Robert Šket[1], Zala Prevoršek[2], Deni Košeto[3], Aleksandar Sebastijanović[4], Simona Konda[5], Jerca Bajuk[6], Blaž Stres[7]

# Analitski in konceptualni izzivi pri raziskovanju človeške črevesne mikrobiote za potrebe personalizirane večnivojske medicine

## *Analytical and Conceptual Challenges in the Investigation of Human Intestinal Microbiota for the Needs of Personalized Multidimensional Medicine*


## IZVLEČEK

KLJUČNE BESEDE: 16S ribosomska ribonukleinska kislina, metagenomika, metabolomika, bioinformatika, personalizirana medicina

Črevesna mikrobiota ima v človeškem prebavnem traktu pomembno vlogo. Omogoča dinamičen odnos z gostiteljem, pomaga ohranjati homeostazo epitelijskih celic, vpliva na pridobivanje hranil in uravnavanje energije, fizično preprečuje kolonizacijo površine črevesa s patogenimi mikroorganizmi in sodeluje pri proizvodnji vitaminov ter kratkih verig maščobnih kislin. Črevesno mikrobioto poleg gliv, arhej in virusov v večini sestavlja $3 \times 10^{13}$ bakterij, ki jih uvrščamo v vsaj tisoč različnih bakterijskih vrst. Sestavo mikrobiote in njeno vlogo v črevesju lahko opredeljujemo s številnimi metodami, ki se razlikujejo v ločljivosti. Uporabnost rezultatov, ki jih dobimo, je tako odvisna od izbrane metode. Z nekaterimi metodami lahko le relativno primerjamo sestavo mikrobiote v različnih vzorcih,



[1] Dr. Robert Šket, mag. mikrobiol., Katedra za mikrobiologijo in mikrobno biotehnologijo, Oddelek za zootehniko, Biotehniška fakulteta, Univerza v Ljubljani, Jamnikarjeva ulica 101, 1000 Ljubljana; robert.sket@kclj.si

[2] Dr. Zala Prevoršek, univ. dipl. inž. zoot., Katedra za genetiko, animalno biotehnologijo in imunologijo, Oddelek za zootehniko, Biotehniška fakulteta, Univerza v Ljubljani, Jamnikarjeva ulica 101, 1000 Ljubljana

[3] Deni Košeto, mag. mikrobiol., Katedra za mikrobiologijo in mikrobno biotehnologijo, Oddelek za zootehniko, Biotehniška fakulteta, Univerza v Ljubljani, Jamnikarjeva ulica 101, 1000 Ljubljana; SINTEF Ocean, Strindveien 4, 7034 Trondheim, Norveška

[4] Aleksandar Sebastijanović, mag. mikrobiol., Katedra za mikrobiologijo in mikrobno biotehnologijo, Oddelek za zootehniko, Biotehniška fakulteta, Univerza v Ljubljani, Jamnikarjeva ulica 101, 1000 Ljubljana; Institut Jožef Stefan, Jamova cesta 39, 1000 Ljubljana

[5] Simona Konda, mag. mikrobiol., Katedra za mikrobiologijo in mikrobno biotehnologijo, Oddelek za zootehniko, Biotehniška fakulteta, Univerza v Ljubljani, Jamnikarjeva ulica 101, 1000 Ljubljana; Lek farmacevtska družba, d. d., Verovškova ulica 57, 1000 Ljubljana

[6] Jerca Bajuk, mag. mikrobiol., Katedra za mikrobiologijo in mikrobno biotehnologijo, Oddelek za zootehniko, Biotehniška fakulteta, Univerza v Ljubljani, Jamnikarjeva ulica 101, 1000 Ljubljana; Krka, d. d., Novo mesto, Šmarješka cesta 6, 8501 Novo mesto

[7] Izr. prof. dr. Blaž Stres, univ. dipl. mikrobiol., Katedra za mikrobiologijo in mikrobno biotehnologijo, Oddelek za zootehniko, Biotehniška fakulteta, Univerza v Ljubljani, Jamnikarjeva ulica 101, 1000 Ljubljana; Inštitut za zdravstveno hidrotehniko, Fakulteta za gradbeništvo in geodezijo, Univerza v Ljubljani, Hajdrihova ulica 28, 1000 Ljubljana; Center za klinično toksikologijo in farmakologijo, Interna klinika, Univerzitetni klinični center Ljubljana, Zaloška cesta 7, 1000 Ljubljana

druge nam do določene mere omogočajo popis sestave mikrobiote (mikromreže, sekvenciranje pomnožkov genov za 16S ribosomsko RNA, sekvenciranje metagenoma) in kot nadgradnja temu: z nekaterimi opredelimo tudi dejavnost posameznih predstavnikov črevesne mikrobiote (mikromreže transkriptov, sekvenciranje metatranskriptoma), ki se odraža v aktiviranju genov za določen proces. Izkaže se, da se linearno s širšo uporabnostjo metode oz. z večanjem obsega rezultatov povečujejo tudi stroški analize. Opisani so primeri metod za posamezen način opredeljevanja črevesne mikrobiote, izpostavljene so njihove pozitivne ter negativne lastnosti in uporabnost teh metod. Taksonomska klasifikacija in popis funkcionalnih genov pa med seboj niso povezani (determinacijski koeficient < 0,3). Prav tako različne taksonomske ravni dajo različne slike odnosov med vzorci. Majhne razlike v taksonomskem popisu so nekonsistentne, saj so za njihov popis potrebovali velike kohorte (število preiskovancev > 2.000). Nasprotno pa lahko na ravni metabolomov enostavno ločimo različne skupine, ki jih z uporabo popisa mikrobioma ne moremo.

## ABSTRACT

KEY WORDS: 16S ribosomal ribonucleic acid, metagenomics, metabolomics, bioinformatics, personalized medicine

Intestinal microbiota plays an important role in the human gastrointestinal tract. It enables dynamic relationship with the host, helps maintain homeostasis of epithelial cells, affects the acquisition of nutrients and regulation of energy, physically prevents the colonization of the intestine surface by pathogenic microorganisms and it is involved in the production of vitamins along with short chain fatty. Human intestinal microbiota, in addition to the fungi, archaea, and viruses, consists of $3 \times 10^{13}$ bacteria, which are classified into at least a thousand different bacterial species. The composition of the microbiota and its role in the gut can be defined using different methods with very different resolutions. Applicability of the obtained results thus depends on the method chosen. With some methods one can only relatively compare the composition of the microbiota in different samples others allow us to some extent absolute description of the microbiota composition (microarrays, amplicon sequencing of 16S ribosomal RNA genes, metagenome sequencing) and as an upgrade to that some methods made it possible to define the activity of individual microbes (microarrays, metatranscriptome sequencing), which is reflected in the activation of genes for a specific process. It turns out that linearly with the broader applicability of the methods, the cost of analysis expands. Here we describe methods of defining the human intestinal microbiota and expose positive and negative features of their usability. Taxonomic classification and the list of functional genes are not well correlated (coefficient of determination < 0.3). Also, different taxonomic levels give different images of the relationship between samples. The small differences in the taxonomic inventory are inconsistent since they needed large cohorts for their delineation (sample size > 2,000). On the other hand, at the level of the metabolites, we can easily distinguish between different groups that could not be separated by microbiome analysis.

## UVOD

Glede na do sedaj zbrane podatke ima črevesna mikrobiota ključno vlogo pri ohranjanju zdravja ljudi. Udeležena je pri oblikovanju homeostaze epitelijskih celic, omogoča proizvodnjo vitaminov in fizično onemogoča razrast patogenih mikroorganizmov *(1, 2)*. Za lažjo opredelitev vloge črevesne mikrobiote je treba opredeliti njeno sestavo v črevesju zdravega človeka. To je cilj raziskav v okviru projektov Človeški mikrobiom in MetaHIT (Metagenomics of the Human Intestinal Tract) *(3, 4)*. Več kot 90 % vseh bakterijskih vrst v črevesni mikrobioti spada v le dve bakterijski debli, to sta *Bacteroidetes* in *Firmicutes (4)*. Nato sledita debli *Proteobacteria* in *Actinobacteria* ter v manjšem deležu debla *Fusobacteria*, *Verrucomicrobia* in *Cyanobacteria (5)*. Kaj lahko povemo o sestavi ali pa dejavnosti mikrobiote v preiskovanem vzorcu, je odvisno od izbrane metode. Z različnimi metodami lahko:

- primerjamo sestavo mikrobiote med različnimi vzorci,
- opredelimo, kateri mikrobi sestavljajo mikrobioto in kakšna so razmerja med njimi,
- ugotovimo, kakšen je presnovni potencial mikrobov z njihovimi nabori genov in medsebojno soodvisnostjo in
- preučimo, kakšna je njihova presnovna vloga oz. kaj pravzaprav počnejo v trenutku odvzema vzorca.

V tej smeri narašča tudi nedorečenost metod, saj za merjenje dejavnosti *in situ* ne obstajajo metode za merjenje velikega števila vzorcev.

## METODE OPREDELJEVANJA KOMPLEKSNOSTI MIKROBNE ZDRUŽBE

Poznamo več metod za opredelitev črevesne mikrobiote, ki jih v grobem delimo na tradicionalne in molekularne. Med tradicionalne metode štejemo tiste, ki vključujejo gojenje. Primer takšnih metod sta štetje kolonij na trdnem (selektivnem) gojišču in metoda najverjetnejšega števila celic. Poleg tega, da moramo pri delu poznati rastne zahteve mikroorganizma, ki ga gojimo, je glavna slabost teh metod anomalija števnih plošč (angl. *great plate count anomaly*). Na gojiščih namreč lahko zraste le 0,01–10 % vseh prisotnih celic v mikrobnem vzorcu, odvisno od vzorca *(6)*. Zaradi tega danes pri ugotavljanju sestave mikrobiote in števila posameznih mikrobov v prebavnem traktu uporabljamo različne molekularne metode (tabela 1) *(7)*. Gojitvene tehnike s pridom uporabljamo pri rutinskih analizah kliničnih vzorcev.

Pri večini molekularnih metod kot filogenetski označevalec za taksonomsko uvrščanje organizmov uporabljamo gene za bakterijsko, arhejsko, glivno in protozojsko ribosomsko RNA (rRNA). Gre za regije 16S in 18S rRNA ter notranji prepisni vmesnik ribosomske DNA (angl. *internal transcribed spacer*, ITS). Edinstvena značilnost genov za rRNA je njihova splošna ohranjenost pri vseh bakterijah, arhejah ter evkariontih, pa tudi analitsko dovolj velika medvrstna raznolikost za ugotavljanje istovetnosti *(8)*. Tako lahko preučujemo mikrobno raznolikost, dobimo podatke o kvalitativni (prisotnost in odsotnost skupin) in kvantitativni zastopanosti bakterijskih vrst ter proučujemo dinamiko sprememb mikrobne združbe v primeru bolezni skozi čas in v populaciji različnih gostiteljev.

