TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOINFORMAATIKA ÕPPETOOL

Viiruste detekteerimine reoveeproovidest

Bakalaureusetöö

12 EAP

Kristi-Rebeka Tiits

Juhendajad

PhD Maido Remm

TARTU 2022

Viiruste detekteerimine reoveeproovidest

Viirused on väiksed lihtsa ehitusega osakesed, mis kõik sisaldavad DNA-d või RNA-d ja valke. Käesoleva töö eesmärk on kirjeldada DNA analüüsil põhinevaid meetodeid viiruste jälgimiseks. Haigustekitajate jälgimiseks keskkonnas on võimalik viiruseid analüüsida ning üks viis selleks on viiruste järjestuste tuvastamine. Reovesi on keskkond, millest saab alustada viiruste uurimist ja tuvastamist. Analüüsiks on loodud erinevaid programme ja tööriistu, mis on kõik olulised viirusanalüüsi läbiviimisel.

Märksõnad: reovesi, viirused, metagenoomika, detekteerimine, sekveneerimine

CERCS: B110 (Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika)

Detection of viruses in wastewater samples

Viruses are small particles of simple structure, all containing DNA or RNA and proteins.
The aim of this work is to describe methods based on DNA analysis for the monitoring of viruses. Viruses can be analysed to monitor pathogens in the environment, and one way to do this, is by sequencing. Wastewater is an environment from which to start the investigation and identification of viruses. A variety of programmes and tools have been developed for analysis, all of which are essential for virus analysis.

Keywords: sewage, viruses, metagenomics, detection, sequencing

CERCS: B110 (Bioinformatics, medical informatics, biomathematics, biometrics)

SISUKORD

SISUKORD 3

SISSEJUHATUS 5

KIRJANDUSE ÜLEVAADE 6

1. VIIRUSTE DNA ANALÜÜS REOVEEPROOVIDEST 6

1.1 Eesmärgid 6

1.2 Praktilised rakendused 6

2. METAGENOOMSE DNA ANALÜÜSIMISE MEETODID 11

2.1 Puhastusprotokollid 13

2.2 PCR 14

2.3 Sekveneerimine 15

3. METAGENOOMSE DNA ANALÜÜSIMISE TARKVARA 16

3.1 Lazypipe 16

3.2 Kraken 2 17

3.3 LABRADOR 17

4. METAGENOOMSED ANDMESTIKUD 19

4.1 Kasutatavad andmebaasid 19

5. METAGENOOMSE ANALÜÜSIGA SEOTUD PROBLEEMID 21

ARUTELU 23

KOKKUVÕTE 25

SUMMARY 26

TÄNUSÕNAD 28

KASUTATUD KIRJANDUSE LOETELU 29

LIHTLITSENTS 36

KASUTATUD LÜHENDID

BLAST - *Basic Local Alignment Search Tool*

GUI - graafiline kasutajaliides ehk *Graphical User Interface*

HTS – suure läbilaskevõimega sekveneerimine ehk *high-throughput sequencing*

INSDC – *The International Nucleotide Sequence Database Collaboration*

MERS – Lähis-Ida respiratoorne sündroom ehk *Middle East respiratory syndrome*

mNGS – metagenoomiline järgmise põlvkonna sekveneerimine ehk *metagenomic next generation sequencing*

NCBI – *National Center of Biotechnology Information*

NGS – teise põlvkonna sekveneerimine ehk *next generation sequencing*

RefSeq (*database*) – referentsjärjestuste (andmebaas) ehk *The Reference Sequence (database)*

RT-PCR - pöördtranskriptsiooniga polümeraasi ahelreaktsioon ehk *reverse transcription polymerase chain reaction*

RVDB – viiruste referentsandmebaas ehk *Reference Viral Database*

SARS – raskekujuline äge respiratoorne sündroom ehk *severe acute respiratory syndrome*

SURPI – ülikiire järjestuspõhine patogeeni tuvastamine ehk *Sequence-based Ultrarapid Pathogen Identification*

ZIKV – Zika-viirus

WBE – reoveepõhine epidemioloogia ehk *Wastewater Based Epidemiology*

SISSEJUHATUS

Metagenoomiliste uuringute kohaselt on viiruste populatsioonid väga mitmekesised ning levivad peaaegu kõikjal maailmas. Inimestele patogeensed viirused saavad sageli alguse põllumajandusloomadelt ning väiksemal määral ka metsloomadelt. Viirused on võimelised püsima erinevates keskkondades pikalt ning mõned neist kujutavad ohtu inimeste tervisele.

Esimene samm viiruste uurimisel erinevates kooslustes on nende tuvastamine. Kõik nakkustekitajad sisaldavad DNA või RNA genoomi, mis tähendab, et sekveneerimine võib olla pädev viis patogeenide tuvastamiseks, sest võimaldab tuvastada kõik proovis esinevad geneetilised üksused. Metagenoomiline sekveneerimine on oluliselt parandanud teadlaste arusaama viiruste kogukondadest, mida varem ei olnud võimalik tuvastada. Tänu järgmise põlvkonna sekveneerimise kasutusele võtmisele, on võimalik ka süsteemselt määratleda proovis esinevate viiruste spektrit.

Keskkonnast võetud proovidest viiruste tuvastamine on aga keerulisem, sest viirustel puudub ühtne fülogeneetiliselt informatiivne geen, mis oleks sarnane bakterite 16S ribosomaalse RNA geeniga. Seetõttu kasutatakse viiruste tuvastamiseks sageli järjestuse homoloogiaotsinguid BLAST-i või Kraken-iga (Gregory et al., 2020). Metagenoomiliste andmete suurenenud kättesaadavus järgmise põlvkonna sekveneerimise ajastul, koos järjest taskukohasemate sekveneerimistehnoloogiatega, rõhutab vajadust usaldusväärsete ja terviklike analüüsi ja andmehalduse meetodite järele. Hiljutine koroonaviiruse pandeemia suurendas ka vajadust uurida ja tuvastada seni vähemlevinud viiruseid, mis võivad siiski kujutada inimkonnale suurt ohtu.

Käesoleva töö kirjanduse ülevaade annab ülevaate DNA-põhistest analüüsimeetoditest, mille abil on võimalik viiruseid reoveest detekteerida ning analüüsida. Töö eesmärk on kirjeldada DNA-l põhinevaid meetodeid viiruste jälgimiseks.

KIRJANDUSE ÜLEVAADE

1. VIIRUSTE DNA ANALÜÜS REOVEEPROOVIDEST

1.1 Eesmärgid

Maailmas on palju erinevaid viiruspatogeene ning nende leviku jälgimiseks on oluline viirustekitajaid põhjalikult analüüsida. Inimeste tervisele kujutab ohtu pinnavee saastumine, mille puhul võivad vees eksisteerivad viirused sattuda ka joogivette (Moreira & Bondelind, 2017). Viiruste leviku alguspunkti, edasise leviku ja muutuvate suundumuste jälgimine on oluline, et kaardistada ühiskonnas levivaid inimestele patogeenseid viiruseid. Lisaks kokkupuute-eelsele profülaktikale või ravile, on kriitilise tähtsusega nakatunud isikute kiire tuvastamine.

Epideemiate puhul on pikka aega olnud peamine lähenemisviis diagnostiliste testide laiaulatuslik rakendamine üksikjuhtumite puhul. Suurte pandeemiate puhul, nagu oli COVID-19 pandeemia, tuleb välja selle meetodi kitsaskoht ehk suurte populatsioonide kiire uurimine. Vee teel levivad viirused on geneetiliselt väga mitmekesised, mis muudab keeruliseks ka viirushaigustele suunatud usaldusväärsete diagnostiliste testide loomise. Viiruste tuvastamiseks on üks võimalus koguda reoveest proovid ning viia läbi põhjalik analüüs. Reoveeproovide analüüsimine annab laiaulatuslikku informatsiooni uuritavas keskkonnas eksisteerivate viiruste ja nende erinevate tüvede leviku kohta (Nieuwenhuijse & Koopmans, 2017). Keskkonnateadlased on välja pakkunud WBE ehk reoveepõhise epidemioloogia, mis võiks suuri haiguspuhanguid leevendada või ohjeldada (Daughton, 2020).

1.2 Praktilised rakendused

Reoveepõhine epidemioloogia pakub aja- ja kulutõhusat viisi inimese seedetrakti kaudu levivate viiruste jälgimiseks ning kliiniliste uuringute täiendamiseks. Viiruseid on erinevates veemaatriksites rohkelt uuritud, sealhulgas näiteks joogivett, kaevuvett, merevett jpm (Corpuz et al., 2020). Üle saja liigi on tuvastatud viiruseid, mis levivad inimese seedetrakti kaudu, muuhulgas näiteks rota- ja enteroviirused. 2016. aasta esimeses pooles tehtud uuringust selgus, et kõige laialdasemalt levisid inimesele patogeensetest viirustest astroviirused. Enteroviiruseid leidus vähem, kuid siiski kogu maailma reoveeproovidest. Noro- ja rotaviiruseid leidus vaid vähesel määral (Nieuwenhuijse et al., 2020).

