo določiti zaporedje čim večjemu deležu genomov vseh organizmov v vzorcu. Še posebno to velja v primeru, da so izhodiščna zaporedja neenakomerno zastopana (npr. nekaj prevladujočih organizmov) in želimo zanesljivo identificirati tudi tista, ki so zastopana slabše. Za zanesljivo sestavitev genoma določene vrste iz kompleksnega metagenoma je potrebna pokritost ocenjena na približno 20-kratno,[11] Iverson in sodelavci[12] pa so poročali o uspešni *de novo* sestavitvi genoma negojljive morske arheje iz metagenoma površinske morske vode, v katerem je ta arheja predstavljala le 1,7% odčitkov.

Stroški določanja zaporedja so za določeno platformo v neposredni povezavi z globino sekvenciranja,[13] kar je razumljivo, saj pomeni povečevanje globine sekvenciranja pravzaprav večkratno ponavljanje naključnega sekvenciranja. Dobra lastnost v povezavi s tem pa je, da lahko globino naknadno povečamo z vnovičnim sekvenciranjem istega vzorca, če se izkaže, da je pokritost premajhna.

Vendar pa povečevanje globine samo po sebi mnogokrat ni dovolj, še posebno v kompleksnih vzorcih. Qin in sodelavci[8] so na primer v raziskavi mikrobioma človeškega črevesa z globino sekvenciranja presegli več milijard nukleotidov na vzorec, vendar je bilo le manj kot pol vseh odčitkov možno sestaviti v soseske (angl. *contigs*) daljše od 500 nukleotidov, večina od teh pa je bila še vedno krajša od 2200 nukleotidov.

Mende in sodelavci[13] pa so v svoji študiji primerjali Sangerjevo metodo, pirosekvenciranje ter Illumina metodo na simuliranih metagenomih; izkazalo se je, da v enostavnejših metagenomih (10 genomov) ni večjih razlik v pokritosti na organizem, v kompleksnejših metagenomih (100 genomov) prednjači Illumina prav na račun večje globine sekvenciranja, pri metagenomih s 400 genomi pa niti velika globina ne more izboljšati slabe pokritosti vsakega od genomov in kot uspešnejša se izkaže Sangerjeva metoda zaradi večjih izhodiščnih dolžin odčitkov.

Tudi Kuczynski in sodelavci[4] so izrazili dvom o zgolj povečevanju globine sekvenciranja kot rešitvi za težave sestavljanja pri metagenomiki in pokazali, da je tako pri taksonomskih kot pri funkcionalnih analizah za iskanje povezav med geni, organizmi ter fiziološkimi in bolezenskimi stanji, bolj informativno sekvenciranje večjega števila vzorcev kot povečevanje globine sekvenciranja posameznega vzorca.

Natančnost odčitkov

Natančnost odčitkov je odvisna od več vrst napak, ki se lahko pojavijo pri določanju nukleotidnega zaporedja, ali pa že predhodno pri pripravi knjižnic oziroma pri pomnoževanju. Napako se običajno oceni z določitvijo zaporedja znanih, referenčnih genomov, za metagenomiko pa je velikega pomena ocena napake na kompleksnejšem vzorcu, na primer na umetni združbi, sintetično pripravljeni z združevanjem genomskih DNA ali kloniranih 16S rDNA fragmentov različnih izolatov.[4,5] Vendar pa je primerjava natančnosti med posameznimi platformami otežena, saj ima vsaka svoje pristranske napake, proizvajalci pa mnogokrat navajajo podatke o napakah ocenjene na podlagi primerov (npr. *E. coli*, Phi X itd.), ki so v korist njihove platforme.[6]