### Hibridizacija *in situ*

Hibridizacija *in situ* omogoča prepoznavanje posameznih mikrobnih celic in s tem filogenetsko prepoznavanje. Temelji na hibridizaciji kratkega fluorescentno označenega oligonukleotida s komplementarno sekvenco v rRNA. Fluorescentno označene celice lahko opazujemo z epifluorescenčnim ali konfokalnim mikroskopom oz. pretočnim citometrom. Metoda je zelo razširjena v mikrobni ekologiji in znana pod imenom fluorescenčna *in situ* hibridizacija *(6)*. Slabost metode je, da ne omogoča odkrivanja še

**Tabela 1.** Pregled uporabljanih metod za karakterizacijo črevesne mikrobiote. qPCR – kvantitativna verižna reakcija s polimerazo (angl. *quantitative polymerase chain reaction*), rRNA – ribosomska RNA, DNA – deoksiribonukleinska kislina, PCR – verižna reakcija s polimerazo (angl. *polymerase chain reaction*), DGGE – gelska elektroforeza v gradientu denaturanta (angl. *denaturing gradient gel electrophoresis),* TGGE – gelska elektroforeza v temperaturnem gradientu (angl. *temperature gradient gel electrophoresis),* T-RFLP – restrikcijski polimorfizem dolžine končnih fragmentov (angl. *terminal restriction fragment lenght polymorphism*), ARISA – avtomatizirana analiza notranje prepisane regije ribosomske DNA (angl. *automated ribosomal intergenic spacer analysis),* FISH – fluorescenčna *in situ* hibridizacija.

| Metoda | Opis | Prednosti | Slabosti |
|---|---|---|---|
| **Gojenje** | osamitev bakterij na selektivnem mediju | semikvantitativna metoda, poceni | veliko laboratorijskega dela, gojiti je mogoče manj kot 30 % črevesne mikrobiote |
| **qPCR** | pomnožitev in kvantifikacija 16S rRNA, reakcijska mešanica vsebuje spojino, ki fluorescira, ko se veže na dvoverižno DNA | omogoča filogenetsko identifikacijo, kvantitativna in hitra metoda | vpliv napak pri PCR, ne moremo prepoznati neznanih vrst |
| **Poliakrilamidna gelska elektroforeza v gradientu denaturanta** | ločevanje 16S rRNA--pomnožkov na gelu z dodatkom denaturanta (DGGE) ali višanjem temperature (TGGE) | semikvantitativna in hitra metoda, lise lahko izrežemo za nadaljnjo analizo | filogenetsko prepoznavanje ni mogoče, vpliv napak pri PCR |
| **T-RFLP** | pomnožitev s fluorescentno označenimi začetnimi oligonukleotidi in nato restrikcija 16S rRNA-pomnožkov z encimi, sledi ločitev fragmentov z gelsko elektroforezo | semikvantitativna, poceni in hitra metoda | filogenetsko prepoznavanje ni mogoče, vpliv napak pri PCR, nizka ločljivost |
| **ARISA** | pomnožitev odseka med 16S in 23S RNA, sledi ločitev fragmentov s kapilarno elektroforezo | semikvantitativna, poceni in hitra metoda | filogenetsko prepoznavanje ni mogoče, vpliv napak pri PCR |
| **FISH** | hibridizacija fluorescentno označenih oligonukleotidnih sond z geni 16S rRNA, zatem spremljanje pojava fluorescence s pretočnim citometrom | omogoča filogenetsko prepoznavanje, semikvantitativna metoda, ni vpliva napak pri PCR | uspešnost metode je odvisna od nukleotidnega zaporedja sonde, ne moremo določiti neznanih vrst |
| **DNA-mikromreže** | hibridizacija fluorescentno označenih oligonukleotidnih sond s komplementarnimi nukleotidnimi zaporedji, spremljanje fluorescence z laserjem | omogoča filogenetsko prepoznavanje, semikvantitativna in hitra metoda | navzkrižna hibridizacija, vpliv napak pri PCR, zahtevno zaznavanje malotevilnih vrst |
| **Sekvenciranje klonirane 16S rRNA-genov** | kloniranje celotnega pomnožka 16S rRNA, sekvenciranje po Sangerju in kapilarna elektroforeza | omogoča filogenetsko identifikacijo, kvantitativna metoda | vpliv napak pri PCR, zahtevno izvedljiva in draga metoda, napake pri kloniranju |
| **Sekvenciranje 16S rRNA-pomnožkov** | globoko sekvenciranje pomnožkov 16S rRNA | omogoča filogenetsko prepoznavanje in prepoznavanje neznanih bakterij, kvantitativna in hitra metoda | draga in zahtevno izvedljiva metoda, vpliv napak pri PCR |

| | | | |
|---|---|---|---|
| **»Shotgun« sekvenciranje metagenoma** | globoko sekvenciranje celotnih genomov celotne mikrobne združbe | omogoča filogenetsko prepoznavanje in prepoznavanje neznanih bakterij ter njihovih funkcionalnih skupin genov, kvantitativna metoda | draga metoda, zahtevna za izvedbo in računsko potratna analiza podatkov |
| **»Shotgun« sekvenciranje metatranskriptoma** | globoko sekvenciranje izraženih genov celotne mikrobne združbe | omogoča filogenetsko prepoznavanje, kvantitativna metoda, omogoča spremljanje bakterijske dejavnosti (izražanje genov) | draga in zahtevno izvedljiva metoda, zahtevna in računsko potratna analiza podatkov |

neznanih oz. neopisanih mikroorganizmov in je zaradi delovne intenzivnosti primernejša za sledenje enega ali nekaj tarčnih mikroorganizmov kot pa za ugotavljanje strukture mikrobiote.

## Poliakrilamidna gelska elektroforeza v gradientu denaturanta

Gelska elektroforeza v gradientu denaturanta (angl. *denaturing gradient gel electrophoresis*, DGGE) je elektroforetska metoda, ki jo največkrat uporabljamo za oceno razlik v sestavi mikrobiote med primerjanimi vzorci, saj omogoča ločevanje med različnimi mikrobnimi skupinami na osnovi razlik v zaporedju DNA, kar vpliva na stabilnost DNA. Zaradi teh razlik pride do različne elektroforetske mobilnosti delno denaturirane dvoverižne molekule DNA (angl. *double stranded DNA*, dsDNA) v poliakrilamidnem gelu z linearnim gradientom denaturanta. Večja kot je vsebnost parov gvanin-citozin, bolj je molekula dsDNA odporna na denaturacijo in kot taka se bo razklenila ter posledično ustavila pri višji koncentraciji denaturanta. To na gelu vidimo v obliki lise (angl. *band*) *(9)*. Nadalje lahko iz DGGE-gela izrežemo določene lise, z verižno reakcijo s polimerazo (angl. *polymerase chain reaction*, PCR) ponovno pomnožimo tam prisotno DNA in s sekvenciranjem izvedemo filogenetsko identifikacijo. Podobno kot DGGE tudi gelska elektroforeza v temperaturnem gradientu (angl. *temperature gradient gel electrophoresis*, TGGE) omogoča ocenitev sestave mikrobiote, le da v tem primeru namesto gradienta denaturanta med potekom analize višamo temperaturo *(10)*.

## Restrikcijski polimorfizem dolžine končnih fragmentov

Restrikcijski polimorfizem dolžine končnih fragmentov (angl. *terminal restriction fragment lenght polymorphism*, T-RFLP) tako kot DGGE in TGGE omogoča oceno sestave mikrobiote v vzorcu. Metoda omogoča ločevanje med različnimi mikrobiotami na podlagi spremljanja velikosti končnih restrikcijskih fragmentov DNA. Metodo izvedemo tako, da tarčno DNA ob pomnoževanju z metodo PCR na koncu 5' označimo s fluorokromom in po restrikciji z endonukleazo ločimo s kapilarno elektroforezo *(11, 12)*.

## Avtomatizirana analiza notranje prepisane regije ribosomske deoksiribonukleinske kisline

Avtomatizirana analiza notranje prepisane regije ribosomske DNA (angl. *automated ribosomal intergenic spacer analysis, ARISA*) je metoda, kjer v postopku PCR pomnožimo raznolike nekodirajoče odseke bakterijske DNA med 16S in 23S RNA, ki jih nato po velikosti ločimo s kapilarno elektroforezo. Metodo uporabljamo za primerjavo sestave mikrobiote med različnimi populacijami *(13, 14)*.

## DNA-mikromreže

Metoda omogoča filogenetsko prepoznavanje črevesne mikrobiote in se uporablja za primerjavo sestave mikrobiote med različnimi populacijami. Princip metode temelji na oligonukleotidnih sondah, ki so pritrjene na stekleni ploščici. Ko na ploščo dodamo fluorescentno označene pomnožke tarčne DNA, pride do hibridizacije med pomnožki in specifičnimi sondami. Fluorescenco vezanih pomnožkov nato spremljamo z laserjem *(15, 16)*.

Glavne težave zgoraj navedenih tehnik so predvsem velik obseg ročnega dela, nestandardizirane priprave materialov ter posledično velika spremenljivost v mejah ozadja, zaznavanja in kvantifikacije ter linearnosti zaznavanja teh tehnik.

## Verižna reakcija s polimerazo v realnem času

Kvantitativna verižna reakcija s polimerazo (angl. *quantitative polymerase chain reaction,* qPCR) je kvantitativna metoda, ki omogoča prepoznavanje že poznanih bakterijskih vrst oz. višjih taksonomskih skupin in oceno števila bakterij v vzorcih. Princip metode je podoben običajnemu PCR, le da je pri qPCR dodano fluorogeno označevanje začetnih oligonukleotidov, sond ali pomnožkov. Slednje omogoča sprotno spremljanje poteka reakcije. Povečanje količine pomnožka DNA spremljamo kot povečanje fluorescentnega signala, ki ga zaznamo z detektorji v napravi za qPCR. Količino sproščene fluorescence prikažemo grafično v odvisnosti od števila ciklov PCR in glede na potek rasti izmerjene fluorescence določimo cikel, kjer je reakcija prešla prag zaznave. Točka, imenovana $C_T$ (angl. *cycle threshold*) je odvisna od začetne količine tarčne DNA. Količino tarčne DNA v vzorcu kvantificiramo glede na standardno krivuljo, ki jo izrišemo na podlagi pomnoževanja standarda z znano količino tarčne DNA *(17, 18)*. Stranski učinek velikega obsega kontrole v delovanju te tehnike je, da lahko kvantitativno spremljamo le posamezne specifične skupine, ki smo jih posebej izbrali za zaznavanje in zanje pripravili ter umerili analitski sistem, zato je taksonomska širina temu primerno ozka, omejena na le nekaj izbranih skupin.

## Sekvenciranje

Sekvenciranje predstavlja zlati standard karakterizacije črevesne mikrobiote in je neodvisno od gojenja mikrobov. Gre za postopek določanja zaporedja deoksiribonukleotidov (angl. *deoxyribonucleotide triphosphate*, dNTP) adenina, gvanina, citozina in timina v molekuli DNA ali RNA, pridobljenih iz bakterij, arhej, rastlin, živali ali kakršnega koli drugega vira genetskih podatkov *(19)*. Uporaba sekvenciranja je zelo široka, tako lahko določimo zaporedja posameznih genov, večjih genetskih regij (klastri ali operoni), celotnih kromosomov ali kompletnih genomov. Informacije, pridobljene s sekvenciranjem, omogočajo prepoznavo sprememb v genih, povezav med boleznimi in fenotipi ter ugotavljanje kazalcev kroničnih bolezni in tarčnih mest delovanja zdravil. Sekvenciranje se uporablja tudi v evolucijski biologiji pri proučevanju razvoja in povezav med različnimi organizmi. Poleg tega, da s sekvenciranjem filogenetsko prepoznavamo in kvantificiramo preiskovano mikrobioto, lahko odkrijemo tudi nepoznane in v majhnih količinah prisotne mikroorganizme *(19, 20)*.

Začetni poskusi sekvenciranja so bili usmerjeni v sekvenciranje najbolj dostopnih, relativno čistih vrst molekul RNA, kot so mikrobna informacijska RNA, prenašalna RNA ali genomi enoverižnih RNA-bakteriofagov *(21)*. Začetke avtomatiziranega sekvenciranja DNA, ki mu pravimo tudi prva generacija sekvenciranja, so predstavili Frederick Sanger in sodelavci leta 1977 *(22)*. Uporabljali so radioaktivno označene kemične analoge dNTP, t. i. dideoksinukleotide (angl. *dideoxynucleotide triphosphate*, ddNTP) Njihova značilnost je, da nimajo 3' hidrok-

silne skupine, potrebne za podaljšanje verige DNA, zato ne pride do njihove vezave s 5' fosfatno skupino naslednjega dNTP. Radioaktivno označeni ddNTP se nahajajo v reakcijski mešanici skupaj s standardnimi dNTP. Vsakič, ko DNA-polimeraza vgradi ddNTP, se sinteza verige DNA ustavi. Kot rezultat dobimo veliko število fragmentov DNA različnih dolžin. Kadar izvajamo štiri vzporedne reakcije, od katerih vsaka vsebuje svojo vrsto radioaktivno označenega ddNTP, lahko po ločevanju vzorcev na poliakrilamidnem gelu s pomočjo avtoradiografije ugotovimo zaporedje nukleotidov v izvirnem zaporedju DNA *(22)*. Ta metoda sekvenciranja je prevladovala od 80. let prejšnjega stoletja do leta 2000. V tem obdobju je prišlo do velikega tehnološkega napredka. Kot izboljšava metod prve generacije sekvenciranja se pojavljajo metode sekvenciranja naslednjih generacij, ki omogočajo učinkovitejše sekvenciranje in nižje stroške postopka *(23)*. Glavna prednost teh metod sekvenciranja je, da za sekvenciranje ni potrebna klonirana DNA kot pri metodi po Sangerju, ampak lahko neposredno ugotavljamo nukleotidno zaporedje skupne DNA mikrobne združbe.