Viiruste tüvede mitmekesisuse uurimiseks on võimalik analüüsida pikki järjestusi ning tuvastatud viiruste iseloomustamine on oluline, et oleks võimalik detekteerida uusi tüvesid ja sellega vähendada riski puhangute tekkeks. Sageli kasutatakse suure koguse vee kontsentreerimist, et vähendada proovivõtu varieeruvust ja suurendada analüüside tundlikkust. Suurema arvu proovide analüüsimine on ka keskkonna osas informatiivsem ehk peegeldab nii inimestele patogeensete kui mittepatogeensete viiruste mitmekesisust uuritavas piirkonnas (Strubbia et al., 2019).

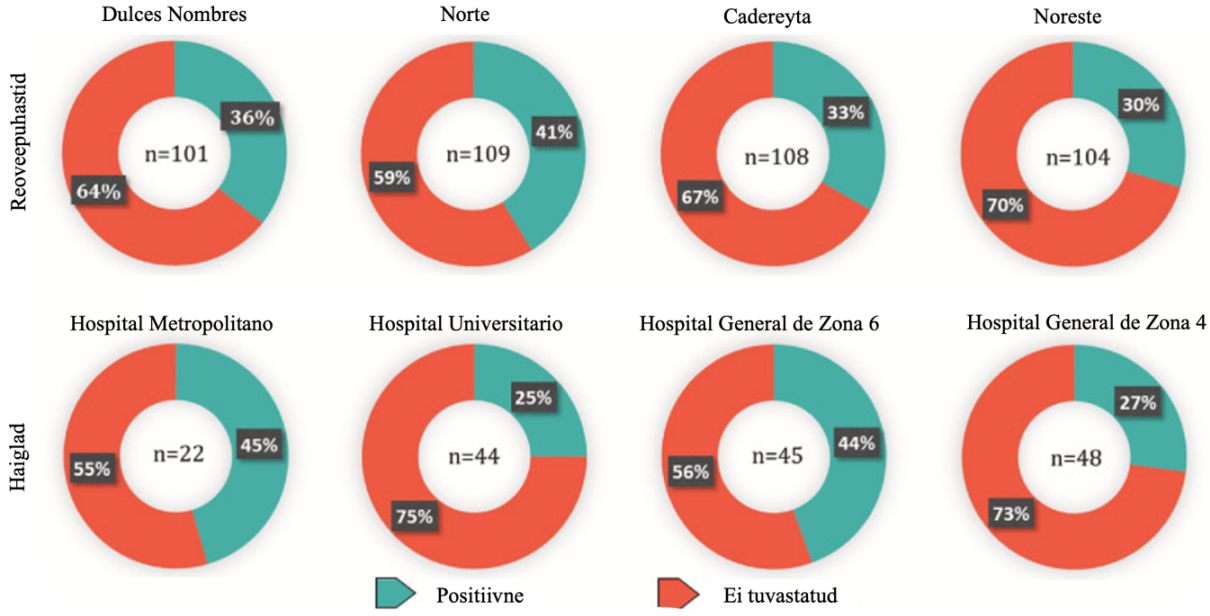
Kaasaegse maailma suurenenud ühenduvus ja ka kliimamuutused võimaldavad viirustel levida uutes piirkondades, mida oli näha näiteks Ebola haiguspuhangu korral. Kirjeldatud on veel linnade reovee uuringut, kus käsitleti metagenoomilist sekveneerimist, kui potentsiaalset strateegiat ülemaailmsete pandeemiate valmisolekuks. Sekveneerimise kasutamisel seires on oluline järjestusi ka õigesti tõlgendada, sest vigane annotatsioon muudab liigi määramise keerulisemaks (Nieuwenhuijse et al., 2020).

Koroonaviirus

Vähem kui kahe aastakümne jooksul on inimkonnas levinud kolm zoonootilist koroonaviirust: (1) äge raskekujulise respiratoorse sündroomiga koroonaviirus (SARS-CoV), (2) Lähis-Ida respiratoorse sündroomi koroonaviirus (MERS-CoV) ning (3) SARS-CoV-2. Koroonaviirused on ümbrisega viiruste rühm, millel on üheahelaline RNA genoom. RNA viirustele iseloomuliku kõrge mutatsioonimäära tõttu on koroonaviirused võimelised kiiresti suurendama oma virulentsust ja kohanema uute peremeesorganismidega (Foladori et al., 2020).

2019. aastal hakkas levima SARS-CoV-2, mis kinnitas seda, et koroonaviirused on võimelised pidevalt muteeruma ning muutuma inimesele veel patogeensemaks kui varem. SARS-CoV-2 suhtes positiivse tulemuse andnud inimeste väljaheites sisalduv viiruskoormus on varieeruv, hinnanguliselt 5 x 103-107,6 koopiat/ml, sõltuvalt nakkuse kulgemisest. SARS-Cov-2 nukleiinhappe osakesi on tuvastatud ja uuritud reoveeproovidest. Analüüsi tulemusena on täheldatud, et üle poolte uuringus osalenud patsientide fekaalproovid püsisid SARS-CoV-2 osas positiivsed isegi siis, kui hingamisteedest võetud proovid näitasid juba negatiivset tulemust. Osadel patsientidel leidus viiruse RNA-d veel isegi 30 päeva pärast haigust, kuid keskmiselt püsisid proovid positiivsed 11,2 päeva kauem, kui hingamisteedest võetud proovid (Foladori et al., 2020). See viitab tõenäoliselt sellele, et viirus suudab replitseeruda seedetraktis ka pärast hävimist hingamisteedes (Wu et al., 2020).

SARS-CoV-2 uurimiseks koguti 2020. aasta juunist detsembrini proovid Mehhikost, Monterrey suurlinna reoveest. Uuringul oli kolm peamist eesmärki: (1) tuvastada ja kvantifitseerida SARS-CoV-2 viiruse esinemist ja selle ajalist dünaamikat reoveepuhastusjaamadest ning haiglate heitveest, (2) uurida, kas pikaajaliselt kogutud andmed kajastavad nakkustaseme arvu elanikkonnas, võttes arvesse ka sõltumatuid muutujaid (viiruskoormus reovees, reovee voolukiirus jne), (3) uurida tuvastamismäärasid ja seire täpsust. Kokku koguti ja töödeldi 422 reoveeproovi veepuhastitest ning 159 heitveeproovi haiglatest. SARS-CoV-2 tuvastamiseks kasutati antud katses RT-PCR ehk pöördtranskriptsiooniga PCR meetodit. Hinnanguline madalaim tuvastuspiir oli SARS-CoV-2 N geeni ~ 1 koopia/reaktsioon ml vee kohta. Positiivsete proovide osakaal oli kõigi puhastusseadmete puhul sarnane (35,0 ± 4,8%). Haigla reoveest kogutud positiivsete proovide osakaal oli samuti ühtlane (35,5 ± 9,5%), kuid natuke suurema varieeruvusega (joonis 1) (Padilla-Reyes et al., 2022).



Joonis 1. SARS-CoV-2 RNA viiruse tuvastamine. Tulemused reoveepuhastitest (ülemine rida) ja haiglatest (alumine rida). Märgitud on proovide koguhulk (n) ja positiivsete tulemuste osakaal (Padilla-Reyes et al., 2022, kohandatud)

Koroonaviiruse hiljutist pandeemiat on palju uuritud. Teostatud on katse, kus mõõdeti viiruse kontsentratsiooni COVID-19 hospitaliseeritute väljaheidetes ning tuvastati, et viiruse kontsentratsioonid olid kõrgeimad just esimesel nädalal pärast sümptomite ilmnemist (Wölfel et al., 2020). SARS-CoV-2 tuvastamiseks kasutatakse kõige rohkem RT-PCR meetodit ning viiruse RNA leidmine proovis ei tähenda, et viirus on elujõuline ja nakkav. SARS-CoV-2 puhul

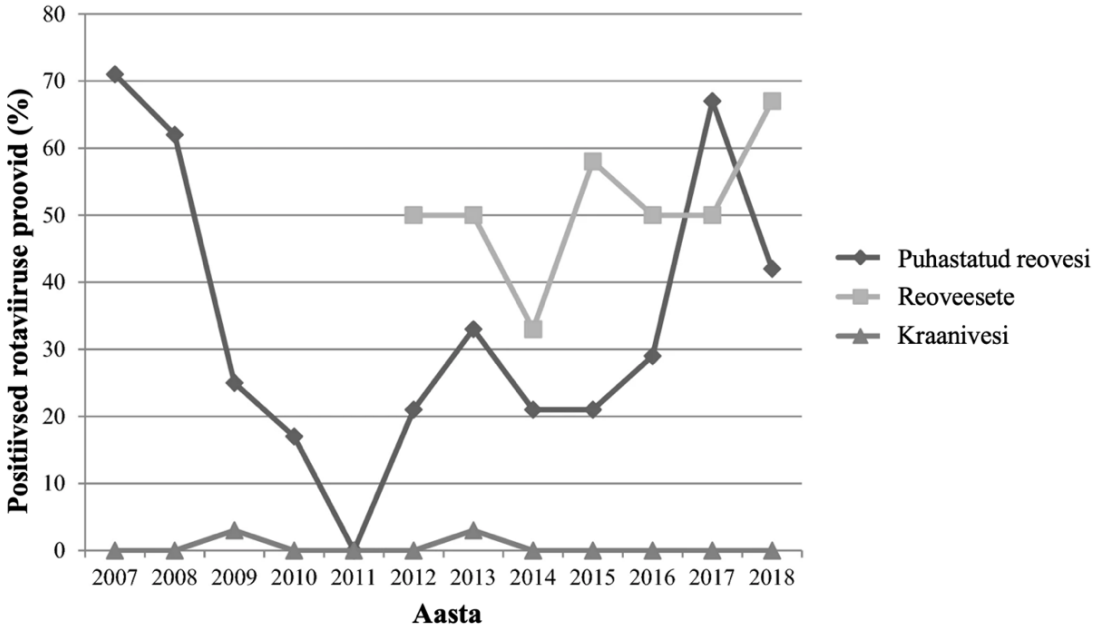
langeb elujõulisus näiteks temperatuuri ja pH tõttu, kuid kõrge inaktiveerituse saab saavutada vaid desinfitseerimise (kloori, UVC-valguse) abil (Foladori et al., 2020). Kliinilise testimise abil PCR-iga püütakse luua ka COVID-19 seiresüsteem (Padilla-Reyes et al., 2022). Viimase 30 aasta jooksul on kogutud rohkelt teavet koroonaviiruste replikatsiooni ja patogeneesi kohta, kuid siiski tuleb tõhusaks ennetuseks koguda veel informatsiooni lisaks (Wölfel et al., 2020).