Za vse platforme NGS velja, da je natančnost odčitkov največja na začetku zaporedja in pada z dolžino odčitka;[7] pravzaprav je maksimalna dolžina odčitkov omejena prav s toleranco napak - če bi bili pripravljeni sprejeti več napak, bi lahko dobili na posameznih platformah daljše odčitke kot jih sicer.[6] Podobno kot pri mutacijah poznamo tudi pri določanju zaporedja tri glavne tipe napak: substitucije, kjer je nukleotid napačno določen, delecije, kjer je eden ali več nukleotidov izpuščenih ter insercije, kjer je eden ali več nukleotidov napačno dodanih.[7] Čeprav je iz že omenjenih razlogov napake pri posameznih platformah za splošne primere težko oceniti, se vse vrste napak pri platformah Illumina in 454 gibljejo malo pod 1%, občutno višje pa so na primer pri sistemu PacBio, kjer prihaja kar do 13% insercijskih napak.[4]

Natančnost odčitkov je pri metagenomskem sekvenciranju še bistveno bolj pomembna kot pri določanju zaporedja posameznega genoma. Običajno je namreč pri sekvenciranju posameznega genoma zaporedje vsake regije genoma določeno večkrat, pri metagenomskem določanju zaporedja pa se lahko zgodi, da je zaporedje posameznih fragmentov določeno le enkrat. Če je torej le-to določeno napačno, lahko napačno sklepamo, da fragment izhaja iz novega organizma.[4] Za povečanje zanesljivosti metagenomskih študij glede na natančnost odčitkov obstaja več pristopov. Ponovno je lahko v veliko pomoč povečevanje globine sekvenciranja; Luo in sodelavci[5] so opisali 10-kraten padec frekvence substitucijskih napak pri povečanju pokritosti sosesk z 2-kratne na 20-kratno, vendar so hkrati opazili, da frekvenca napak pri več kot 20-kratni pokritosti ostaja enaka. Zgolj povečevanje globine sekvenciranja ima torej tudi v tem primeru omejen domet.

Drugi, zelo pomemben pristop temelji na bioinformatskih filtrih za odkrivanje in popravljanje napak, ki so v neprestanem razvoju in so običajno že vgrajeni v potek obdelave podatkov (angl. *data processing pipeline*).[5] To velja tako za znane pristranske oziroma sistemske napake posameznih platform, kot za naključne napake pri določanju nukleotidov. V splošnem so odčitki s predpostavljenim relativno visokim deležem napak zavrženi že v prvih filtracijskih korakih na osnovi povprečne ocene kvalitete zaporedja (angl. *quality score*), števila in dolžine homopolimerov (t.j. zaporedij, ki vsebujejo ponavljajoče se identične nukleotide, na katerih se pojavlja še posebno veliko napak pri določanju zaporedja), števila napačno določenih nukleotidov v začetnih oligonukleotidih in dolžine samega zaporedja.[4]

Pomen parametrov NGS v metagenomskih študijah – konkretna primera

Metagenomska študija ustnega mikrobioma

Metagenomska študija ustnega mikrobioma, ki so jo opravili Lazarevic in sodelavci,[9] je po navedbah avtorjev prva metagenomska študija, v kateri je bila za določanje zaporedja uporabljena platforma Illumina. Namen študije je bil oceniti potencial NGS platforme Illumina za preučevanje raznolikosti človeškega ustnega mikrobioma.

Za klasifikacijo bakterij iz človeških ustnih vzorcev so avtorji uporabili delna zaporedja dobro karakteriziranega in evolucijsko ohranjenega gena za 16S rRNA. V prvem koraku so poravnali več kot 750 16S rDNA zaporedij iz podatkovne baze HOMD (Human Oral Microbiome Database, www.homd.org) in izbrali začetne oligonukleotide na ohranjenih področjih, konkretno na robnem območju okoli 82 nukleotidov dolge hipervariabilne V5 regije. Namen tega postopka je bil zagotoviti amplikone čim večjega deleža bakterij (ohranjeno zaporedje za začetne oligonukleotide), ki bi hkrati odražali filogenetsko pripadnost mikroba, iz katerega izvirajo (hipervariabilna regija znotraj amplikona). Nato so iz vzorcev, pridobljenih iz ustne votline treh zdravih oseb, takšne 16S V5 amplikone pripravili ter jih enostransko (angl. *single-end*) sekvencirali na platformi Illumina GAII s 76 cikli.