V sklop druge generacije sekvenciranja uvrščamo tržno dostopne platforme, kot so Roche/454, Illumina/Solexa, Applied Biosystems/SOliD in Helicos BioSciences. Omenjene metode sekvenciranja omogočajo karakterizacijo tarčnega genoma s precej nižjimi stroški v primerjavi s sekvenciranjem prve generacije *(24)*.

Metoda 454-pirosekvenciranja (Roche) temelji na pomnoževanju molekul DNA znotraj vodnih kapljic v oljni raztopini (PCR v emulziji), pri čemer vsaka vodna kapljica vsebuje eno molekulo DNA, ki je pritrjena na kroglico, na kateri so začetni oligonukleotidi. Po končanem pomnoževanju na kroglici dobimo klonsko kolonijo. Sekvenator vsebuje številne pikotitrske vdolbine, v katerih se nahajajo po ena kapljica z DNA in encimi, ki omogočajo kemiluminiscenčno zaznavanje. Sekvenciranje poteka tako, da DNA-polimeraza dodaja posamezne nukleotide v verigo glede na originalni zapis DNA. Ko je posamezen nukleotid dodan v verigo, s pomočjo encimov ATP-sulfurilaze in luciferaze nastane svetloba, ki jo merimo *(25)*.

Pri metodi sekvenciranja Illumina na naključne fragmente DNA vežemo adapterje in fragmente imobiliziramo na trdo podlago, na kateri poteka metoda mostovnega pomnoževanja (angl. *bridge PCR*) *(26)*. Molekule DNA se množijo tako, da se formirajo lokalne klonske kolonije DNA ali »klastri DNA«. Za določitev zaporedja DNA dodamo štiri vrste reverzibilnih terminacijskih baz, vse ostale nukleotide, ki se niso vezali, pa izperemo. Kamera naredi sliko fluorescenčno obeleženih nukleotidov. Za tem se barva, vključno z blokatorjem na koncu 3', kemično odcepi od molekule DNA in je na ta način omogočena nastavitev procesa oz. začetek novega cikla. Slednje omogoča optimalno prepustnost in teoretično neomejeno zmogljivost sekvenciranja. Tako je danes prepustnost lahko več kot 1.000.000 nukleotidov na sekundo, kar je približno enako enemu človeškemu genomu z enkratno pokritostjo na uro delovanja instrumenta *(27)*.

Ion Torrent polprevodniško sekvenciranje temelji na standardni tehnologiji sekvenciranja, vendar z uporabo novega polprevodniškega sistema zaznavanja. Ta metoda sekvenciranja temelji na zaznavanju vodikovih ionov, ki se sprostijo med polimerizacijo DNA. Uporablja se plošča z mikroluknjami, ki vsebujejo eno verigo molekule DNA, ki se sekvencira. Dodaja se en nukleotid naenkrat. V luknji, v kateri je prišlo do vezave nukleotida zaradi homologije, pride do sprostitve vodikovega iona, ki ga zazna izjemno občutljivi senzor. V primeru, da so prisotne homopolimerne ponovitve, se bo več istih nukleotidov vezalo na istem mestu v enem ciklu, več vodikovih ionov pa se bo sprostilo, kar bo pripeljalo do sorazmerno višjega elektronskega signala *(28)*.

V drugo generacijo sekvenciranja spadata še metodi SOLiD in DNA nanoball. Pri prvi sekvenciramo v kroglicah pomnožene molekule DNA kot pri pirosekvenciranju in na koncu dobimo sekvence, ki imajo primerljivo količino in dolžino kot pri Illumina sekvenciranju *(29, 30)*. Pri drugi metodi pa uporabljamo tehniko krožnega pomnoževanja (angl. *rolling circle replication*) za pomnoževanje majhnih fragmentov genomske DNA v t. i. nanožoge DNA (angl. *DNA nanoballs*). Ta metoda omogoča, da se veliko število nanožog sekvencira naenkrat, pri čemer so stroški reagentov v primerjavi z drugimi metodami sekvenciranja naslednje generacije zelo nizki *(31)*. Po drugi strani pa se iz vsake nanožoge določijo le kratka zaporedja, zaradi česar je mapiranje kratkih odčitkov z referenčnim genomom oteženo *(32)*.

Hkrati z razvojem tehnologij se povečuje prepustnost metod in se nižajo stroški ter čas, potrebni za pridobivanje rezultatov, kar je cilj metod sekvenciranja tretje generacije *(33)*. Sem vključujemo metode, ki temeljijo na:

- sintezi,
- branju zaporedja DNA, dokler molekula DNA prehaja skozi nanoporo in
- mikroskopskih tehnikah, kot sta mikroskopiranje z atomsko silo in transmisijska elektronska mikroskopija.

Te metode se uporabljajo za ugotavljanje položaja posameznih nukleotidov znotraj dolgih fragmentov DNA, ki so večji od 5.000 baznih parov. Komercialno najpogosteje uporabljeni platformi sta PacBio (Pacific Biosciences) in MinION (Oxford Nanopore Technologies).

Enomolekulsko sekvenciranje v realnem času (angl. *single molecule real time sequencing*) podjetja Pacific Biosciences temelji na pristopu sekvenciranja s sintezo. DNA se sintetizira v majhnih luknjicah (angl. *zero-mode wave-guides*), v katerih je na dno pritrjena DNA-polimeraza. Nato se pri dodajanju fluorescenčno označenih nukleotidov v verigo DNA sprošča fluorescenca. Rezultat so odčitki 20.000 ali več nukleotidov s povprečno dolžino 5.000 baznih parov *(34)*.

Nanoporna metoda sekvenciranja DNA podjetja Oxford Nanopore Technologies temelji na odčitku električnih signalov, ki se pojavljajo v času prehoda nukleotidov skozi beljakovinsko nanoporo. Koncept izvira iz ideje, da enoverižne molekule DNA ali RNA v linearnem zaporedju elektroforetsko prevajamo skozi biološko poro, pri tem pa se spremeni elektronski tok. Sprememba je odvisna od oblike, velikosti in dolžine zaporedja DNA, saj vsak od nukleotidov spremeni elektronski tok za različno časovno obdobje. Signali, ki ustrezajo zaporedju nukleotidov, se nato ovrednotijo *(35)*. Za uspešno izvajanje je ključnega pomena natančen nadzor prehoda molekule DNA skozi poro *(36)*. Obstajata dve metodi nanopornega sekvenciranja; metoda sekvenciranja v trdnem stanju in metoda, ki temelji na uporabi beljakovin *(35, 37)*.

Pravilen odvzem vzorca in kakovostna ekstrakcija nukleinskih kislin (DNA, RNA, deplecija rRNA, reverzna transkripcija v komplementarno DNA) sta osnova pravilne razlage rezultatov. Naslednji korak je sekvenciranje DNA z zgoraj opisanimi pristopi. Postopek sekvenciranja je sestavljen iz priprave knjižnic, pomnoževanja in sekvenciranja dobljenih pomnožkov ter nato bioinformatske in statistične analize dobljenih podatkov. Glede na način izvedbe sekvenciranja lahko določamo:

- mikrobne sekvence določenega mikrobnega odseka mikrobov v združbi,
- naključne odseke iz celotnih sekvenc teoretično vseh mikrobov v združbi in
- sekvence mikrobnih molekul rRNA celotne združbe.

Analizo podatkov sekvenciranja lahko opravimo na več načinov in z uporabo več programskih orodij, odvisno od tega, ali smo določali nukleotidno zaporedje mikrobnega pomnožka ali celokupnih mikrobnih

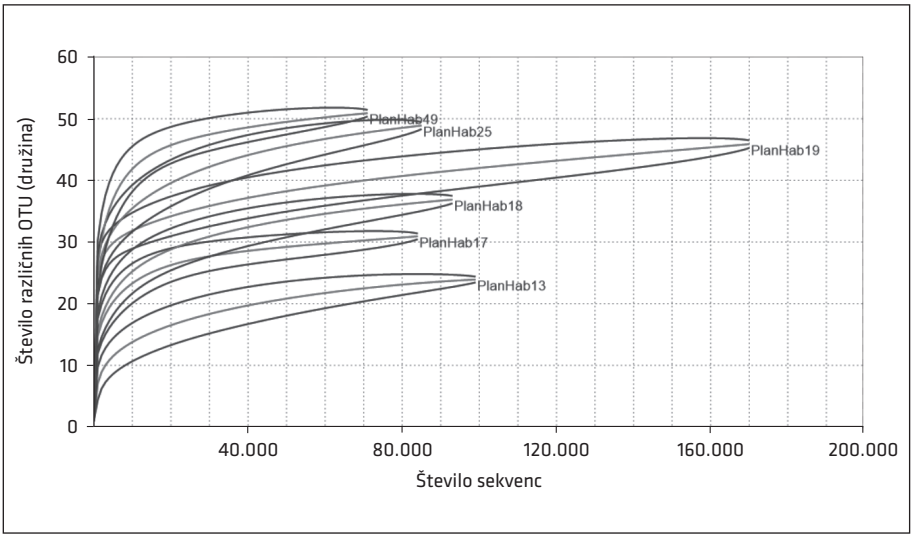
genomov v vzorcu. Pri tem želimo odgovoriti na vprašanja:

- Kako raznolika je mikrobiota?
- Kako se dve ali več mikrobiot razlikujeta/jo med sabo?
- Kdo so člani preiskovanih mikrobiot?
- Kako so ti člani med seboj povezani?

Pri iskanju odgovorov uporabljamo številna statistična in bioinformatska orodja *(38–40)*. Odgovori so kvalitativni, če preiskujemo, ali je določen član prisoten ali odsoten, ter kvantitativni, če preučujemo, npr. kako velike so skupine, gledano pri enakem številu uporabljenih kakovostnih sekvenc.

Za dovolj informativne podatke sekvenciranja je treba zagotoviti, da je določen odsek preiskovanega mikrobnega gena dovoljkrat prebran oz. sekvenciran. Tehnologije sekvenciranja nove generacije kot rezultat podajo od 250 in do več 1.000 baznih parov velik odsek. Število branj, ki jih dobimo pri sekvenciranju, opredelimo z izrazom globina sekvenciranja. S tehničnega vidika izvedbe sekvenciranja mora biti globina sekvenciranja tako velika, da je mogoče razlikovati med maloštevilnimi mikrobnimi skupinami in napakami sekvenciranja. Pokaže se namreč, da je pri vsaki izvedbi sekvenciranja določen odstotek branj napačen, izkustveno lahko tudi več kot 50 % vseh sekvenc v določenem vzorcu. Zato z ustrezno globino sekvenciranja zagotovimo, da imajo tudi maloštevilne skupine mikrobov zadostno število branj, in jih lahko nato ločimo od napak. Z biološkega vidika globino sekvenciranja posameznega vzorca grafično ponazorimo z rarefakcijsko krivuljo, kjer filogenetska raznolikost narašča z globino sekvenciranja (slika 1) *(41)*. Optimalna globina sekvenciranja je dosežena takrat, ko se filogenetska raznolikost, kljub večanju globine sekvenciranja, ne povečuje več *(42)*.