Noroviirus

Noroviiruste genoomid on väga mitmekesised ning viiruse kontsentratsioon on nakatunud inimese proovis üldiselt suur. Läbi on viidud katseid hindamaks viiruste tuvastamise meetodeid metagenoomika abil, keskendudes just noroviirustele. Ühe uurimuse eesmärk oli tuvastada pikki fragmente, et hinnata populatsioonis esinevate tüvede mitmekesisust (Strubbia et al., 2019). Lisaks üldlevinud GII.4 tüvele, mis on levinud kogu maailmas, tuvastati ka GII.P17-GII.17 genotüübiga noroviiruse tüve, mis põhjustas mitmetes riikides gastroenteriidi puhanguid (Matsushima et al., 2015; Koo et al., 2017). Samuti leiti ka GIV.1 tüve, mida sageli alahinnatakse gastroenteriidi põhjustajana, arvatavasti sellepärast, et GenBank-is (Benson et al., 2012) on saadaval vaid väike arv järjestusi ehk praimerite kujundamine on selle tüve puhul keeruline (Strubbia et al., 2019).

Rotaviirus

Rotaviirused on inimestele nakkusohtlikud ning kanduvad edasi kas otse inimeselt inimesele või saastunud toidu või vee kaudu (Greenberg & Estes, 2009). Kirjeldatud on uuringut Tais läbi viidud rotaviiruste seirest reovees. Uurimiseks koguti proovid kraaniveest, puhastatud reoveest ja reoveesettest aastatel 2007-2018 ning viiruste esinemissagedused olid väga varieeruvad (joonis 2). Proovid koguti ning neid töödeldi rotaviiruse VP7 geeni tuvastamiseks *nested* PCR-i abil. Analüüsiks kasutati ka VP7-spetsiifilisi praimereid. Rotaviiruse genotüübid tuvastati sekveneerimise ja fülogeneetilise analüüsi abil. *Nested* PCR-i tundlikkus määrati kindlaks nii, et proovidesse külvati teadaolevad rotaviiruse kogused. Kogutud keskkonnaproovide põhjal tuvastati näiteks, et 50% reoveesettest kogutud proovidest olid rotaviiruse suhtes *nested* PCR-iga tuvastades positiivsed (Kittigul & Pombubpa, 2021).



Joonis 2. Rotaviiruse sagedusjaotus kraanivees, puhastatud reovees ja reoveesettes (Kittigul & Pombubpa, 2021, kohandatud).

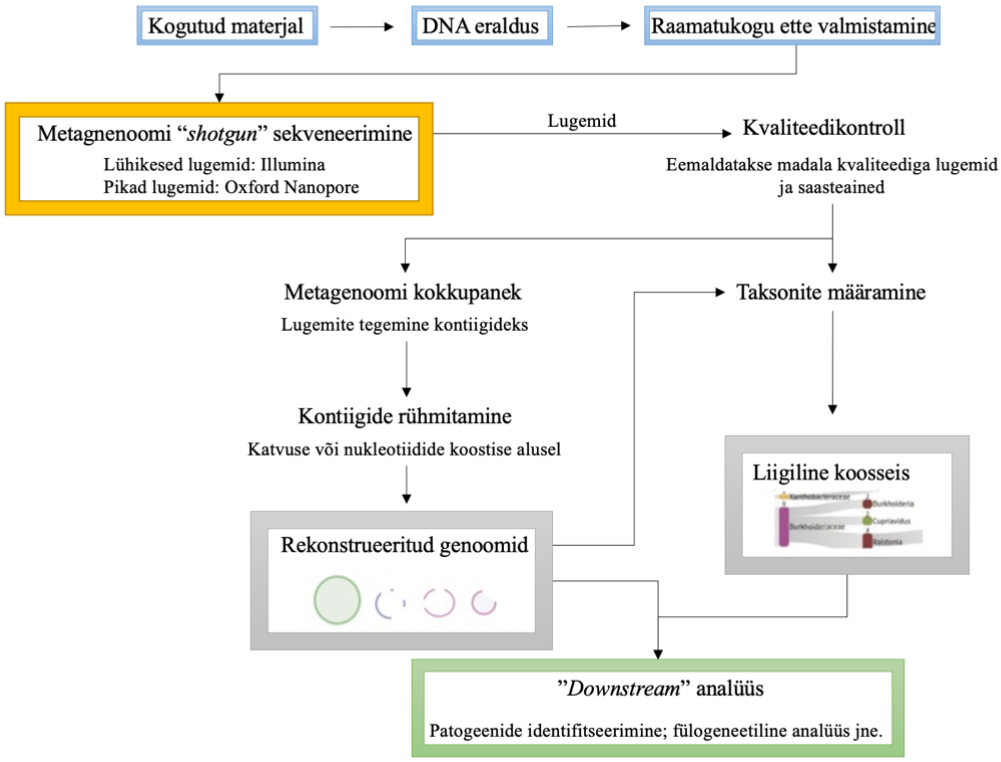
Enteroviirus

Enteroviirus on üheahelaline ümbriseta RNA viirus, mis levib inimese soolestiku kaudu. Keskkonnaseire abil on võimalik tuvastada enteroviiruse levikut ja sellist lähenemist kasutatakse tänapäeval kogu maailmas. Reovee kasutamist viiruste tuvastamiseks, mille puhul suremus on väike, kuid mõju elanikkonnale suur, on testitud näiteks poliomüeliidi likvideerimise edendamiseks (Majumdar et al., 2018). 14 aastat kestnud enteroviiruse seire proovidest tuvastati 20% positiivseid proove. Kokku analüüsiti 5450 proovi ning 1089 nendest olid positiivsed. Positiivsete proovide koguarv ulatus 8,9%-st 2016. aastal kuni 26,6%-ni 2012. aastal. Keskkonnaseire, mis koosneb suures osas reoveeseirest, võimaldab tuvastada polioviiruste levikut elanikkonnas isegi kliiniliste poliomüeliidi juhtumite puudumisel. See on väga oluline epidemioloogilise hindamise ja varajase reageerimise seisukohalt (Ivanova et al., 2019).

2. METAGENOOMSE DNA ANALÜÜSIMISE MEETODID

Keskkonna DNA-d on aja jooksul hakatud maailmas aina rohkem uurima ja metagenoomika võimaldab kirjeldada kõiki proovis esinevaid nukleiinhappe järjestusi (Strubbia et al., 2019). Enamikel inimese seedetrakti kaudu levivatel viirustel on väike RNA genoom, mistõttu on selliste viiruste tuvastamine keeruline. Madal arvukus võrreldes keskkonnas esinevate bakterite, faagide ja parasiitidega teeb nende viiruste tuvastamise veel keerulisemaks (Adriaenssens et al. 2018), mistõttu eemaldatakse proovist võimalikult palju peremeesorganisme ja baktereid. Sellele järgneb nukleaasi töötlus vabade nukleiinhapete eemaldamiseks (Strubbia et al., 2019). Viiruse tuvastamise täpsus sõltub veel ka muudest aspektidest - näiteks viiruse kontsentratsioonist proovis ja nukleiinhappe ekstraheerimise saagise puhtusest. Hjelmsø et al., 2017 artiklis läbi viidud uuring näitas, et nii nukleiinhapete kontsentratsioon, kui ekstraheerimiseks kasutatud meetodid võivad samuti mõjutada oluliselt viirusliku metagenoomilise analüüsi tulemusi (Corpuz et al., 2020).

Suur osa klassikalisi molekulaarseid analüüse on suunatud piiratud patogeenide arvule, mille puhul kasutatakse spetsiifilisi praimereid või sonde (Ren et al., 2017). Näiteks PCR meetodi puhul saabki tuvastada ainult ühe patogeeni korraga, sest analüüsiks on vaja teada uuritavat järjestust. Metagenoomilised analüüsimeetodid aga iseloomustavad kogu proovis leiduvat DNA-d või RNA-d ning nende abil on võimalik saada ülevaade kõikidest uuritavas keskkonnas eksisteerivatest viirustest (H.-H. Lin & Liao, 2017). Andmete analüüsimiseks on erinevaid võimalusi ning tavaliselt koosneb see protsess kindlaksmääratud osadest (joonis 3) (Breitwieser et al., 2019). Juhusliku järjestamise laialdane kasutus annab võimaluse hinnata viiruse levikut populatsioonis ka pandeemia vaheaegadel (J. Lin et al., 2017).