V bioinformatski fazi so nato najprej odstranili odčitke, ki so vsebovali nedoločene nukleotide ali nepravilno določena zaporedja začetnih oligonukleotidov ter tiste odčitke, ki so vsebovali več kot 12 enakih zaporednih nukleotidov. Odčitki so bili dolgi 72 nukleotidov, čemur so odvzeli 13 nukleotidov dolge začetne oligonukleotide, s čimer so ostala zaporedja dolžine 59 nukleotidov. Od nekaj manj kot 1,4 milijona odčitkov jih je prestalo preverjanje kvalitete malo več kot 1,2 milijona. Da bi omejili vpliv napak sekvenciranja, so avtorji naknadno odstranili še vse odčitke, katerih zaporedja so se pojavila manj kot trikrat, s čimer je ostalo okoli 865.000 odčitkov, ki so predstavljali približno 26.000 ločenih zaporedij.

Tako pridobljena zaporedja so nato medsebojno poravnali (angl. *multiple alignment*) in analizirali z več različnimi orodji za oceno raznolikosti ter taksonomsko analizo. Prevladovala so zaporedja, pripisana mikroorganizmom iz debel *Firmicutes* in *Proteobacteria*, ki so predstavljala po 30% vseh zaporedij, ter v nekoliko manjši meri mikroorganizmom iz debel *Actinobacteria*, *Fusobacteria* ter *TM7* (po 1 do 5% vseh zaporedij); mikroorganizmom ostalih debel je bil pripisan manj kot 1% zaporedij, nedoločenih pa je ostala približno tretjina; takšna razporeditev je v skladu s tistimi iz drugih raziskav človeškega ustnega mikrobioma.[14-16] Za odreditev zaporedja posamezni vrsti so upoštevali 3% raznolikost zaporedja (pri dolžini 59 nukleotidov to pomeni ločljivost dveh nukleotidov, kar je konzervativna ocena). Na podlagi tega so ocenili, da je v naboru podatkov okoli 8000 različnih filotopov (angl. *phylotypes*). S pomočjo krivulje razredčenja (angl. *rarefaction curve*) so avtorji predvideli, da bi bilo za odkritje novega unikatnega filotipa potrebnih 30.000 dodatnih odčitkov oziroma 120.000 dodatnih odčitkov za odkritje novega filotipa pri 3% raznolikosti. Z nadaljnjimi taksonomskimi analizami so določili pripadnost 135 rodovom, z najpogostejšima rodovoma *Neissera* in *Streptococcus*, ki sta skupno vsebovala okoli 70% vseh zaporedij, 43 rodov pa je bilo takšnih, ki v prejšnjih študijah oralnega mikrobioma niso bili določeni in jih ni bilo v podatkovni bazi Human Oral Microbiome Database.

Povzetek rezultatov te študije je, da je kljub kratki dolžini odčitkov taksonomska ocena na ravni debel s platformo Illumina (na amplikonih V5 regije 16S rDNA) dovolj zanesljiva za učinkovito primerjavo vzorcev. K še večji zanesljivosti in ločljivosti bi pripomogla uporaba obojsetranskih odčitkov, ki bi povečala dolžino odčitkov in kvaliteto zaporedij.

Iskanje novih genov za termo-stabilno razgradnjo celuloze

Xia in sodelavci[17] so izvedli študijo, v kateri so s pomočjo metagenomike znotraj obogatene termofilne, celulozo razgrajujoče brozge iskali nove gene za razgradnjo celuloze. Izhodiščni material je predstavljala specifična, delno umetno pripravljena mikrobna združba: anaerobna brozga iz obrata za ravnanje z odpadnimi vodami je bila v laboratorijskem bioreaktorju obogatena s celulolitičnimi in metanogenimi skupinami mikrobov na način, da je bila tretirana dve leti pri temperaturi 55°C in pH > 6,0 na substratu mikrokristalne celuloze z glukozo kot so-substratom. Brozga je bila zmožna dnevne predelave 1,15 kg celuloze na kubični meter. Za takšno izhodišče so se avtorji odločili iz praktičnega razloga, saj bi za genomsko zaporedje zgolj najbolj dominantne populacije iz vzorca prsti morali sekvencirati za okoli 6 milijard nukleotidov zaporedij in še mnogokrat več za določitev genomov manj zastopanih vrst, kar bi pomenilo prevelike stroške.[17] V primeru z želeno funkcijo obogatene biomase pa je spekter organizmov mnogo manjši in iskano zaporedje lahko iščemo skoraj izključno med potencialnimi kandidati.