Z uporabo sekvenciranja druge generacije, kjer se proizvedejo sekvence dolžine do 1.000 baznih parov z visoko kakovostjo, moramo pri metodi sekvenciranja določenega mikrobnega odseka zagotoviti vsaj 30.000–40.000 branj na vzorec, pri kompleksnih mikrobnih združbah pa tudi do $10^5$

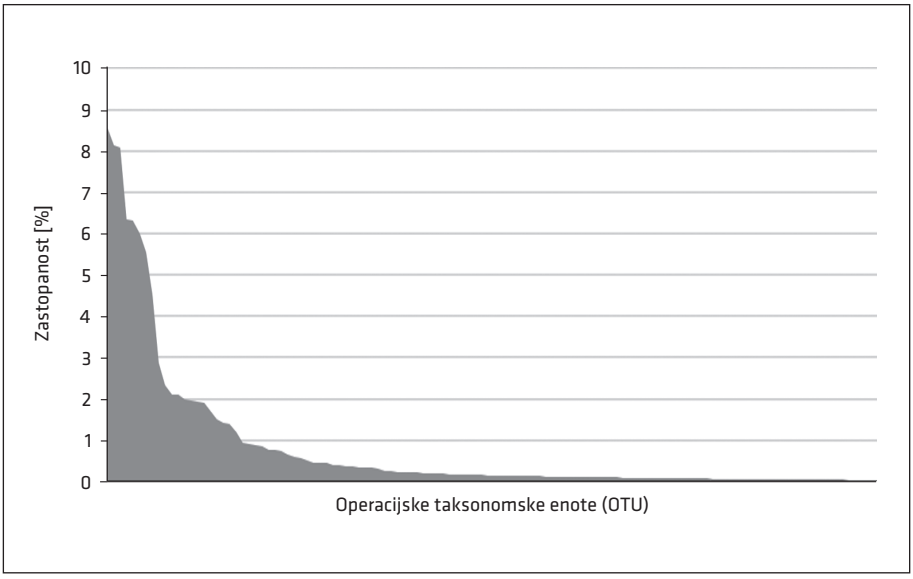



**Slika 1.** Primer rarefakcijskih krivulj iz reanalize bakterijskih mikrobnih združb v blatu preiskovancev. Krivulja ponazarja naraščanje filogenetske raznolikosti ob povečevanju globine sekvenciranja, ki pa ne doseže popolnega platoja. Zunanje črte ponazarjajo 95 % intervale zaupanja *(41)*. OTU – operacijska taksonomska enota (angl. *operational taxonomic unit*).

branj na vzorec. Pri sekvenciranju metagenoma je potrebna globina sekvenciranja vsaj $10^6$ branj na vzorec in metatranskriptoma vsaj $10^7$ branj na vzorec po odstranitvi 16S rRNA in reverzni transkripciji v DNA *(43)*. Med naključno sekvenciranimi odseki genov bo prisotnih veliko branj genov tistih bakterij, ki so v združbi prisotne v velikem številu, pri maloštevilnih bakterijah pa je pokritost slaba. Pokritost sekvenciranja, ki je izračunana kot razmerje med skupno dolžino branj, ki jih uspemo združiti v skupke, in velikosti teh skupkov, pa bo predstavljala podatek o relativni zastopanosti bakterijskih predstavnikov v združbi (slika 2) *(44)*.

## SEKVENCIRANJE 16S RIBOSOMSKE RIBONUKLEINSKE KISLINE ALI RAJE METAGENOMIKA?

Navkljub vsesplošnemu nižanju cen sekvenciranja se z večanjem zahtev po globini sekvenciranja določenega vzorca večajo tudi stroški sekvenciranja, saj lahko z istim orodjem sedaj analiziramo manjše število vzorcev. Pridobljene sekvence mikrobnih filogenetskih označevalcev, izbranih funkcionalnih genov ali celotnega metagenoma je treba nato bioinformatsko in statistično obdelati *(38–40)*. Pri tem je potrebna računska moč za izračun, sorazmerna s količino pridobljenih podatkov oz. z globino sekvenciranja. Z večanjem globine sekvenciranja se ne povečuje le potreba po večji računski moči, ampak je potrebna tudi drugačna arhitektura centralnih procesorskih enot od klasičnih HPCC (angl. *high-performance computing cluster*), saj analize vsebujejo veliko preiskav različnih zaporedij z več podatkovnimi bazami. Pot naprej v tem primeru pomeni najem prostora in kapacitet za izračune v oblaku, s čimer pa se odprejo težave, ko gre za ohranjanje anonimnosti podatkov o preiskovanih bolnikih.




**Slika 2.** Krivulja porazdelitve mikroorganizmov v posamezne skupine v vzorcu iztrebkov (število vzorcev = 340). Graf prikazuje majhno število visoko zastopanih skupin, srednje število srednje zastopanih skupin ter dolg rep nizko zastopanih skupin v vzorcu (angl. *species abundance curve*). Z večanjem resolucije proti 97 % operacijskim taksonomskim enotam se delež sekvenc vzorca, ki predstavljajo določeno operacijsko taksonomsko enoto, zmanjša na nivo, manjši od 0,1 %. OTU – operacijska taksonomska enota (angl. *operational taxonomic unit*).

## Sekvenciranje pomnožkov 16S ribosomske ribonukleinske kisline

Pri določanju bazne sestave mikrobnega odseka določenega gena največkrat določamo sestavo genov za 16S rRNA, npr. bakterij, lahko pa tudi drugih (arheje, glive, protozoji, posamezni funkcionalni geni). Prvotna obdelava pridobljenih sekvenc je v grobem sestavljena iz naslednjih sklopov.

Najprej sekvence z uporabo filtra očistimo molekularnih signalov pod mejo zahtevane kakovosti, poravnamo na referenčno poravnavo in odstranimo nebakterijske (arhejske, mitohondrijske, kloroplastne in evkariontske) sekvence ter umetne konstrukte DNA, ki so nastali kot posledica napak med potekom sekvenciranja. Težave pri razlagi rezultatov predstavljajo obdelava dobljenih branj in odprava ter popravljanje sistemskih napak, ne da bi odstranili tudi signale, ki pripadajo resničnim sekvencam. Brez pravilne obdelave lahko dobimo precenjene podatke o raznolikosti mikrobiote v vzorcu in napačno združevanje v skupke ter napačno klasifikacijo. Temu se želimo izogniti, zato *(45)*:

- Uporabimo filter za odstranitev napak sekvenciranja, s katerim odstranimo vse sekvence, ki niso zadoščale kriterijem. To so minimalna dolžina sekvence, maksimalna dolžina sekvence, maksimalna dolžina sekvenc z identičnimi bazami ter število sekvenc, kjer je bila med potekom sekvenciranja katera od štirih baz (adenin, timin, gvanin, citozin) dvoumno določena (angl. *ambiguous base calls*) ali pa je bil signal pod zahtevano mejo jakosti in tako program ne more jasno določiti, katera baza se nahaja na dotičnem mestu.
- Z uporabo programa (npr. UCHIME) odstranimo umetne konstrukte molekul DNA, t. i. himere, ki med potekom sekvenciranja nastanejo kot posledica združitve različnih sekvenc.
- Prepoznamo ter odstranimo sekvence, ki se niso uvrstile v preiskovano deblo mikroorganizmov, npr. med bakterije, ali so se uvrstile kot arheje, evkarionti, kloroplasti ali pa mitohondriji.

Podatke sekvenciranja po čiščenju lahko nadalje analiziramo tudi tako, da sekvence razdelimo v operacijske taksonomske enote (angl. *operational taxonomic unit,* OTU) različnih taksonomskih resolucij (od domene do vrste, 97 % OTU). Do teh pridemo na tri načine: na osnovi podobnosti zaporedja (razponi 80–97 % identičnost sekvenc), glede na taksonomsko uvrstitev v definirane kategorije sprejete taksonomije, npr. NCBI (National Center for Biotechnology Information), Bergey, UNITE (User-friendly Nordic ITS Ectomycorrhiza) ter glede na filogenetsko drevo. Prva je filogenetsko neodvisna ali kar metoda OTU, druga je odvisna od reprezentativnih sekvenc, uvrščenih v določeno taksonomsko skupino, tretja pa je od filogenije odvisna ali kar filogenetska metoda *(46–48)*. Prednost pristopa, ki temelji na OTU, je, da ni taksonomske pristranskosti, vendar je računsko zelo intenzivna in podvržena velikim vplivom neodstranjenih napak iz sekvenciranja. Medtem ko je filogenetski pristop uporabnejši za preučevanje podobnosti in razlik v združbi, je jasno, da je računska stopnja generiranja ogromnih filogenetskih dreves ena najbolj vprašljivih in približnih *(47)*. Kljub temu da še ni sprejetega soglasja, katera metoda je boljša, trenutna slika v objavah kaže, da večina raziskovalcev ubira srednjo pot in v analizah pomnožkov uporablja taksonomsko klasifikacijo v obstoječe skupine. Težava, ki se pojavi, je, da je taksonomija konstrukt človeške klasifikacije in ne odraža stanja v naravi, saj število kategorij eksponentno narašča z večanjem resolucije proti ravni vrste, medtem ko na ravni sevov odstopa (slika 3). Prav tako znotraj taksonomskega nivoja niso vsi taksoni (taksonomski nivoji) definirani na enak način, ampak so nekatere skupine filigransko razdelane, čeprav med njimi na nivoju 16S rRNA ni velikih razlik.

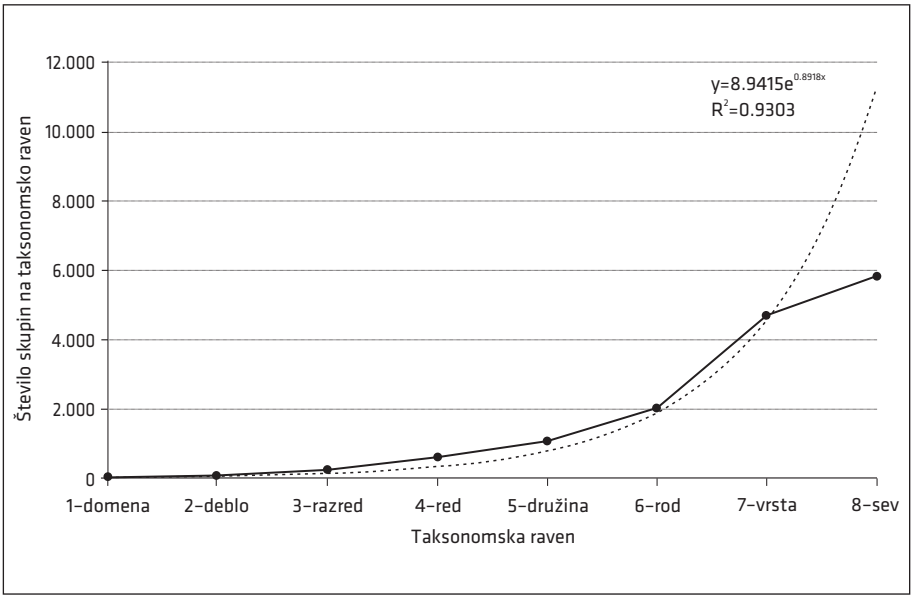
Prosto dostopni programski platformi, ki nam omogočata izvedbo zgoraj omenjenih analiz, sta QIIME in mothur *(46, 49)*. Razlika med platformama je predvsem v tem, da QIIME primarno temelji na filogenetski metodi, mothur pa primarno na metodi OTU, vendar omogoča vse tri (taksonomsko, filogenetsko, OTU). Obe platformi omogočata analizo raznolikosti mikrobiote in, pomembneje, omogočata izvedbo analize, s katero poskušamo pojasniti, kako so člani združbe medsebojno povezani. Razlikujeta se predvsem po tem, do kakšne mere lahko uporabnik nadzira posamezne stopnje analiz. QIIME izvaja svoje analize v programskem jeziku python, medtem ko je mothur napisan v preglednem programskem jeziku C++, kjer uporabnik sam sprejema odločitve in nastavlja parametre posameznih stopenj analiz, za kar pa je potrebnega več znanja in izkušenj.

Za prepoznavanje in številčno določanje zastopanosti posameznih taksonov v preiskovani mikrobni združbi lahko v prvem koraku očiščene sekvence sedaj razvrstimo z razvrščevalcem rRNA. Primer je naivni Bayesov razvrščevalec rRNA, ki je del baze RDP *(50, 51)*. V uporabi so še ostale baze podatkov: Greengenes, SILVA in NCBI, med katerimi se stopnja ažuriranosti razlikuje *(52–54)*. Prav slednje se pogosteje izpostavlja kot veliko težavo, predvsem pri podatkovni bazi Greengenes, ki je bila zadnjikrat osvežena maja 2013, in tako predstavlja precej zastarelo sliko taksonomskih razvrstitev.