Joonis 3. Tavalised metagenoomika andmete analüüsi osad (Breitwieser et al., 2019, kohandatud).

Viirusjärjestuste tuvastamine metagenoomides, mis sisaldavad nii viiruse kui peremeesorganismi kontiige, on proovide viirusanalüüsil esimene samm. Järjestuste tuvastamiseks on välja töötatud mitmeid tööriistu ja protokolle, mis põhinevad erinevatel algoritmidel. Näiteks peetakse järjestust viiruslikuks, kui see on märkimisväärselt sarnane teadaolevate viirustega (Roux et al., 2019). VirFinder (Ren et al., 2017) on üks k-meeri sagedusel põhinev tööriist, mida kasutatakse viiruse kontiigide tuvastamiseks. Ren et al., 2017 kirjeldati katset, milles testiti VirFinder-i jõudlust viirusjärjestuste tuvastamisel ning süsteemi võrreldi ka VirSorter-iga (Roux et al., 2015), mis on tunnustatud geenipõhiste viiruste klassifitseerimise tööriist. Katsest järeldati, et VirFinder tuvastas märkimisväärselt paremini viiruste kontiige kui VirSorter, kui hindamine toimus täielike genoomide või simuleeritud inimese soolestiku metagenoomi põhjal kokku pandud kontiigide abil. Samuti leiti, et VirFinder töötab väikeste kontiigide puhul tunduvalt paremini kui VirSorter (Ren et al., 2017).

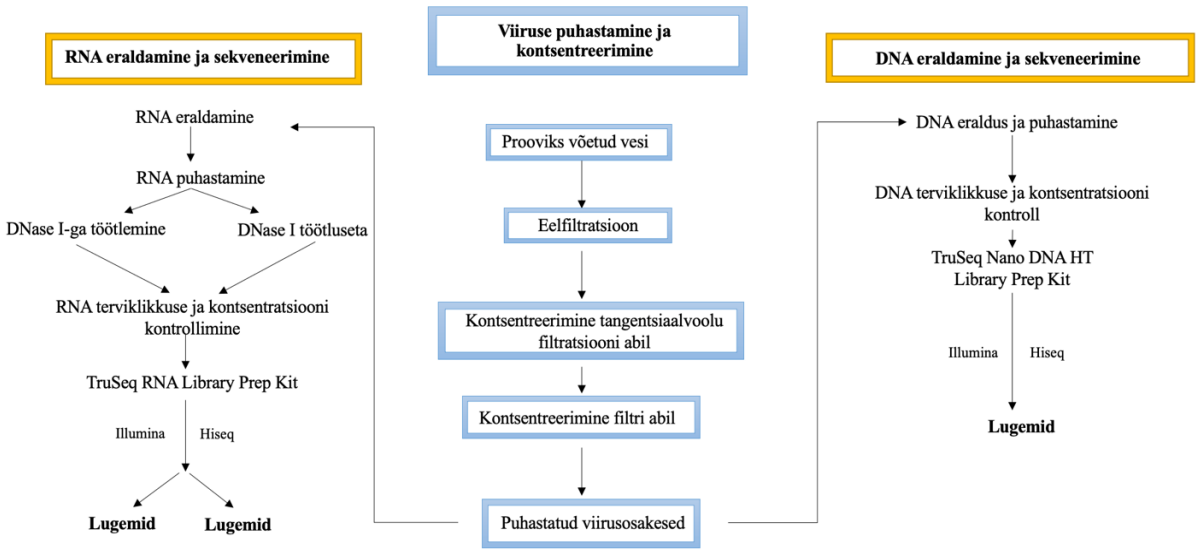
Viirusjärjestusi on võimalik küll analüüsida ka andmekogumeid rikastamata, muuhulgas näiteks rakkudest, kudedest või keskkonnaproovidest. Andmekogumite analüüsimisel ilma viiruste rikastamiseta on erinevaid eeliseid. Võimalik on tuvastada näiteks erineva infektsiooni

intensiivsusega haigeid, mis kõrvaldab genoomi suurusest tingitud kallutatuse. Pärast viiruste rikastamist saadud andmekogumite peamised eelised on aga (1) nii arvukate kui haruldaste viiruste hea genoomide järjestuste *de novo* kokkupanek ja (2) kindlus, et järjestus on viirusliku päritoluga (Roux et al., 2019).

Kultiveerimata viiruste genoomide ehk UViG-de analüüs on väärtuslik viiruste genoomijärjestuste uurimiseks. UViG-de *de novo* assambleerimisel on tehtud edusamme ning seda nii biootilisest kui abiootilisest keskkonnast. Viimase paari aastaga on metagenoomide ja metatranskriptoomide andmekogumites tuvastatud üle 750 000 kultiveerimata viiruse genoomi ja see arv on viis korda suurem, kui viiruse isolaatidest sekveneeritud genoomide koguarv. UViG-d moodustavad praegu juba suure osa taksonoomilisest mitmekesisusest avalikult kättesaadavates viirusjärjestustes (Roux et al., 2019).

2.1 Puhastusprotokollid

Reoveeproovide analüüsimiseks tuleb läbi viia spetsiifilised eeltöötlusetapid, et proovist oleks viiruste avastamine tõhusam. Selleks on võimalik näiteks puhastada või kontsentreerida viirust (Foladori et al., 2020). Kirjeldatud on katset, kus viiruste kontsentreerimine viidi läbi adsorptsiooni teel lõssipulbrile, kuid võimalik on ka materjali sadestamine (Hellmér et al., 2014). Sageli ei ole metagenoomika uuringute jaoks standardset lähenemisviisi või üht parimat protokolli, kuid sobimatud DNA puhastusprotokollid võivad põhjustada vigu viiruste analüüsitulemuste hindamisel. Erinevad metagenoomikas kasutatavad meetodid võivad viia osade taksonite üleesindatuseni ning vähese arvukusega taksonite välja jätmiseni, mistõttu on väga oluline õige DNA ekstraheerimismeetodi valik ja see peab vastama ka uurimisküsimustele. Mitmed uuringud näitavad, et erinevate ekstraheerimismeetodite kombineerimine ühe keskkonna suhtes aitab vähendada üksikute meetoditega seotud taksonite esindatuse kallutatust (Fidler et al., 2020). Läbi on viidud katse viiruse kontsentratsiooni kahekordistamisest, kus kirjeldati analüüsile eelnevaid puhastusprotokolle. Veest kogutud DNA ja RNA viroomide puhastamiseks ja sekveneerimiseks järgiti kindlaid etappe ning metoodikaid (joonis 4) (Wolf et al., 2020).



Joonis 4. DNA ja RNA viroomide puhastamise ja sekveneerimise skeem. DNase I=deoksüribonukleaas I. (Wolf et al., 2020, kohandatud).

2.2 PCR

Polümeraasi ahelreaktsiooni ehk PCR-põhiseid meetodeid on aastate jooksul kohandatud paljude molekulaarbioloogiliste rakenduste jaoks ning nendest on saanud olulised igapäevased töövahendid teaduslaborites. PCR-i peetakse lihtsaks ja usaldusväärseks meetodiks, et paljundada kindlat nukleiinhappe järjestust ning tuvastada seda proovist. Viiruse esinemist reoveeproovis võib kinnitada PCR meetod. PCR-i abil on võimalik tuvastada viirust, mida esineb uuritavas proovis väga vähe ning PCR-põhised meetodid on küll tundlikud viiruste nukleiinhapete tuvastamisel, kuid suudavad tuvastada ainult neid viiruseid, mille jaoks on loodud praimerid ja reaktsioonitingimused optimeeritud. Enne PCR testi välja töötamist on vaja teada antud viiruse genoomseid DNA järjestusi, sest nende alusel tuleb disainida ka praimerid, mis on üldjuhul oligonukleotiidid. Kvantitatiivne ehk reaalaja PCR on suurendanud PCR meetodite usaldusväärsust veelgi tänu oma kiirusele ja tundlikkusele (Mackay et al., 2002). Proovis olevat koopiate arvu hinnatakse tsükli läviväärtuste abil, mis on pöördvõrdeline viiruse hulgaga (Foladori et al., 2020).

2.3 Sekveneerimine

Arvestades, et nukleiinhappejärjestus praimerite kujundamiseks ei pruugi olla alati saadaval, võib potentsiaalne saastav viirus jääda PCR-põhiste meetodite puhul tähelepanuta. Sellisel juhul on võimalik rakendada suure läbilaskevõimega sekveneerimist, mis võimaldab tuvastada mistahes DNA või RNA molekuli proovist, sõltumata selle järjestusest (Fabiańska et al., 2021). Viroloogid on enamasti keskendunud viiruste uurimisele, mis põhjustavad erinevaid haigusi inimestel ja koduloomadel, kuid metagenoomilise sekveneerimise abil on võimalik uurida viroomi kogu biosfääris. Kui varasemalt oli uue viiruse kirjeldamiseks vaja teavet peremehe, replikatsioonitsükli ja ka viiruse osakeste kohta, siis tänu suure läbilaskevõimega sekveneerimisele ja metagenoomilistele meetoditele on järjestusandmete põhjal tuvastatud väga palju uusi viiruseid. Sekveneerimise abil on võimalik tuvastada viiruse genoomset materjali ka sihtimata konkreetset viiruspatogeeni ning selle meetodi põhjal on iseloomustatud suur hulk genoomijärjestusi (Simmonds et al., 2017).