V prvem koraku metagenomske študije so iz izolirane DNA pripravili knjižnico okoli 180 nukleotidov dolgih zaporedij in izvedli sekvenciranje na platformi Illumina HiSeq2000 s 100 nukleotidov dolgimi obojestranskimi odčitki, pri čimer je primarno kontrolo kvalitete prestalo skupno 12 milijonov odčitkov (1,2 milijarde nukleotidov). Nato so zaporedja *de novo* sestavili s

pomočjo sestavljalnega programa Velvet, za kar je bilo uporabljenih 75% skupnih odčitkov, 96% od teh pa je bilo sestavljenih v soseske daljše od 1000 nukleotidov, v skupni dolžini 28,5 milijonov nukleotidov. Najdaljša soseska je obsegala nekaj čez 200.000 nukleotidov. Iz omenjenih sosesk so s pomočjo spletnega orodja MetaGeneMark predvideli 31.500 odprtih bralnih okvirjev s povprečno dolžino 850 nukleotidov; 64% od teh naj bi predvidoma predstavljalo celotno gensko zaporedje. Da bi potrdili veljavnost sestavitve, so naključno izbrali deset domnevnih celulaznih genov (v prevedeni dolžini 98 do 917 aminokislin) in jih s pomočjo v ta namen pripravljenih začetnih oligonukleotidov skušali z verižno reakcijo s polimerazo (PCR) pomnožiti iz izhodiščnega vzorca. Pri tem jim je uspelo izolirati devet od desetih kandidatnih genov, ki so jih nato sekvencirali po Sangerjevi metodi in ugotovili nad 99% identičnost z napovedanimi geni. Primerjava zaporedij 31.500 bralnih okvirjev z anotiranimi zaporedji v podatkovnih zbirkah je pokazala, da ima okoli polovica predvidenih genov iz metagenoma brozge manj kot 50% podobnosti z znanimi geni iz podatkovne baze, kar kaže na veliko število potencialnih genov za doslej neznane termo-stabilne celulolitične encime.

Podobno kot pri zgoraj obravnavani študiji ustnega mikrobioma so tudi v tej študiji avtorji skušali določiti strukturo metagenoma s pomočjo genov rRNA, le da so v tem primeru izhajali iz celotnih 16S in 18S rRNA genov, ki so predstavljali okoli 0,15% vseh odčitkov metagenoma. Dobljena zaporedja so v 83,4% primerov pripisali bakterijam, v 11,1% arhejam, 1,3% evkariontom, 0,3% virusom, 4,0% zaporedij pa ni bilo določljivih. Kar 55% celotne populacije je predstavljal rod *Clostridium* in se s tem izkazal za glavnega razgrajevalca celuloze v mikrobiomu brozge, medtem ko sta bila glavna metanogena rodova *Methanothermobacter* (11,2%) ter *Methanosarcina* (1,3%). Tudi pri tej študiji je krivulja razredčenja pokazala, da je že bila dosežena globina sekvenciranja, pri kateri se s povečevanjem globine število novih vrst mikroorganizmov povečuje le počasi.

V nadaljevanju raziskave so avtorji s pomočjo primerjave določenih bralnih okvirjev s podatki iz podatkovnih baz odkrili več kot 200 kandidatnih genov za nove termo-stabilne encime za razgradnjo celuloze.