Ne nazadnje, na enak način, kot analiziramo sekvence 16S ali 18S rRNA ali ITS, lahko analiziramo tudi sekvence, pridobljene iz globokega sekvenciranja funkcionalnih genov *(38)*.

## Sekvenciranje metagenomov

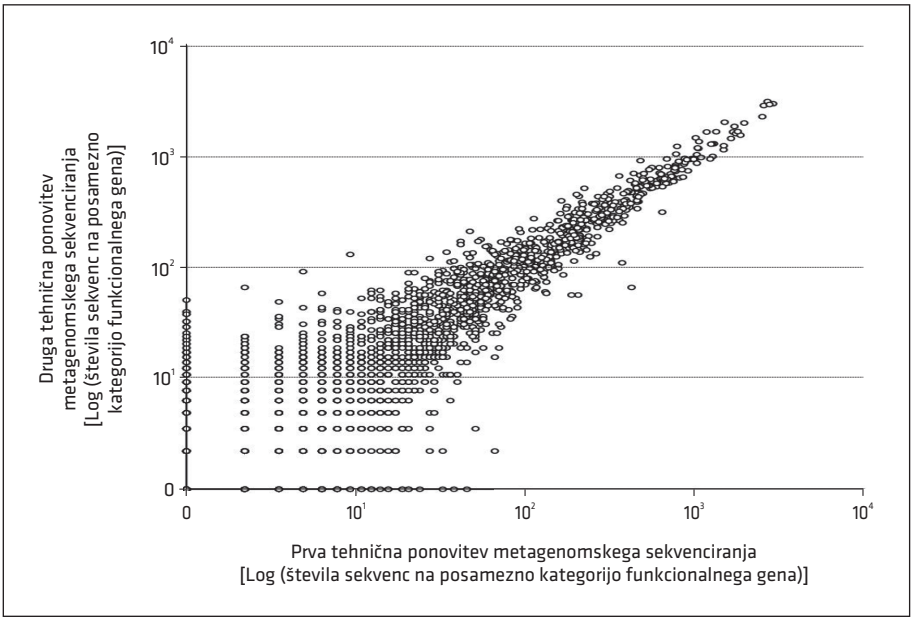
Določanje zaporedji genov 16S rRNA zagotavlja veliko podatkov o preiskovani mikrobni združbi, vendar je glavna težava v tem, da je le del gena za 16S rRNA pomnožen in sekvenciran. Z razvojem tehnologij sek-




**Slika 3.** Odnos med številom skupin na določeni taksonomski ravni v odvisnosti od izbire taksonomske ravni. Raven seva odstopa od splošnega eksponencialnega naraščanja števila taksonov v posamezni kategoriji. $R^2$ – regresijski koeficient.

venciranja se je povečala količina pridobljenih podatkov in zmanjšali so se stroški izvedbe take analize. Zaradi tega čedalje več raziskav vključuje sekvenciranje celotnega metagenoma, torej ugotavljamo nukleotidno zaporedje fragmentov mikrobnih genov, ki so zbrani v skupku vseh genomov preiskovane mikrobne združbe. Takšno sekvenciranje omogoča anotacijo večine najbolj zastopanih mikrobnih genov znotraj vzorca, od katerih lahko večji del predstavljajo gospodinjski geni mikroorganizmov, ki so nujni za preživetje in so kot takšni ohranjeni pri večini mikroorganizmov. Z izvedbo sekvenciranja metagenoma tako opredelimo tudi funkcionalno vlogo mikrobne združbe v določenem ekosistemu in izvemo, ali so spremembe v sestavi mikrobne združbe vzrok ali posledica npr. določenega bolezenskega stanja *(42, 55)*. Podatki metagenomskih raziskav so pridobljeni v velikih količinah in zelo razdrobljeni, kar predstavlja velik izziv pri iskanju, zbiranju in predelovanju koristnih bioloških podatkov *(56, 57)*. Primer je sekvenciranje metagenoma človeškega črevesnega mikrobioma, kjer je bilo prepoznanih 3.300.000 genov, sestavljenih iz 567,7 giga baznih parov sekvenciranih podatkov *(4)*. Dodaten vpogled v delovanje in domet metagenomskih analiz nam pokaže tudi enostavna povezanost med tehničnima ponovitvama sekvenciranja istega vzorca, ki pokaže, da je za kakovostne analize potrebnih precej več sekvenc na posamezno kategorijo zaznanega funkcionalnega gena (slika 4). Pri nizkem številu sekvenc na kategorijo gena veliko skupin genov namreč zaznamo v eni od ponovitev sekvenciranja, v drugi pa sploh ne. To je pokazatelj, da se limite zaznavanja in kvantifikacije sekvenciranja genov iz določenih kategorij funkcionalnih genov med seboj razlikujejo in jih v praksi ne znamo dobro nadzorovati.

Metagenomska analiza je sestavljena iz več korakov. Po sekvenciranju je treba pridobljene odčitke najprej očistiti in nato zložiti v večje skupke (angl. *contigs*). Odvisno




**Slika 4.** Povezanost med številom sekvenc na posamezno kategorijo funkcionalnega gena med dvema tehničnima ponovitvama sekvenciranja istega vzorca.

od naših podatkov sestavljanje opravimo *de novo* ali glede na referenco. Slednje je pri analizah kompleksnih metagenomov težavno, saj pri sestavljanju metagenomov praviloma ne vemo, katerim mikroorganizmom sekvence pripadajo. Zato se načeloma uporabljajo od reference neodvisni pristopi. Skupke genov grupiramo v »draft genome«, napovemo kodirajoče in nekodirajoče gene ter opravimo funkcionalno anotacijo teh s primerjavo s podatkovnimi bazami. Na koncu sledi statistična obdelava podatkov in razlaga rezultatov *(58)*. Izbira programov za izvedbo analize metagenomskega sekvenciranja je sedaj še vedno v fazi preverjanja sistemskega vpliva glede na izbiro določenega programa, s skupnim ciljem zmanjševanja algoritemskega šuma in povečevanja standardizacije protokolov za analizo podatkov, kar bi nas v idealnem primeru pripeljalo do primerljivosti podatkov med različnimi raziskavami, ne glede na to, kateri program oz. protokol je bil uporabljen pri analizi.

Pred začetkom bioinformacijske analize moramo najprej preveriti kakovost dobljenih sekvenc. Podatek o tem je kodiran v FASTQ formatu sekvenc, ki nam poleg bazne sestave pove tudi verjetnost, da je določena baza na določenem mestu v zaporedju napačna. Orodja, ki omogočajo analizo in prikaz razporeditev napak teh verjetnosti, so NGS QC Toolkit, Kraken in HTQC *(59–61)*. V koraku filtriranja sekvenc odstranimo dvojnike sekvenc, artefakte in nukleotidne baze z nizko kakovostjo ter gostiteljeve sekvence. Gostiteljevo DNA prepoznavamo z orodjem, ki poišče ujemanje z gostiteljevo DNA na podlagi referenčnega genoma. Primera orodij za odstranjevanje evkariontskih genomskih zaporedij DNA sta Eu-Detect in DeConseq *(62, 63)*.

V naslednjem koraku očiščene sekvence združimo v skupke. Z vidika računske moči ta stopnja predstavlja ozko grlo analize in ni popolnoma zanesljiva, saj so metagenomski podatki zelo zgoščeni in kratkih dolžin, kar pa predstavlja visoko verjetnost za napako. Obstaja več programov za sestavljanje odčitkov. Programi, kot sta Velvet assembler in SOAPdenovo, so bili optimizirani za krajše odčitke, ki jih proizvaja sekvenciranje naslednje generacije z uporabo grafov De Bruijn *(64)*. Pri sestavljanju težave povzročajo ponavljajoče se sekvence DNA, predvsem zaradi razlike v relativni številčnosti vrst, ki so prisotne v vzorcu, in nastanek himernih skupkov, kjer pride do sestavljanja odčitkov iz več kot ene mikrobne vrste *(65)*. Pri sestavljanju si lahko pomagamo z uporabo referenčnih genomov, ki omogočajo sestavljanje čedalje večjega števila mikrobnih vrst, saj narašča število mikrobnih debel, za katere so sekvencirani genomi dostopni. Kljub temu težavo v tem koraku lahko predstavljajo pomanjkljivo sestavljeni referenčni genomi, katerih anotacije se skozi čas spreminjajo, zato se preferenčno v praksi uporabljajo programi, ki temeljijo na analizi De Bruijnovih grafov *(65, 66)*.

Sestavljanju sledi napoved genov oz. označitev kodirajočih regij v sestavljenih skupkih in anotacija. Napoved genov lahko opravimo na dva načina *(65)*. Prepoznava genov v prvem pristopu temelji na homologiji z geni, katerih zaporedja so že javno dostopna v podatkovnih bazah. Tako pregledamo funkcionalne motive in celične lokacijske signale za napovedane beljakovinske sekvence z orodji, kot npr. PRIAM za encimsko uvrščanje, HMM-Pfam in TIGRFAM za funkcionalne motive, TMHMM za prepoznavanje transmembranskih potencialnih domen, ter z uporabo brskalnika BLAST, kjer za referenco uporabimo nukleotidno in beljakovinsko bazo PANDA *(67–70)*. To vrsto pristopa uporablja program MEGAN *(71, 72)*. Pri drugem pristopu se uporabljajo bistvene značilnosti zaporedja za napoved kodirajočih regij, ki temeljijo na sklopu genov iz sorodnih organizmov. Takšen pristop je implementiran v programih, kot sta GeneMark in GLIMMER *(73)*. Glavna prednost

take napovedi je, da omogoča odkrivanje kodirajočih regij, ki nimajo homolognih zaporedij v podatkovnih bazah. Najbolj natančno napovedovanje pa opravimo, ko imamo na voljo velike regije sosednje genomske DNA za primerjavo *(57)*. Vseeno pa ni nič nenavadnega, če delež neanotiranih zaporedij v metagenomu predstavlja tudi 60 % vseh sekvenc tega vzorca. Tudi podatkovne baze znanih beljakovin se neprestano razvijajo in dopolnjujejo, kar kaže, da je posodabljanje podatkovnih baz nova prioriteta pri analizah metagenomskih podatkov na strežnikih velikih zmogljivosti.

Da bi povezali raznolikost mikrobne združbe in njeno funkcijo v metagenomih, je treba sekvence spojiti. Spajanje (angl. *binning*) je proces povezovanja sekvenc z določenim organizmom *(65)*. Pri spajanju, ki temelji na podobnosti, z metodo BLAST iščemo filogenetske označevalce ali podobna zaporedja v obstoječih podatkovnih bazah. Rezultate iz baze BLAST uvozimo neposredno v program MEGAN, kjer sekvence z uporabo baze podatkov NCBI taksonomsko uvrstimo in nato opravimo še funkcionalno analizo in napoved predvidenih presnovnih poti z uporabo sistema razvrščanja SEED ali KEGG *(71, 72, 74, 75)*. V uporabi je tudi orodje PhymmBL, ki uporablja interpolirane Markove modele za dodelitev ali nakazovanje vloge branj sekvenciranja *(57)*. MetaPhlAn in AMPHORA sta dve metodi, ki za ocenjevanje relativne obilnosti organizmov uporabljata edinstvene označevalce taksonomskih skupin *(76)*. Ko so sekvence spojene, je mogoče izvesti primerjalno analizo.