Sekveneerimistehnoloogiad arenevad pidevalt ning tänu sellele on jõudsalt suurenenud ka sekveneeritud genoomide arv ja nende mitmekesisus. 2005. aastal kasutusele võetud NGS ehk teise põlvkonna sekveneerimine andis metagenoomika valdkonnale olulise tõuke, sest esimest korda oli võimalik luua miljoneid kuni miljardeid lugemeid ühe korraga ning see võimaldas analüüsida kogu kliinilise- või keskkonnaproovi geneetilist sisu (Chiu & Miller, 2019). Viiruste genoomid on suhteliselt väikesed võrreldes bakterite või eukarüootide genoomidega, mistõttu on NGS-i rakendamisel oluline viiruse genoomi rikastamine (Khan et al., 2020). Tänapäeval kasutavad teise põlvkonna sekveneerimismeetodid kogu genoomi „*shotgun*“ sekveneerimist ning NGS-i puhul on kaks peamist paradigmat: lühikese- ja pika lugemiga sekveneerimine (Khan et al., 2020). Lühikese lugemiga meetodid on üldjuhul odavamad ja võimaldavad täpsemaid andmeid. Pika lugemiga sekveneerimise abil on aga võimalik analüüsida keerulisi või korduvaid piirkondi ühe korraga (Goodwin et al., 2016).

3. METAGENOOMSE DNA ANALÜÜSIMISE TARKVARA

Bioinformaatilised tööriistad ja andmebaasid on vajalikud viiruste tõhusaks tuvastamiseks sellistest järjestustest, mis võivad sisaldada nii viirusjärjestusi, viiruste sarnaseid järjestusi kui bakterite järjestusi. Viiruste avastamist mõjutab oluliselt bioinformaatiliste vahendite valik ja parameetrite optimeerimine. Parameetrid peavad alati olema kohandatud vastavalt kasutatavale programmile või platvormile. Töökindlad bioinformaatilised vahendid on olulised ka viiruste kiireks klassifitseerimiseks. Olemasolevad arvutuslikud vahendid keskenduvad sageli kas uute viiruste avastamisele või viroomi koostise analüüsile, kuid mitte mõlemale. Näiteks VirusSeeker (Zhao et al., 2017) on BLAST-il (*Basic Local Alignment Search Tool)* (Altschul et al., 1997) põhineva NGS andmeanalüüsi töövoog, mida on võimalik kasutada mõlemaks otstarbeks (Zhao et al., 2017).

Sekveneerimisel on andmete analüüs oluline osa protsessist ning selleks on võimalik kasutada viiruste tuvastamist (1) spetsiifiliste k-meeride abil otse lugemitest, (2) lugemite mäppimine/paigutamine referentsgenoomile või (3) viiruse lugemid omavahel kokku assambleerida ja seejärel pikemaid osasid kasutada homoloogiaotsingus (Ren et al., 2017). Teise põlvkonna sekveneerimisandmete analüüsimiseks on loodud näiteks SURPI (Naccache et al., 2014), mis on arvutuslik töövoog patogeenide identifitseerimiseks NGS-andmetest. Samuti on loodud Taxonomer (Flygare et al., 2016) tööriist, mis võimaldab lugemite kiiret klassifitseerimist ning VirusTAP (Yamashita et al., 2016), mis on veebipõhine integreeritud NGS analüüsivahend viiruse genoomi koostamiseks metagenoomilistest lugemitest (H.-H. Lin & Liao, 2017). Kuna metagenoomiline sekveneerimine on pidevalt arenev tehnoloogia, ei ole see veel täielikult standardiseeritud. Erinevused võivad esineda näiteks proovide eeltöötlemisel kui andmeanalüüsil, mis lõpuks mõjutavad lugemite arvukust, mitmekesisust ja ka annoteeritud järjestuste osakaalu (Nieuwenhuijse et al., 2020).

3.1 Lazypipe

Lazypipe (Plyusnin et al., 2020) on töövoog järgmise põlvkonna sekveneerimise raamatukogude automaatseks kokkupanekuks ja taksonoomiliseks profileerimiseks. Programm toimib lugemite asemel kokkupandud kontiigide päringu abil, millega on võimalik tuvastada viiruste taksoneid väga täpselt. Nii kliinilistest kui keskkonna proovidest suudab Lazypipe avastada ka uudseid viiruseid ning neid iseloomustada. Näiteks SARS-CoV-2 pandeemia ajal testiti Lazypipe-ga Illumina positiivseid raamatukogusid ning kinnitati, et üheksas

raamatukogus kümnest tuvastati SARS-CoV ja seda ilma SARS-CoV-2 referentsgenoomita. Homoloogide otsimiseks kasutab Lazypipe UniProt (Consortium, 2011) andmebaase. Lazypipe paneb ka kokku ja annoteerib viiruste kontiigid. Programm toetab andmeformaate, mida saavad kasutada nii teadlased kui automatiseeritud tööriistad. Tulemused väljastatakse näiteks Exceli tabelite ja interaktiivsete graafikute kujul, aga ka taksonoomiliste profiilide kujul (Plyusnin et al., 2020).

3.2 Kraken 2

Metagenoomiliste järjestusandmete kiiret taksonoomilist klassifitseerimist võimaldab näiteks Kraken-i (Wood ja Salzberg, 2014) k-meeridel põhinev lähenemisviis. Kuna võrdlusandmed on mahukad ja mäluvajadus aina suurem, siis täiendati klassikalist Kraken-it ning loodi Kraken 2 (Wood et al., 2019) tööriist. Kraken 2 vähendab oluliselt mälukasutust, säilitades suure täpsuse ja suurendades kiirust viis korda. Kraken 2 suudab klassifitseerida paarikaupa andmeid üle 93 miljoni lugemi minutis. Kraken-iga analüüsi tehes on sisendiks FASTQ järjestused ning tulemusena saab vaadata, et mitu protsenti segust on näiteks viiruseid, baktereid jne. Tehtud uuringute põhjal võib öelda, et Kraken 2 täpsus on sarnasel tasemel Kraken-i täpsusega. Transleeritud otsingurežiim Kraken 2X kasutab vähendatud aminohapete tähestikku ning suurendab tundlikkust viiruslikel andmekogumitel võrreldes nukleotiidipõhise otsinguga. Otsingu teostamiseks koostab Kraken 2X andmebaasi referentsvalkudest samamoodi nagu Kraken 2 teeb seda nukleotiidjärjestuste puhul. Tranlseeritud otsing on kasulik suure geneetilise varieeruvuse ja kättesaadavate referentsgenoomide vähesuse korral (Wood et al., 2019).

3.3 LABRADOR

Arvutuslik andmeanalüüs ning võrdlusandmebaasi sidusus ja täpsus mõjutavad tugevalt suure läbilaskevõimega sekveneerimise uuringute tulemust. LABRADOR (Fabiańska et al., 2021) on arvutuslik töövoog viiruse tuvastamiseks suure läbilaskevõimega sekveneerimisandmetes. Programm tuvastab viiruseid kasutades sekveneerimislugemite mustripõhist klassifitseerimist koos kaardistamisega vastavale nukleotiidjärjestusele. Selline lähenemine võimaldab tuvastada madala kontsentratsiooniga või ka lühikesi viirusfragmente. LABRADORi analüüs bioloogilise proovi Illumina paarislugemiga võtab aega umbes 150 minutit. Graafiline kasutajaliides (GUI) võimaldab esitada sisendina FASTQ-faile. LABRADOR-iga viidi läbi taksonoomiline

klassifikatsioon ning leiti, et programm suudab klassifitseerida ka madala sekveneerimise sügavusega viiruseid. Samuti tuvastati, et tööriista tundlikkus on 90,5% liigi ja 93,3% perekonna tasandil ehk programmi saab nende andmete põhjal pidada usaldusväärseks (Fabiańska et al., 2021).

4. METAGENOOMSED ANDMESTIKUD

Suure läbilaskevõimega sekveneerimistehnoloogia pakub võimalust uurida inimese viroomi, kuid võmalikult täpse viirusanalüüsi läbi viimiseks on vaja usaldusväärseid võrdlusandmestikke. Andmete analüüsimiseks on kriitilise tähtsusega bioinformaatilise platvormi kvaliteet ning samuti võimalus saada andmete hindamiseks ligipääs ekspertteadmistele (Khan et al., 2020). HTS ehk suure läbilaskevõimega sekveneerimine võib toota miljoneid lugemeid, mida on võimalik tuvastada võrdlemisel viiruste võrdlusandmestikega. Paljud laborid on loonud enda võrdlusandmebaasid, kuid järjestusi võrdlemiseks kogutakse näiteks NCBI (*National Center for Biotechnology Information*) või UniProt-i ressurssidest (Gleizes et al., 2020). Andmebaasi suurus ja annotatsioonide täpsus mõjutab tugevalt viiruse liikide ja tüvede tuvastamist.

4.1 Kasutatavad andmebaasid

NCBI *Taxonomy* on suur andmebaas, mis sisaldab organismide nimesid ja klassifikatsioone iga järjestuse kohta, mis on esindatud INSDC (*The International Nucleotide Sequence Database Collaboration*) (Cochrane et al., 2016) nukleotiidide ja valkude järjestuste andmebaasides (Schoch et al., 2020). NCBI *Viral Genomes Resource* on vahend, mille abil on võimalik korrastada suurt hulka järjestusi ning lihtsustada ka viiruste järjestusandmete kasutamist. Andmebaas sisaldab kõiki avalikult kättesaadavaid genoomijärjestusi ning kureerib referentsgenoomijärjestusi (Brister et al., 2015).