Zaključek in perspektiva

Metagenomika je z metodami NGS pridobila svoj pravi potencial in postaja vse dostopnejša tudi manjšim raziskovalnim skupinam. Očiten je tudi zelo hiter napredek področja. Že pri predstavljenih študijah, med katerima je štiri leta razlike, je opazen napredek tehnologije, v tem primeru platforme Illumina. Medtem ko so pri študiji ustnega mikrobioma[9] iz leta 2009 avtorji razpolagali z 1.4 milijona 72 nukleotidov dolgih, enostranskih odčitkov, je pri študiji iz leta 2013, iskanju novih genov za razgradnjo celuloze[17] izhodiščne podatke predstavljalo 12 milijonov 100 nukleotidov dolgih, obojestranskih odčitkov. K večji zanesljivosti poznejše izmed študij tako zagotovo prispeva količina in kvaliteta izhodiščnih podatkov, ki je občutno večja glede na drugo predstavljeno študijo, kot tudi protokoli, bioinformatski filtri in uveljavljanje dobrih praks pri sekvenciranju na platformah nove generacije. Izbira pravilnega ravnovesja parametrov NGS ima pomemben vpliv na zanesljivost metagenomskih študij in kot je značilno za vse nove tehnologije se postopoma uveljavljajo dobre prakse ter priporočljive metodologije ter se izostrujejo kvalitetne in premišljene zasnove eksperimentov. Vedno bolj je poudarjena bioinformatska plat analize, ki postaja tudi čedalje bolj zahtevna in kompleksnejša, pri čemer gre za problematiko od primernega shranjevanja podatkov do beleženja verzij programov in uporabljenih parametrov, kot je dobro povzeto v preglednem članku Nekrutenka in Taylorja.[18]

S številnimi področji uporabe in relativno dostopnostjo predstavlja metagenomika v kombinaciji z metodami sekvenciranja naslednje generacije obetavno področje in tako vir

uporabnih inovacij kot tudi velike količine novih znanj.

Zahvala

Zahvala za pomoč in svetovanje pri pripravi članka prof. dr. Tanji Kunej, prof. dr. Gregorju Anderluhu ter prof. dr. Branki Javornik. Posebna zahvala prof. dr. Marku Dolinarju za pomoč pri prevajanju strokovnih terminov.

Literatura

1. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM: Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* 1998; 5(10): R245-R249.
2. Chen K, Pachter L: Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol* 2005; 1(2): 106-112.
3. Teeling H, Glöckner FO: Current opportunities and challenges in microbial metagenome analysis - a bioinformatic perspective. *Brief Bioinform* 2012; 13(6): 728-742.
4. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, Knight R: Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet* 2012; 13(1): 47-58.
5. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT: Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* 2012; 7(2): e30087.
6. Glenn TC: Field guide to next-generation DNA sequencers. *Mol Ecol Resour* 2011; 11(5): 759-769.
7. Hoff KJ: The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics* 2009; 10: 250.
8. Qin J, Li R, Arumugam M et al.: A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010; 464(7285): 59-65.
9. Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, Osteras M, Schrenzel J, Francois P: Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods* 2009; 79(3): 266-271.
10. Wang Z, Gerstein M, Snyder M: RNA-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009; 10(1): 57-63.
11. Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT: Individual genome assembly from complex community short-read metagenomic datasets. *ISME J* 2012; 6(4): 898-901.
12. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV: Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* 2012; 335(6068): 587-590.
13. Mende DR, Waller AS, Sunagawa S, Jarvelin AI, Chan MM, Arumugam M, Raes J, Bork P: Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS One* 2012; 7(2): e31386.
14. Huyghe A, Francois P, Charbonnier Y et al.: Novel microarray design strategy to study complex bacterial communities. *Appl Environ Microbiol* 2008; 74(6): 1876-1885.
15. Keijser BJ, Zaura E, Huse SM et al.: Pyrosequencing analysis of the oral microflora of healthy adults. *J Dent Res* 2008; 87(11): 1016-1020.
16. Nasidze I, Li J, Quinque D, Tang K, Stoneking M: Global diversity in the human salivary microbiome. *Genome Res* 2009; 19(4): 636-643.
17. Xia Y, Fang HHP, Zhang T: Mining of novel thermo-stable cellulolytic genes from a thermophilic cellulose-degrading consortium by metagenomics. *PLoS One* 2013; 8(1): e53779.
18. Nekrutenko A, Taylor J: Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat Rev Genet* 2012; 13(9): 667-672.