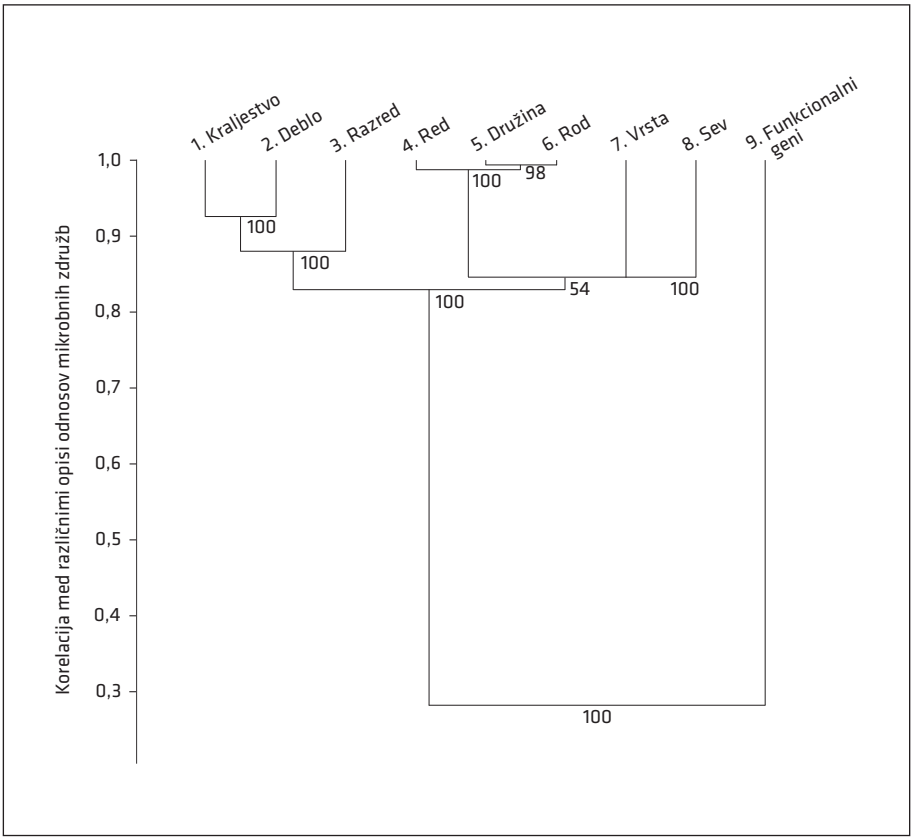
Ogromna količina eksponentno rastočih metagenomskih podatkov in pripadajočih metapodatkov predstavlja izjemen izziv, ki pa ima zelo velik potencial za razumevanje medsebojnega delovanja mikrobov ter medsebojnega delovanja med mikrobi in gostiteljem na različnih nivojih, ki jih pokriva sistemska medicina. Metapodatki vključujejo podatke o poskusu, preiskovancih, kemijske/okoljske lastnosti vzorca, fizične podatke o mestu in metodologiji vzorčenja itd. in so nujno potrebni, tako za zagotavljanje ponovljivosti analize kot tudi za njeno uspešnost. Primer orodja za združitev zaporedij DNA in metapodatkov, ki omogočajo primerjalne analize različnih naborov podatkov z uporabo ekoloških indeksov, je MG-RAST *(77)*. Danes, 10 let od ustanovitve, strežnik vsebuje 135,08 tera baznih parov oz. 286.340 metagenomov s 1.024 milijardami sekvenc. Server omogoča tudi anotacijo, izračun taksonomske porazdelitve vrst, številčnosti in $\alpha$-raznolikosti ter sestave napovedanih kodirajočih regij genov v funkcionalne kategorije in podsisteme. Prav tako vsebuje orodja za primerjavo in vizualizacijo podatkov tako iz lastnih kot tudi iz prej naloženih baz podatkov. Zato taka integracija računskih orodij s shranjevanjem podatkov prestavlja infrastrukturo za metagenomiko in metatranskriptomiko na nivoju urejenosti, kot jo je včasih pomenila uporaba NCBI GeneBank za analize posameznih sekvenc, le da je sedaj raven kompleksnosti nekaj stopenj višja. Zaradi poenotene analize so rezultati vsebinsko bolj primerljivi med različnimi raziskavami, saj je analitski šum občutno zmanjšan. Sistem IMG/M tudi zagotavlja zbirko orodij za funkcionalno analizo mikrobnih združb, ki temelji na njihovem metagenomskem zaporedju ter na genomskem zaporedju referenčnih izolatov, ki so vključeni v sistem IMG ter v projekt GEBA (Genomic Encyclopedia of Bacteria and Archaea) *(78)*. Zaporedje ukazov in orodij za anotacijo posameznih branj ali pa skupkov branj, pridobljenih z metagenomskim sekvenciranjem, predlaga tudi Inštitut J. Craig Venter *(79)*. Tako postajajo orodja, ki so bila namenjena hkratnim analizam taksonomskega in funkcionalnega opisa mikrobiote v drugih okoljih posredno vir informacij in podatkov za analize, ki vključujejo podatke z več nivojev človeškega telesa, torej sistemske in bolj personalizirane medicine, ki se analizirajo v sklopu projektov,

kot je npr. CA15120 – Open Multiscale Systems Medicine *(80)*. Hkrati nam za razvoj novih pristopov na področju mikrofluidike služi tudi PortASAP *(81)*.

Seveda pa ravno uporaba istih podatkov na različnih taksonomskih nivojih vodi v začetno zmedo, ko uporabniki še ne sprejemajo dejstva, da analize istih podatkov na različnih nivojih taksonomske ali funkcionalne resolucije lahko generirajo zelo različne odgovore in odnose med vzorci kot tudi med preiskovanimi skupinami, ki jim ti vzorci pripadajo (slika 5).

## Statistična analiza

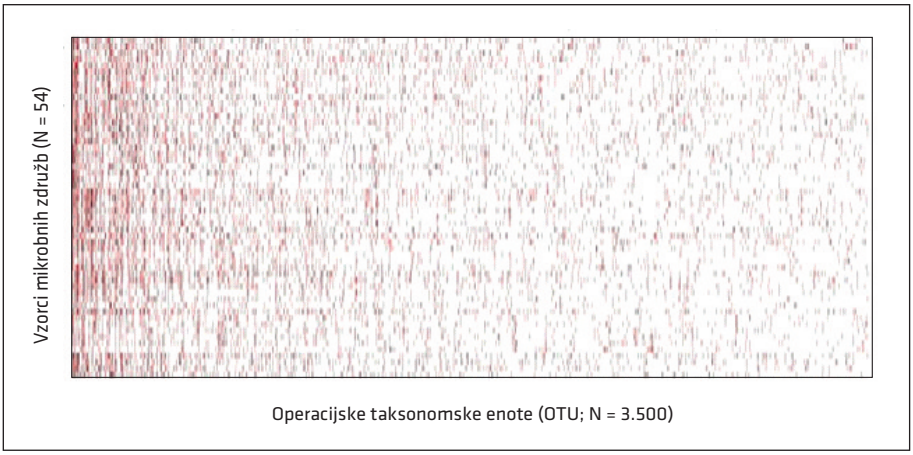
Pri preučevanju raznolikosti mikrobiote želimo izvedeti, koliko različnih članov je v posamezni združbi (α-raznolikost). To lahko predstavimo na grafu z rarefakcijsko krivuljo, kjer je število v vzorcu odkritih taksonov funkcija števila sekvenc (slika 2). Na ta način opišemo bogatost združbe in preverimo, ali smo zajeli celotno raznolikost vzorca. V tem primeru krivulja doseže plato, kar pomeni, da se z večanjem števila sekvenc število taksonov ne veča več. Kazalca α-raznolikosti sta tudi Chao1, ki s pomočjo ekstrapolacije izračunava vrstno bogatost,



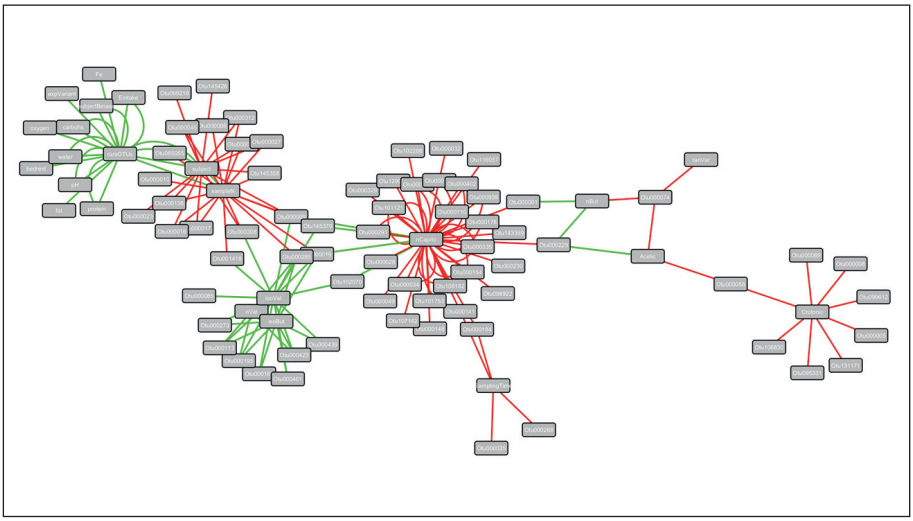

**Slika 5.** Povezava med opisi odnosov mikrobnih združb 40 metagenomov prebavnega trakta. Metagenome smo analizirali na različnih taksonomskih in funkcionalnih nivojih in na vsakem nivoju izračunali distančne matrike razdalj med posameznimi vzorci. Dendrogram prikazuje korelacije med temi matrikami, kako se z večanjem taksonomske resolucije spreminjajo odnosi med vzorci znotraj posameznega nivoja, predvsem pa ilustrira nizko povezavo med taksonomskim in funkcionalnim nivojem obravnavanih metagenomov. Številke na vejiščih kažejo podporo vejanja (angl. *bootstrap*).

in Simpson, ki odraža število taksonov, kar opredelimo z izrazom bogatost združbe, ter relativno zastopanost taksonov v vzorcu, kar opredelimo z izrazom enakomernost združbe *(82–84)*. Težava pri analizi so velike asimetrične matrike (majhno število vzorcev in veliko število kategorij mikroorganizmov) z velikim številom ničel (slika 6).

Za opis razlik ali podobnosti med združbami (β-raznolikost) uporabimo kazalce podobnosti, s katerimi izračunamo matriko razdalj, to pa nato uporabimo za grafični prikaz uvrščanja združb v skupine. Pri tem uporabljamo metode, ki za izračun matrike razdalj in testiranje značilnosti razlik uporabljajo evolucijsko povezanost zaporedij, npr. metoda UniFrac, ter kazalce raznolikosti kot npr. Bray-Curtis, Morisita-Horn in Sörensen *(47, 83)*. Nadalje s testom HOMOVA testiramo, ali je genetska pestrost med dvema združbama homogena in s testom analize molekularne variance AMOVA preverimo, ali je genetsko odvisna pestrost med dvema ali več mikrobnimi združbami različna od pestrosti vseh združb skupaj *(38–40, 84–86)*. Kadar primerjamo več skupin vzorcev, moramo opraviti tudi popravek večkratnih primerjav *(87, 88)*. Prav tako lahko s pristopi strojnega učenja, kot so umetne nevronske mreže, naključni gozd (angl. *random forest*), testi Lefse, Metastats in indikatorske vrste prepoznamo taksonomske ali funkcionalne skupine, ki se statistično značilno razlikujejo med vzorci *(38–40, 89–91)*. S testom AWKS lahko testiramo prisotnost taksonomskih ali funkcionalnih skupin v različnih vzorcih in to grafično predstavimo na osnovi podatkov o prisotnosti, številčnosti in pogostosti preiskovanih skupin v vzorcih *(39, 92)*. Za grafični prikaz raznolikosti lahko uporabimo metodo glavnih komponent ali pa nemetrično večdimenzionalno lestvičenje, ki ga lahko med drugim izvedemo s paketom vegan v programskem okolju R *(93)*. Ravno tako lahko razširimo naše analize na analize metabolnih mrež in mrež sopojavnosti mikroorganizmov ali funkcionalnih genov z določenimi lastnostmi gostitelja ali njegovega okolja (slika 7).




**Slika 6.** Pregled zastopanosti in razporeditve prisotnosti in odsotnosti sekvenc po prvih 3.500 najbolj zastopanih operativnih taksonomskih enotah v 54 vzorcih. Najbolj zastopane skupine, ki so prisotne v večini vzorcev z največjim številom, so obarvane rdeče (levo), čedalje manj zastopane skupine so obarvane sivo, bela polja prikazujejo odsotnost določene operativne taksonomske enote v določenem vzorcu. Sama podolgovata oblika matrike z naraščajočim deležem belih lis v desno smer najbolj prikaže omenjeno numerično asimetrijo obravnavanih matrik podatkov. Za primerjavo glej sliko 2. OTU – operacijska taksonomska enota (angl. *operational taxonomic unit*).