RefSeq (O'Leary et al., 2016) on NCBI referentsjärjestuste projekt. Andmebaas on laialdaselt kasutatav ning sisaldab kureeritud järjestusi, näiteks kogu genoomi. RefSeq sisaldab ligi 200 000 bakterite ja arhede genoomi ning 150 miljonit valku koos asjakohase annotatsiooniga (Li et al., 2021). RefSeq-i poolt kureeritud ja stabiilselt annoteeritud võrdlusgenoomide ja valkude pakkumine viiruste, mikroobide ja eukarüootsete organismide jaoks on võimaldanud teadlastel keskenduda parimatele järjestusandmetele ning viidata üheselt konkreetsetele geneetilistele järjestustele. RefSeq-i andmekogumit kasutatakse laialdaselt võrdlusstandardina erinevate analüüside jaoks, sealhulgas kliinilistes rakendustes, genoomikas ja järjestuse varieeruvuse tõlgendamisel. Referentsjärjestuste standardid on olulised, et oleks võimalik sooritada täpset aruandlust ja reprodutseerimist. Biomeditsiinilistes uuringutes on sellised standardid kriitilise tähtsusega. Näiteks viiruslikud RefSeq dokumendid läbivad NCBI-s

kõigepealt taksonoomilise kontrolli, et neid oleks võimalik üldse avalikustada (O'Leary et al., 2016).

Andmebaasi täielikkus ja täpsus mõjutavad otseselt järjestuste määramist, mis põhinevad järjestuste identsusel andmebaasis olevate viiruste järjestustega. Viiruste referentsandmebaasi ehk RVDB (Goodacre et al., 2018) välja töötamisel püüti luua kõikehõlmav viiruste andmebaas, mis põhineb kindlatel valikukriteeriumidel, et kaasata kõiki viiruslikke, viiruslaadseid ja viirusega seotud järjestusi (Fabiańska et al., 2021). Teine näide on Virosauruse andmebaas, mis on loodud selleks, et pakkuda automatiseeritud ja annoteeritud andmebaasi kliiniliste metagenoomika uuringute ja diagnoosimise jaoks. Virosauruse andmebaas sisaldab eukarüootide viirusjärjestusi, mis on kompaktsemaks muudetud 90% või 98% identsusega rühmitamise teel, et eemaldada redundantsed järjestused (Fabiańska et al., 2021). Rakenduse tööks on kasutatud GenBank-ist saadud toorjärjestusi, mis on puhastatud, et eemaldada potentsiaalseslt vigased järjestused (Gleizes et al., 2020).

5. METAGENOOMSE ANALÜÜSIGA SEOTUD PROBLEEMID

Maailma rahvaarvu pidev kasv toob kaasa suurema nõudluse põllumajandussaaduste ja vee järele, mistõttu on reovee korduvkasutamine vajalik. See toob endaga kaasa ohu keskkonna viirusreostuseks ja mõjutab ka viiruste laialdast levikut. Viiruste leviku ja mitmekesisuse uurimine reoveest on pädev meetod kogukonnas ringlevate viiruste jälgimiseks, kuid reoveeproovide viroomi kirjeldamise teeb keeruliseks see, et vees on suur hulk mikroorganisme, mida on uuritavas proovis sageli rohkem kui viiruseid. Just puhangute vaheaegadel esineb proovis sageli baktereid ja faage rohkem, kui inimeste seedetrakti kaudu levivaid viiruseid (Strubbia et al., 2019). Sageli on keeruline saada reoveest esinduslik proov viiruskoormuse ja mitmekesisuse kohta. Proove kogutakse küll suurest kogusest veest, et tõsta analüüside tundlikkuse lävendit, kuid see võib kaasa tuua analüüsides kasutatavate ensüümide inhibiitorite kontsentratsiooni kasvu (Strubbia et al., 2019). Vees võib leiduda näiteks ravimeid, pesemisvahendeid, toidujäätmeid ja muid kemikaale, mis nõuab omakorda vee puhastamise etappe.

Siiani leitakse seirete abil sageli viiruseid, mille esinemise kohta keskkonnas ja toiduahelas ei ole piisavalt informatsiooni ehk võib esineda raskusi seireproovide korrektsete tulemuste tõlgendamisel (Nieuwenhuijse & Koopmans, 2017). Selleks, et oleks võimalik paremini mõista viiruste põhjustatud otseseid ja kaudseid mõjusid peremeesorganismile, on vaja täiustada meetodeid viiruste tuvastamiseks.

Hetkel ei ole rakendatud standardeid näiteks selle kohta, et kuidas pärast viiruste tuvastamist tõlgendatakse kontiigid nn liigitasandil järjestusteks, mida oleks võimalik kasutada järgnevateks analüüsideks. Viirusanalüüsi standardite puudumine võib osaliselt seletada ka uuringutevahelisi väga erinevaid (14%-87%) viiruste avastamise määrasid (Mirzaei & Maurice, 2017). Lisaks sellele võivad erinevused analüüsimisel tekkida ka proovide töötlemisel, nende alaesindatusel võrdlusandmebaasides, indiviidivahelisel varieeruvusel jne (Gregory et al., 2020).

Metagenoomika puhul on kriitilise tähtsusega ka uute bioinformaatiliste lähenemisviiside välja arendamine, et viiruste järjestusi oleks võimalik tõhusamalt tuvastada. Tehnoloogilised edusammud raamatukogude ettevalmistamise meetodite, järjestuste genereerimise ja arvutusliku bioinformaatika vallas võimaldavad kiiremaid ja põhjalikumaid metagenoomilisi analüüse madalamate kuludega (Chiu & Miller, 2019). Mitmed metagenoomilised uuringud

põhinevad kaubanduslikel kergesti kättesaadavatel komplektidel, sest need on kiired ning kergesti kohandatavad olemasolevate platvormidega (Fidler et al., 2020).

Metagenoomilise sekveneerimise igapäevasel rakendamisel on ühest küljest olulised kulud, kuid kindlasti ka meetodi kiirus, usaldusväärsus ja tulemuste võimalik võrdlemine strandardmeetoditega. Viirusjärjestuste pidev kureerimine ja laiendamine avalikes andmebaasides toetab lähitulevikus HTS-põhiste viiruste avastamist (Fabiańska et al., 2021). Esimese suure läbilaskevõimega sekveneerimisplatvormi kasutuselevõtuga 2000. aastate keskel kaasnes ka sekveneerimiskulude langus, kuid suure läbilaskevõimega meetoditega kaasnevad ka suured andmemahud, mis võivad tähendada vigaste annotatsioonide kasvu. Uuringute puhul on väga oluline osata tõlgendada annoteeritud järjestusi (Goodwin et al., 2016).

ARUTELU

Metagenoomika abil on tuvastatud rohkelt viiruseid, kuid viiruste taksonoomiat võiks aina edasi uurida ja analüüsimeetodeid vastavalt ka täiendada. Arvatakse, et valdav enamus viiruste taksonitest on veel kirjeldamata. Viirused on väga mitmekesised ning võimelised muteeruma, mis teeb nende analüüsimise keeruliseks. Maailmas kasvab pidevalt ka piirkondade asustustihedus, mis tekitab veelgi soodsamad tingimused viiruste levikule.

Eesmärk haigustekitajate ära tundmisel ja teadlikkuse tõstmisel, on panustada rohkem ressursse teaduse arengusse. Viiruste laialdasteks uuringuteks on oluline koguda seireandmeid, mille abil on võimalik saada informatsiooni kõikide elanike kohta mingis kindlas piirkonnas. Nakkushaiguste seired põhinevad sageli üldlevinumatel patogeenidel ehk see annab inimkonnas levivatest haigustest vaid piiratud ülevaate (Aarestrup et al., 2012). Sageli tekitab analüüsil probleeme just vähelevinud viiruste tuvastamine ehk oluline samm valdkonna arengus on laiendada seireid ka viirustele, mis ei ole veel inimkonnas nii laialtlevinud. Seirete puhul võiks kasutada koos metagenoomilisi sekveneerimisteholoogiaid ja bioinformaatilisi analüüsivahendeid, et saada võimalikult realistlik ülevaade toidu- ja vee kaudu levivate viirushaiguste levikust ja selle mõjust inimestele.

Oluline samm DNA viirusanalüüsi ja selle meetodite edasi arendamisel on olnud teise põlvkonna sekveneerimistehnoloogiate areng. Sekveneerimistehnoloogiaid täiustatakse pidevalt ja see aitab kaasa analüüsi kvaliteedi ja saagikuse paranemisele. Tulevikus on oodata teise põlvkonna sekveneerimise veel laialdasemat kasutamist ja kindlasti ka võimaluste suurenemist. Edasiminekud kuludes, läbilaskevõimes ja juurdepääsetavuses viivad uute tehnoloogiate tekkimiseni, mis aitavad lahendada sekveneerimisel ja selle analüüsil tekkivaid probleeme (Goodwin et al., 2016). Sekveneerimisandmete analüüsi abil on võimalik tuvastada erinevaid olemasolevaid ja uusi viiruseid uuritavast keskkonnast. Suure läbilaskevõimega sekveneerimise puhul nõuab andmete analüüs tugevat ja pidavat infrastruktuuri, mis omakorda tekitab vajaduse pideva ekspertiisi ja investeeringute järele (Spjuthetal.,2016).

Lisaks, võib reaalajas sekveneerimine olla kliinilises meditsiinis paljumuutev tehnoloogia, sest laboritel on võimalus kasutada neid vahendeid ebaharilike nakkuste diagnoosimiseks ja patogeenide puhangute jälgimiseks. Seda näitas ka reaalajas nanopoorse sekveneerimistehnoloogia kasutamine Ebola ja ZIKV epidemioloogilise kaugseire jaoks (Chiu & Miller, 2019). Nanopoorne sekveneerimine võimaldab DNA või RNA reaalajas analüüsimist.

See toimib jälgides elektrivoolu muutuseid, kui DNA/RNA oligonukleotiidid läbivad nanopoori. Saadud signaal dekodeeritakse, et anda konkreetne DNA või RNA järjestus (Sun et al., 2020).

Selleks, et järjestusandmeid töödelda või andmebaasidest võrdlemiseks järjestusi otsida, on sageli kasutusel erinevad bioinformaatilised tööriistad. Järjestusandmete klassifitseerimiseks on levinud Kraken 2 tööriist ning viiruse tuvastamiseks suure läbilaskevõimega sekveneerimisandmetest on olemas LABRADOR-i töövoog. Andmete otsimiseks viiruste, nende liikide ja erinevate tüvede kohta kasutatakse erinevaid andmebaase. Näiteks NCBI *Taxonomy* sisaldab suurt hulka organismide nimesid ja klassifikatsioone ning RefSeq sisaldab kureeritud referentsjärjestusi. Usaldusväärsed ja hinnatud andmebaasid on olulised, et viirusjärjestuste annotatsioonid oleksid korrektsed ning neid oleks võimalik ka tulevastes uuringutes kasutada. Bioinformaatika töövoogude puhul on oluline kasutajasõbralikkus ja mNGS (*metagenomic next generation sequencing*) andmete analüüsiks on vajalik ka personal, kes oskab arendada, valideerida ja hooldada töövoogusid kliiniliseks kasutamiseks. Tarkvara seadistamine on keeruline ning nõuab asjakohaste meetmete rakendamist konfidentsiaalsete andmete ja teabe kaitsmiseks (Chiu & Miller, 2019).

Vaatamata metagenoomika potentsiaalile on kliinilised diagnostilised rakendused jäänud teadusuuringute edusammudest maha. Meetodi puhul ei ole universaalseid võrdlusstandardeid ning lisaks võiks arendada lähenemisviise kliiniliste metagenoomiliste testide valideerimise, reprodutseeritavuse ja kvaliteedi tagamise tõendamiseks. Kliiniline kasulikkus ning kulud mängivad samuti suurt rolli mNGS rutiinsel rakendamisel raviasutustes (Chiu & Miller, 2019). Teadlike otsuste abil võib olla võimalik vähendada viiruskoormust inimkonnale.

KOKKUVÕTE

Käesolevas referatiivses töös on antud ülevaade teadustöödest, mis on seotud viiruste detekteerimisega reoveest. Reovee kaudu levivad paljud viirused, nende hulgas ka inimestele patogeensed viirused, mille tuvastamine on kriitilise tähtsusega. Reoveest proovide kogumise abil on võimalik tuvastada viiruseid ning jälgida nende epidemioloogilist levikut.

Kirjeldatud on nii DNA eraldamise kui bioinformaatilise analüüsiga seotud meetodeid ning tarkvara. Eraldi lõiguna on töös refereeritud erinevate riikide seniseid kogemusi ja teadmisi, mis on saadud koroona-, rota-, noro- ja enteroviiruste reovee monitooringus. Viiruste analüüsil on esimene samm viiruste tuvastamine. Tuvastamismeetodeid on erinevaid, kuid laialdaselt kasutatakse viiruste tuvastamisel PCR meetodit või suure läbilaskevõimega sekveneerimist. PCR analüüsimeetod on kõrge tundlikkusega, kuid selle abil on võimalik tuvastada vaid kindel patogeen, mille järjestusandmed on teada ja mida antud analüüsi käigus otsitakse. Suure läbilaskevõimega sekveneerimise abil on võimalik tuvastada ka selliseid viiruseid, mille nukleiinhappe järjestus ei ole enne teada ehk meetodi abil on tuvastatud ka uusi viiruseid. Viiruste tuvastamisel on oluline ka vajadusel teostada puhastusprotsessid ning muuta selle abil viiruse tuvastamine täpsemaks ja tõhusamaks.

Võrdlusandmestike kasutamine ja saadud analüüsitulemuste uurimine on kogu muu protsessi osa ning kuna andmemahud on suured ja analüüsi kiirus on samuti oluline, siis kasutatakse enamasti nende protsesside jaoks bioinformaatilist tarkvara ja andmebaase. On loodud erinevaid töövoogusid ja andmebaase, mida arendatakse ja täiendatakse pidevalt, et need oleksid asjakohased.

Kogu metagenoomika valdkond areneb pidevalt ning kiire tehnoloogia arenguga muutuvad ka nukleiinhapete analüüsimeetodid. See annab teadlastele võimaluse kasutada aina paremaid ja täiustatumaid meetodeid, mis lõpuks mõjutab positiivselt inimeste tervislikku heaolu.

SUMMARY

Detection of viruses in wastewater samples

Kristi-Rebeka Tiits

Summary

There are many enteric viruses in the world that are pathogenic to humans. It is therefore important to find ways to detect and analyse these viruses. Viruses are highly diverse and capable of mutating, which makes their analysis more difficult. It is believed that the vast majority of viral taxa have yet to be described. A large number of viruses have been identified using metagenomics, but the taxonomy of viruses could be further explored and the methods of analysis improved accordingly. The world is also experiencing a steady increase in population density in regions, which creates even more favourable conditions for the spread of viruses.

Many viruses, including viruses pathogenic to humans, are spread through sewage. For example, rotaviruses, noroviruses and coronaviruses. These viruses affect human health, which makes it important to identify those viruses and achieve control over their spread. The collection of wastewater samples, which can be used to detect viruses, is one possible method to monitor the epidemiological spread of viruses. To analyse viruses, they must first be identified. PCR is a sensitiive method, but it can only identify the pathogen for which the sequence data are known and which is being searched for in this analysis. High-throughput sequencing can also be used to identify viruses whose nucleic acid sequence is not yet known.

Various bioinformatics tools are often used to process sequence data or to search for sequences in databases for comparison. For example, the Kraken 2 tool is commonly used to classify sequence data, and LABRADOR is available for virus identification from high-throughput sequencing data. Different databases are used to search for data on viruses, their species and different strains. For example, the NCBI Taxonomy contains a large number of organism names and classifications and RefSeq contains curated reference sequences. Reliable and valued databases are essential to ensure that annotations of viral sequences are accurate and can be used in future research.

Different DNA-based methods of analysis are very important for the detection of viruses. The field of metagenomics is constantly evolving, and with the rapid development of technology, the methods of nucleic acid analysis are also changing. This will allow researchers to use ever better and more advanced methods, which will ultimately have a positive impact on people's health.

TÄNUSÕNAD

Soovin tänada oma juhendajat Maido Remmi, kes oli väga abivalmis ja leidis alati aega, et töös sisalduvat arutada ja analüüsida. Samuti soovin tänada Reidar Andresoni ja Age Brauerit, kes olid abiks programmide kasutamisel ning bioinformaatika algtõdede tutvustamisel. Lisaks veel tänusõnad kogu bioinformaatika õppetoolile toreda koostöö ja seminaride eest.

KASUTATUD KIRJANDUSE LOETELU

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389–3402.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic acids research*, *41*(D1), D36–D42.

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in bioinformatics*, *20*(4), 1125–1136.

Brister, J. R., Ako-Adjei, D., Bao, Y., & Blinkova, O. (2015). NCBI viral genomes resource. *Nucleic acids research*, *43*(D1), D571–D577.

Chiu, C. Y., & Miller, S. A. (2019). Clinical metagenomics. *Nature Reviews Genetics*, *20*(6), 341–355.

Cochrane, G., Karsch-Mizrachi, I., Takagi, T., & Sequence Database Collaboration, I. N. (2016). The international nucleotide sequence database collaboration. *Nucleic acids research*, *44*(D1), D48–D50.

Consortium, T. U. (2011). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, *40*(D1), D71–D75. https://doi.org/10.1093/nar/gkr981

Corpuz, M. V. A., Buonerba, A., Vigliotta, G., Zarra, T., Ballesteros Jr, F., Campiglia, P., Belgiorno, V., Korshin, G., & Naddeo, V. (2020). Viruses in wastewater: Occurrence, abundance and detection methods. *Science of the Total Environment*, *745*, 140910.

Daughton, C. G. (2020). Wastewater surveillance for population-wide Covid-19: The present and future. *Science of the Total Environment*, *736*, 139631.

Fabiańska, I., Borutzki, S., Richter, B., Tran, H. Q., Neubert, A., & Mayer, D. (2021). LABRADOR—A Computational Workflow for Virus Detection in High-Throughput Sequencing Data. *Viruses*, *13*(12), 2541.