**Slika 7.** Primer metabolne mreže mikrobnih operativnih taksonomskih enot s parametri mikrobnega okolja v prebavnem traktu. Z rdečo in zeleno barvo so označene statistično značilne negativne in pozitivne povezave med metaboliti v okolju ter mikrobnimi operativnimi taksonomskimi enotami. S poglobljenimi analizami kompleksnih grafov lahko ugotovimo, kateri mikroorganizmi ali njihovi funkcionalni geni so ključni pri medsebojnih povezavah znotraj mikrobnih združb, kateri okoljski parametri ključno vplivajo (pozitivno ali negativno) na določene mikrobne skupine ter kateri parametri so ključni za razlikovanje med posameznimi skupinami preiskovancev. OTU – operacijska taksonomska enota (angl. *operational taxonomic unit*).

Pri celotni analizi podatkov se je treba zavedati omejitev, ki izvirajo iz *(92)*:

- narave podatkov,
- tehnološke (ne)ponovljivosti sekvenciranja, povezane z najmanjšimi preiskovanimi skupinami, ki jih še lahko natančno določimo z uporabljenim analitskim pristopom in so opisane z:
  - limito sposobnosti zaznave,
  - limito sposobnosti kvantifikacije,
  - limito zaznavanja praznih skupin in
  - limito linearnega odziva za posamezno taksonomsko ali funkcionalno skupino,
- neposredne neprimerljivosti postopkov in približne ocenitve posameznih algoritmov ter
- razvoja statističnih metod, ki zagotavljajo določeno statistično moč raziskav.

Narava podatkov nam kaže presenetljivo ugotovitev, da imamo lahko na nivoju taksonomskih analiz velike razlike med proučevanimi skupinami preiskovancev, medtem ko mikrobne združbe na nivoju funkcionalnih genov vsebujejo praktično identične funkcionalne gene *(95)*. Prav tako ugotovimo, da so rutinske velikosti kohort, ki se uporabljajo v medicini za mikrobiološke raziskave z uporabo sekvenciranja tarčnih filogenetskih genov v splošnem premajhne. Za to, da bi lahko ločili centroidi oz. središčni točki opisanih mikrobnih združb med zdravimi in debelimi preiskovanci za 1 %, bi morali v vsaki skupini uporabiti več kot 2.000 preiskovancev *(96)*. Zato uporaba le nekaj pet ali deset preiskovancev z zelo različnimi prehranskimi, življenjskimi navadami, cirkadianimi ritmi, socialnimi nivoji in osebno zgodovino, kot je pogosto zaslediti v medicinski literaturi, enostavno ne zadostuje za poglobljene raziskave, hkrati pa kaže, da nekateri parametri, ki jih spremljajo, morda niti nimajo tako obsežnega vpliva. Tako se npr. kaže, da so se spremembe v človeški fiziologiji zgodile dva do tri tedne prej, preden je prišlo do konsistent-

nih sprememb v strukturi mikrobnih združb ali metagenomskem funkcionalnem opisu mikrobnih združb, hkrati pa je bilo na metabolomskem nivoju možno zaznati spremembe mikrobne presnove veliko prej, torej sočasno s spremembami v človeški fiziologiji *(38–40)*. Iz tega bi lahko izpeljali enostaven predlog, da je pri kompleksnih bolezenskih stanjih pomembneje, kaj mikroorganizmi počnejo na metabolnem nivoju, kot pa njihova taksonomska uvrstitev.

Uporaba začetnih oligonukleotidov pri globokem sekvenciranju je tudi povzročila, da dolgo časa ni bilo nobenega napredka pri odkrivanju še preostalih mikrobnih skupin. Namesto velikega števila večinoma izmišljenih ocen raznolikosti mikrobnih skupin iz sekvenciranja pomnožkov je ravno metagenomika omogočila odkrivanje celih družin bakterij in arhej, ki jih do takrat ni bilo možno opisati, saj se vezavna mesta za začetne oligonukleotide v PCR enostavno razlikujejo od znanih vezavnih mest in jih torej nismo mogli zaznati s PCR *(95)*. Ravno kombinacija globokega sekvenciranja celotnih metagenomov nam danes omogoča rekonstrukcijo »draft genomov« mikroorganizmov brez gojenja in njihovo evolucijsko analizo v kontekstu do sedaj znanih genomov, razvoja otokov znotraj genomov, horizontalnih prenosov ter regulacije ekspresije in presnove.

Razvoj algoritmov je pripeljal celotno skupnost do točke zavedanja raznolikosti v rezultatih, ki je posledica uporabe različnih algoritmov. Ni presenetljivo, da potekajo celotne raziskave, znotraj katerih primerjajo različne algoritme z istimi podatki, z namenom standardizacije analitskih poti in izborom algoritmov, ki dajejo najboljše rezultate *(96, 97)*.

Razvoj statističnih metod kaže, da živimo v izredno zanimivih časih, ko se tehnološki razvoj iz analitske kemije in mikrofluidike začne odražati tudi na količini podatkov v mikrobiologiji, s čimer so povezani tudi zelo pomembni preboji pri uporabi statističnih metod pri analizi tovrstnih podatkov. Tipi izvedenih analiz se sedaj premikajo s področja GWAS (angl. *genome wide association studies*) na področja, kjer se metagenomski, metabolomski in genomski podatki integrirajo znotraj skupine ter med skupinami podatkov. Zato mikrobiologija kot taka danes predstavlja le eno izmed ravni analize kompleksnih sistemov, kot je človek. V takem sistemu medsebojno reagirajo genom, transkriptom, proteom in metabolom človeka, z metagenomi, metatranskriptomi, metaproteomi in metametabolomi mikrobnih združb ter njihovimi kompleksnimi ekstracelularnimi vezikli, zunajcelično DNA in mikro RNA. Odzivi sistema se spreminjajo skozi prostor in čas, v odvisnosti od aktivnosti, prehrane, cirkadianega ritma in drugih lastnosti gostitelja.

Zaradi kompleksnosti analiz nam napredek v tem trenutku omogoča le pristop od zgoraj navzdol, v katerem spremljamo množice podatkov na vseh teh nivojih ter jih integriramo v skupen model celotnega sistema *(38–40, 98–100)*. Potrditveni faktorialni pristopi od spodaj navzgor pa so izvedeni v nadzorovanih eksperimentih, kjer množice prej spremenljivih parametrov lahko nadzorujemo in ohranjamo nespremenljive ter izluščimo vpliv posameznih spremenljivk, prepoznanih v sklopu pristopa od zgoraj navzdol *(101)*.

## ZAKLJUČEK

Med metodami za opredeljevanje sestave mikrobne združbe v prebavnem traktu človeka najbolj zanesljive in povedne podatke ponuja sekvenciranje. Med različnimi izvedbami sekvenciranja se najpogosteje uporablja sekvenciranje informativnih delov mikrobnih genomov. Za uspešno in hitro izvedbo analize so med drugimi ključni dejavniki: velikost vzorca, ustrezno ravnanje z vzorci, uporaba kontrol, onesnaženje z gostiteljevo DNA ali DNA drugih vzorcev ali med potjo odvzema, ekvimolarna razdelitev količine DNA pri več vzorcih in nazadnje

dovolj velika računska moč za izvedbo bioinformatskih in statističnih analiz. Kot slabost se pri preučevanju večjega števila vzorcev pokaže velik finančni vložek, ki je potreben za izvedbo analize. V primeru, da velik finančni zalogaj ni omejitev in da pričakovano onesnaženje z gostiteljevo DNA ni preveliko, potem je smiselno na izbranih vzorcih izvesti sekvenciranje metagenomov. Tako pridemo do dodatnih taksonomskih in funkcionalnih napovedi. Orodij, ki omogočajo obdelavo metagenomskih podatkov, je veliko, saj je dotično področje izredno aktivno in se hitro razvija. Kljub temu do sedaj še ni razvitega programa, ki bi omogočal poenoteno analizo podatkov sekvenciranja celotnega genoma. Slednje pogosto privede do tega, da je dobljene rezultate v različnih raziskavah zelo težko primerjati med sabo.

Zaradi nepredstavljivega razvoja tehnik, pristopov, statistike ter ne nazadnje načina razmišljanja zato niti ni več presenetljivo, da se v skupnosti pojavlja občutek, da tovrstne analize niso več del mikrobiologije. To je v bistvu napačen okvir razmišljanja, ki predvsem zavrača spremembe na področju mikrobiologije. Če imamo pred očmi težave Kocha in Pasteurja pri uveljavljanju znanstvenega mikrobiološkega načina razmišljanja s percepcijami kemikov, fizikov in medicincev takratnega časa, se ne moremo izogniti spoznanju, da je mikrobiologija tehnološko podprta in poganjana veda, ki se plemeniti s prestopanjem psevdotradicionalnih okvirjev gojenja, z razvojem lastnih ter s sprejemanjem idej z drugih področij znanosti. Na enak način je v preteklosti sprejela mikroskopijo od astronomije ter DGGE iz medicinske analize mutacij, tako mikrofluidiko iz kemije, strojno krmiljenje iz elektrotehnike, robote iz strojništva, bioinformatiko in HPCC od računalništva ter neparametrično vejo analiz iz statistike. Šele skupek znanj z drugih področij je tako omogočil premik mikrobiologije naprej s področij, kjer je koristno delovala preteklo stoletje, na področja, kjer je njena pomoč še toliko bolj potrebna: sladkorna bolezen tipa 2, astma, prezgodnji porod, alergije, kronična obstruktivna pljučna bolezen, debelost, presnovni sindrom, depresija itd. Prav povezava znanj z različnih področij, ki obravnavajo različne velikostne razrede, pa ponuja možnosti za bolj ciljane analize posameznih primerov ter njihovo bolj individualno obravnavo. Na ta področja vstopajo nove tehnologije s področja sekvenciranja beljakovin, oligosaharidov, maščob in drugih kompleksnih molekul.

## LITERATURA


1. Hooper LV, Gordon JI. Commensal host-bacterial relationships in the gut. Science. 2001; 292 (5519): 1115–8.
2. Tjalsma H, Boleij A, Marchesi JR, et al. A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. Nat Rev Microbiol. 2012; 10 (8): 575–82.
3. Turnbaugh PJ, Ley RE, Hamady M, et al. The human microbiome project. Nature. 2007; 449 (7164): 804–10.
4. Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010; 464 (7285): 59–65.
5. Eckburg PB, Bik EM, Bernstein CN, et al. Diversity of the human intestinal microbial flora. Science. 2005; 308 (5728): 1635–8.
6. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. Microbiol Rev. 1995; 59 (1): 143–69.
7. Fraher MH, O'Toole PW, Quigley EM. Techniques used to characterize the gut microbiota: a guide for the clinician. Nat Rev Gastroenterol Hepatol. 2012; 9 (6): 312–22.
8. Kolbert CP, Persing DH. Ribosomal DNA sequencing as a tool for identification of bacterial pathogens. Curr Opin Microbiol. 1999; 2 (3): 299–305.
9. Fischer SG, Lerman LS. DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: correspondence with melting theory. Proc Natl Acad Sci USA. 1983; 80 (6): 1579–83.
10. Muyzer G, Smalla K. Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. Antonie Van Leeuwenhoek. 1998; 73 (1): 127–41.
11. Osborn AM, Moore ER, Timmis KN. An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. Environ Microbiol. 2000; 2 (1): 39–50.
12. Stres B. The first decade of terminal restriction fragment length polymorphism (T-RFLP) in microbial ecology. Acta Agriculturae Slovenica. 2006; 88 (2): 65–73.
13. Fisher MM, Triplett EW. Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. Appl Environ Microbiol. 1999; 65 (10): 4630–6.
14. Franke-Whittle IH, Manici LM, Insam H, et al. Rhizosphere bacteria and fungi associated with plant growth in soils of three replanted apple orchards. Plant and Soil. 2015; 395 (1–2): 317–33.
15. Palmer C, Bik EM, Eisen MB, et al. Rapid quantitative profiling of complex microbial populations. Nucleic Acids Res. 2006; 34 (1): e5.
16. Novak D, Franke-Whittle IH, Pirc ET, et al. Biotic and abiotic processes contribute to successful anaerobic degradation of cyanide by UASB reactor biomass treating brewery waste water. Water Res. 2013; 47 (11): 3644–53.
17. Carey CM, Kirk JL, Ojha S, et al. Current and future uses of real-time polymerase chain reaction and microarrays in the study of intestinal microbiota, and probiotic use and effectiveness. Can J Microbiol. 2007; 53 (5): 537–50.
18. Henry S, Bru D, Stres B, et al. Quantitative detection of the nosZ gene, encoding nitrous oxide reductase, and comparison of the abundances of 16S rRNA, narG, nirK, and nosZ genes in soils. Appl Environ Microbiol. 2006; 72 (8): 5181–9.
19. Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. Genomics. 2016; 107 (1): 1–8.
20. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009; 55 (4): 641–58.
21. Holley RW, Apgar J, Merril SH, et al. Nucleotide and oligonucleotide compositions of the alanine-, valine-, and tyrosine-acceptor soluble ribonucleic acids of yeast. J Am Chem Soc. 1961; 83 (23): 4861–2.
22. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 1977; 74 (12): 5463–7.
23. NHGRI: The cost of sequencing a human genome [internet]. National Human Genome Research Institute; [citirano 2017 Apr 20]. Dosegljivo na: https://www.genome.gov/sequencingcosts
24. Chaucheyras-Durand F, Ossa F. The rumen microbiome: abundance, diversity, and new investigative tools. The Prof Anim Sci. 2014; 30 (1): 1–12.
25. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in open microfabricated high density picoliter reactors. Nature. 2005; 437 (7057): 376–80.