Fidler, G., Tolnai, E., Stagel, A., Remenyik, J., Stundl, L., Gal, F., Biro, S., & Paholcsek, M. (2020). Tendentious effects of automated and manual metagenomic DNA purification protocols on broiler gut microbiome taxonomic profiling. *Scientific reports*, *10*(1), 1–16.

Flygare, S., Simmon, K., Miller, C., Qiao, Y., Kennedy, B., Di Sera, T., Graf, E. H., Tardif, K. D., Kapusta, A., Rynearson, S., & others. (2016). Taxonomer: An interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome biology*, *17*(1), 1–18.

Foladori, P., Cutrupi, F., Segata, N., Manara, S., Pinto, F., Malpei, F., Bruni, L., & La Rosa, G. (2020). SARS-CoV-2 from faeces to wastewater treatment: What do we know? A review. *Science of the Total Environment*, *743*, 140444.

Gleizes, A., Laubscher, F., Guex, N., Iseli, C., Junier, T., Cordey, S., Fellay, J., Xenarios, I., Kaiser, L., & Mercier, P. L. (2020). Virosaurus a reference to explore and capture virus genetic diversity. *Viruses*, *12*(11), 1248.

Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M., & Khan, A. S. (2018). A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. *MSphere*, *3*(2), e00069-18.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351.

Greenberg, H. B., & Estes, M. K. (2009). Rotaviruses: From pathogenesis to vaccination. *Gastroenterology*, *136*(6), 1939–1951.

Gregory, A. C., Zablocki, O., Zayed, A. A., Howell, A., Bolduc, B., & Sullivan, M. B. (2020). The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell host & microbe*, *28*(5), 724–740.

Hellmér, M., Paxéus, N., Magnius, L., Enache, L., Arnholm, B., Johansson, A., Bergström, T., & Norder, H. (2014). Detection of pathogenic viruses in sewage provided early warnings of hepatitis A virus and norovirus outbreaks. *Applied and environmental microbiology*, *80*(21), 6771–6781.

Hjelmsø, M. H., Hellmér, M., Fernandez-Cassi, X., Timoneda, N., Lukjancenko, O., Seidel, M., Elsässer, D., Aarestrup, F. M., Löfström, C., Bofill-Mas, S., & others. (2017). Evaluation of methods for the concentration and extraction of viruses from sewage in the context of metagenomic sequencing. *PloS one*, *12*(1), e0170199.

Ivanova, O. E., Yarmolskaya, M. S., Eremeeva, T. P., Babkina, G. M., Baykova, O. Y., Akhmadishina, L. V., Krasota, A. Y., Kozlovskaya, L. I., & Lukashev, A. N. (2019). Environmental surveillance for poliovirus and other enteroviruses: Long-term experience in Moscow, Russian Federation, 2004–2017. *Viruses*, *11*(5), 424.

Khan, A. S., Blümel, J., Deforce, D., Gruber, M. F., Jungbäck, C., Knezevic, I., Mallet, L., Mackay, D., Matthijnssens, J., O'Leary, M., & others. (2020). Report of the second international conference on next generation sequencing for adventitious virus detection in biologics for humans and animals. *Biologicals*, *67*, 94–111.

Kittigul, L., & Pombubpa, K. (2021). Rotavirus surveillance in tap water, recycled water, and sewage sludge in Thailand: A longitudinal study, 2007–2018. *Food and Environmental Virology*, *13*(1), 53–63.

Koo, E. S., Kim, M. S., Choi, Y. S., Park, K.-S., & Jeong, Y. S. (2017). Occurrence of novel GII. 17 and GII. 21 norovirus variants in the coastal environment of South Korea in 2015. *PLoS One*, *12*(2), e0172237.

Li, W., O'Neill, K. R., Haft, D. H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., Chitsaz, F., Derbyshire, M. K., Durkin, A. S., & others. (2021). RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic acids research*, *49*(D1), D1020–D1028.

Lin, H.-H., & Liao, Y.-C. (2017). DrVM: a new tool for efficient genome assembly of known eukaryotic viruses from metagenomes. *Gigascience*, *6*(2), gix003.

Lin, J., Kramna, L., Autio, R., Hyöty, H., Nykter, M., & Cinek, O. (2017). Vipie: Web pipeline for parallel characterization of viral populations from multiple NGS samples. *BMC genomics*, *18*(1), 1–11.

Mackay, I. M., Arden, K. E., & Nitsche, A. (2002). Real-time PCR in virology. *Nucleic acids research*, *30*(6), 1292–1305.

Majumdar, M., Klapsa, D., Wilton, T., Akello, J., Anscombe, C., Allen, D., Mee, E. T., Minor, P. D., & Martin, J. (2018). Isolation of vaccine-like poliovirus strains in sewage samples from the United Kingdom. *The Journal of Infectious Diseases*, *217*(8), 1222–1230.

Matsushima, Y., Ishikawa, M., Shimizu, T., Komane, A., Kasuo, S., Shinohara, M., Nagasawa, K., Kimura, H., Ryo, A., Okabe, N., & others. (2015). Genetic analyses of GII. 17 norovirus strains in diarrheal disease outbreaks from December 2014 to March 2015 in Japan reveal a novel polymerase sequence and amino acid substitutions in the capsid region. *Eurosurveillance*, *20*(26), 21173.

Mirzaei, M. K., & Maurice, C. F. (2017). Ménage à trois in the human gut: Interactions between host, bacteria and phages. *Nature Reviews Microbiology*, *15*(7), 397–408.

Moreira, N., & Bondelind, M. (2017). Safe drinking water and waterborne outbreaks. *Journal of water and health*, *15*(1), 83–96.

Naccache, S. N., Federman, S., Veeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., Bouquet, J., Greninger, A. L., Luk, K.-C., Enge, B., & others. (2014). A cloud-

compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome research*, *24*(7), 1180–1192.

Nieuwenhuijse, D. F., & Koopmans, M. P. (2017). Metagenomic sequencing for surveillance of food-and waterborne viral diseases. *Frontiers in Microbiology*, *8*, 230.

Nieuwenhuijse, D. F., Oude Munnink, B. B., Phan, M. V., Munk, P., Venkatakrishnan, S., Aarestrup, F. M., Cotten, M., & Koopmans, M. P. (2020). Setting a baseline for global urban virome surveillance in sewage. *Scientific reports*, *10*(1), 1–13.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., & others. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, *44*(D1), D733–D745.

Padilla-Reyes, D. A., Álvarez, M. M., Mora, A., Cervantes-Avilés, P. A., Kumar, M., Loge, F. J., & Mahlknecht, J. (2022). Acquired insights from the long-term surveillance of SARS-CoV-2 RNA for COVID-19 monitoring: The case of Monterrey Metropolitan Area (Mexico). *Environmental Research*, *210*, 112967.

Plyusnin, I., Kant, R., Jääskeläinen, A. J., Sironen, T., Holm, L., Vapalahti, O., & Smura, T. (2020). Novel NGS pipeline for virus discovery from a wide spectrum of hosts and sample types. *Virus evolution*, *6*(2), veaa091.

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). VirFinder: A novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, *5*(1), 1–20.

Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Lavigne, R., Brister, J. R., Varsani, A., & others. (2019). Minimum information about an uncultivated virus genome (MIUViG). *Nature biotechnology*, *37*(1), 29–37.

Roux, S., Enault, F., Hurwitz, B. L., & Sullivan, M. B. (2015). VirSorter: Mining viral signal from microbial genomic data. *PeerJ*, *3*, e985.

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., & others. (2020). NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database*, *2020*.

Simmonds, P., Adams, M. J., Benk\Ho, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davison, A. J., Delwart, E., Gorbalenya, A. E., Harrach, B., & others. (2017). Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, *15*(3), 161–168.

Strubbia, S., Phan, M. V., Schaeffer, J., Koopmans, M., Cotten, M., & Le Guyader, F. S. (2019). Characterization of norovirus and other human enteric viruses in sewage and stool samples through next-generation sequencing. *Food and environmental virology*, *11*(4), 400–409.

Sun, X., Song, L., Yang, W., Zhang, L., Liu, M., Li, X., Tian, G., & Wang, W. (2020). Nanopore sequencing and its clinical applications. *Precision Medicine*, 13–32.

Zhao, G., Wu, G., Lim, E. S., Droit, L., Krishnamurthy, S., Barouch, D. H., Virgin, H. W., & Wang, D. (2017). VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology*, *503*, 21–30.

Wolf, Y. I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kazlauskas, D., Krupovic, M., Fire, A., Dolja, V. V., & Koonin, E. V. (2020). Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nature microbiology*, *5*(10), 1262–1270.

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome biology*, *20*(1), 1–13.

Wu, Y., Guo, C., Tang, L., Hong, Z., Zhou, J., Dong, X., Yin, H., Xiao, Q., Tang, Y., Qu, X., & others. (2020). Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. *The lancet Gastroenterology & hepatology*, *5*(5), 434–435.

Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., Niemeyer, D., Jones, T. C., Vollmar, P., Rothe, C., & others. (2020). Virological assessment of hospitalized patients with COVID-2019. *Nature*, *581*(7809), 465–469.

Yamashita, A., Sekizuka, T., & Kuroda, M. (2016). VirusTAP: viral genome-targeted assembly pipeline. *Frontiers in microbiology*, 32.

LIHTLITSENTS

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kristi-Rebeka Tiits

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

Viiruste detekteerimine reoveest, mille juhendaja on Maido Remm,

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kristi-Rebeka Tiits
26.05.2022