26. Kawashima EH, Farinelli L, Mayer P. Method of nucleic acid amplification. Google patents [internet]. 2005; 8476044: B2.
27. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008; 9: 387–402.
28. Rusk N. Torrents of sequence. Nature Methods. 2010; 8: 44.
29. Valouev, A, Ichikawa J, Tonthat T, et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res. 2008; 18 (7): 1051–63.
30. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods. 2008; 5 (1): 16–8.
31. Porreca GJ. Genome sequencing on nanoballs. Nature Biotech. 2010; 28 (1): 43–4.
32. Drmanac R, Sparks AB, Callow MJ, et al. Human genome sequencing using unchained base reads in self-assembling DNA nanoarrays. Science. 2010; 327 (5961): 78–81.
33. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010; 19 (2): 227–40.
34. Rhoads A, Au KF. PacBio sequencing and its application. Genomics, Proteomics & Bioinformatics. 2015; 13 (5): 278–89.
35. Pathak B, Lofas H, Prasongkit J, et al. Double-functionalized nanopore-embedded gold electrodes for rapid DNA sequencing. Appl Phys Letters. 2012; 100 (2): 023701.
36. Korlach, J, Marks PJ, Cicero RL, et al. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc Natl Acad Sci USA. 2008; 105 (4): 1176–81.
37. dela Torre R, Larkin J, Singer A, et al. Fabrication and characterization of solid-state nanopore arrays for high-throughput DNA sequencing. Nanotechnology. 2012; 23 (38): 385308.
38. Sket R, Treichel N, Debevec T, et al. Hypoxia and inactivity related physiological changes (constipation, inflammation) are not reflected at the level of gut metabolites and butyrate producing microbial community: The PlanHab study. Front Physiol. 2017; 8: 250.
39. Sket R, Treichel N, Kublik S, et al. Hypoxia and inactivity related physiological changes precede or take place in absence of significant rearrangements in bacterial community structure: The PlanHab randomized pilot trial study. PLoS One. 2017; 12 (12): e0188556.
40. Sket R, Debevec T, Kublik S, et al. Intestinal metagenomes and metabolomes in healthy young males: inactivity and hypoxia generated negative physiological symptoms precede microbial dysbiosis. Front Physiol. 2018; 9: 198.
41. Bajuk J. Metaanaliza podatkov o humani mikrobioti [magistrsko delo]. Ljubljana: Univerza v Ljubljani; 2017.
42. Mende DR, Waller AS, Sunagawa S, et al. Assessment of metagenomic assembly using simulated next generation sequencing data. PLoS One. 2012; 9 (11): e114063.
43. Liu Y, Ferguson JF, Xue C, et al. Evaluating the impact of sequencing depth on transcriptome profiling in human adipose. PLoS One. 2013; 8 (6): 1–10.
44. Gill SR, Pop M, DeBoy R. T, et al. Metagenomic analysis of the human distal gut microbiome. Science. 2006; 312 (5778): 1355–9.
45. Edgar RC, Haas BJ, Clemente JC, et al. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011; 27 (16): 2194–200.
46. Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009; 75 (23): 7537–41.
47. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005; 71 (12): 8228–35.
48. Lozupone C, Lladser ME, Knights D, et al. UniFrac: an effective distance metric for microbial community comparison. ISME J. 2011; 5 (2): 169–72.
49. Kuczynski J, Stombaugh J, Walters WA, et al. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr Protoc Bioinformatics. 2011; 10: 934–5.
50. Wang Q, Garrity GM, Tiedje JM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007; 73 (16): 5261–7.
51. Cole JR, Wang Q, Cardenas E, et al. The Ribosomal Database Project: improved alignments, new tools for rRNA analysis. Nucleic Acids Res. 2009; 37 (Database issue): 141–5.
52. McDonald D, Price MN, Goodrich J, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012; 6 (3): 610–8.

53. Pruesse E, Quast C, Knittel K, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 2007; 35 (21): 7188–96.
54. Federhen S. The NCBI taxonomy database. Nucleic Acids Research. 2012; 40: 136–43
55. Turnbaugh PJ, Gordon JI. An invitation to the marriage of metagenomics and metabolomics. Cell. 2008; 134 (5): 708–13.
56. Segata N, Boernigen D, Tickle TL, et al. Computational meta'omics for microbial community studies. Mol Syst Biol. 2013; 9 (1): 666.
57. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput Biol. 2010; 6 (2): e1000667.
58. Thomas T, Gilbert J, Meyer F. Metagenomics – a guide from sampling to data analysis. Microb Inform Exp. 2012; 2: 3.
59. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One. 2012; 7 (2): e30619.
60. Davis MP, van Dongen S, Abreu-Goodger C, et al. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. Methods. 2013; 63 (1): 41–9.
61. Yang X, Liu D, Liu F, et al. HTQC: a fast quality control toolkit for Illumina sequencing data. BMC Bioinformatics. 2013; 14: 33.
62. Mohammed MH, Chadaram S, Komanduri D, et al. Eu-Detect: an algorithm for detecting eukaryotic sequences in metagenomic data sets. J Biosci. 2011; 36 (4): 709–17.
63. Schmeider R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 2011; 6 (3): e17288.
64. Li R, Zhu H, Ruan J, et al. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010; 20 (2): 265–72.
65. Kunin V, Copeland A, Lapidus A, et al. A Bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev. 2008; 72 (4): 557–78.
66. Wang M, Ye Y, Tang H. A de Bruijn graph approach to the quantification of closely-related genomes in a microbial community. J Comput Biol. 2012; 19 (6): 814–25.
67. Claudel-Renard C, Chevalet C, Faraut T, et al. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res. 2003; 31 (22): 6633–9.
68. Bateman A, Coin L, Durbin R, et al. The Pfam protein families database. Nucleic Acids Res. 2004; 32 (Database issue): 138–41.
69. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res. 2003; 31 (1): 371–3.
70. Sonnhammer EL, Von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 1998; 6: 175–82
71. Huson DH, Auch AF, Qi J, et al. MEGAN analysis of metagenomic data. Genome Res. 2007; 17 (3): 377–86.
72. Huson DH, Mitra S, Ruscheweyh HJ, et al. Integrative analysis of environmental sequences using MEGAN4. Genome Res. 2011; 21 (9): 1552–60.
73. Zhu W, Lomsadze A, Borodovsky M. *Ab initio* gene identification in metagenomic sequences. Nucleic Acids Res. 2010; 38 (12): e132.
74. Mitra, S, Rupek P, Richter CD, et al. Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. BMC Bioinformatics. 2011; 12 Suppl 1: 21.
75. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28 (1): 27–30.
76. Segata N, Waldron L, Ballarini A, et al. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012; 9 (8): 811–4.
77. Meyer F, Paarmann D, D'Souza M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008; 9: 386.
78. Markowitz VM, Chen IM, Chu K, et al. IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res. 2012; 40: 123–9.
79. Tanenbaum DM, Goll J, Murphy S, et al. The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. Stand Genomic Sci. 2010; 2 (2): 229–37.
80. CA15120 OpenMultiMed – Open Multiscale Medicine [internet]. COST European Cooperation in Science and Technology; c2012–2018 [citirano 2018 Jun 4]. Dosegljivo na: http://openmultimed.net/

81. PortASAP - European network for the promotion of portable, affordable and simple analytical platforms [internet]. COST European Cooperation in Science and Technology; c2012–2018 [citirano 2018 Jun 4]. Dosegljivo na: http://portasap.eu/
82. Chao A. Non-parametric estimation of the number of classes in a population. Sca J Sta. 1984; 11 (4): 265–70.
83. Magurran AE. Ecological diversity and its measurement. New York: Princeton University Press; 1988.
84. Stres B. Composition and diversity of denitrifying and total microbial community in cultivated soil with genetic markers nosZ and 16S rDNA [doktorsko delo]. Ljubljana: Univerza v Ljubljani; 2003.
85. Kozich JJ, Westcott SL, Baxter NT, et al. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the Miseq Illumina sequencing platform. Appl Environ Microbiol. 2013; 79 (17): 5112–20.
86. Schloss PD. Evaluating different approaches that test whether microbial communities have the same structure. ISME J. 2008; 2 (3): 265–75.
87. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser. 1995; 57 (1): 289–300.
88. Benjamini Y, Yekutieli, D. The control of the false discovery rate in multiple testing under depencency. Ann Stat. 2001; 29 (4): 1165–88.
89. Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011; 12 (6): R60.
90. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol. 2009; 5 (4): e1000352.
91. Dufrene M, Legendre P. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol Monogr. 1997; 67 (3): 345–66.
92. Li K, Bihan M, Methé BA. Analyses of the stability and core taxonomic memberships of the human microbiome. PLoS One. 2013; 8 (5): e63139.
93. Oksanen J, Blanchet FG, Kindt R, et al. Package »vegan«. Version 2.2-1 [internet]. Nairobi: World Agroforestry; c2015 [citirano 2018 Jun 4] Dosegljivo na: http://outputs.worldagroforestry.org/cgi-bin/koha/opac-detail.pl?biblionumber=39154.
94. Armbruster DA, Pry T. Limit of blank, limit of detection and limit of quantitation. Clin Biochem Rev. 2008; 29 Suppl 1: 49–52.
95. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012; 486: 207–14.
96. Sczyrba A, Hofmann P, Belmann P, et al. Critical assessment of metagenome interpretation - a benchmark of metagenomics software. Nat Methods. 2017; 14 (11): 1063–71.
97. Critical Assessment of Metagenome Interpretation, CAMI Challenge [internet]. Wien: Division of Computational Systems Biology; c2017–2018 [citirano 2018 Jun 4]. Dosegljivo na: https://data.cami-challenge.org/
98. Sze MA, Schloss PD. Looking for a signal in the noise: revisiting obesity and the Microbiome. MBio 2016; 7 (4): e01018–16.
99. Cakir T, Khatibipour MJ. Metabolic network discovery by top-down and bottom-up approaches and paths for reconciliation. Front Bioeng Biotechnol. 2014; 2: 62.
100. Perez-Riverola Y, Baia M, Felipe da VL, et al. Discovering and linking public »omics« datasets using the omics discovery index. Nat Biotechnol. 2017; 35 (5): 406–9.
101. Edwards LM, Thiele I. Applying systems biology methods to the study of human physiology in extreme environments. Extrem Physiol Med. 2013; 2: 8.

Prispelo 7. 8. 2018