TARTU ÜLIKOOL
LOODUS- JA TÄPPISTEADUSTE VALDKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

Erinevate töövoogude analüüs viiruste tuvastamisel

Bakalaureusetöö
12 EAP

Brigitta-Robin Raudne

Juhendajad:
MSc Mikk Puustusmaa
MSc Mihkel Vaher

TARTU 2018

Erinevate töövoogude analüüs viiruste tuvastamisel

Viirused on väikesed rakusisesed parasiidid, mis vajavad paljunemiseks peremeesorganismi – olgu selleks ühe- või hulkraksed organismid. Viirused on väga mitmekesised ning uusi viiruseid avastatakse pidevalt juurde. Käesoleva töö eesmärgiks on analüüsida erinevaid programme ja töövooge, mis on loodud uute viiruste avastamiseks ja teadaolevate viiruste tuvastamiseks proovidest. Erinevate positiivsete ja negatiivsete külgede välja toomine aitab leida sobivaimat töövoogu viiruste detekteerimiseks. Viiruste tuvastamine erinevatest keskkondadest aitab meil koguda täiendavaid andmeid selliste viiruste kohta, mida oleks võimalik rakendada ka meditsiinis.

Märksõnad: viirused, klassifikatsioon, tuvastamine, avastamine, sekveneerimine, k-meer
CERCS: B110 (Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika)

Various pipeline analysis for virus detection

Viruses are small intracellular parasites that require a host for replication – be it a unicellular or a multicellular organism. Viruses are very diverse life forms and new viruses are constantly discovered. The purpose of this work is to analyze various programs and pipelines designed to discover new viruses and detect known viruses from samples. Bringing out the various positive and negative sides will help finding the most suitable workflow for detecting viruses. Detection of viruses from different environments will help us to gather additional data on viruses that could be applied in medicine as well.

Keywords: viruses, classification, detection, discovery, sequencing, k-mer
CERCS: B110 (Bioinformatics, medical informatics, biomathematics, biometrics)

SISUKORD

KASUTATUD LÜHENDID

ELISA	*enzyme-linked immunosorbent assay*	immunoensüümmeetod
EM	*electron microscopy*	elektronmikroskoopia
EVE	*endogenous viral element*	endogeenne viiruslik element
FISH	*fluorescent in situ hybridization*	fluorestsents *in situ* hübridisatsioon
GVP	*Global Virome Project*	ülemaailmne viroomi projekt
HBV	*hepatitis B virus*	B-hepatiidi viirus
HCV	*hepatitis C virus*	C-hepatiidi viirus
ICTV	*International Committee on Taxonomy of Viruses* Rahvusvaheline viiruse taksonoomia komitee	
lm	*reads per minute*	lugemit minutis
MDA	*multiple displacement amplification*	mitmekordne nihke amplifikatsioon
MS	*mass spectrometry*	massispektomeetria
NGS	*next-generation sequencing*	teise põlvkonna sekveneerimine
STP	*sewage treatment plant*	reoveepuhasti
SVG	*single virus genomics*	ühe viiruse genoomika
vSAG	*viral single-amplified genomes*	ühe viiruse amplfitseeritud genoom

SISSEJUHATUS

Viirused on nukleiinhappest ja valkudest koosnevad bioloogilised üksused, mille suurus jääb enamasti vahemikku 20 kuni 120 nm. Viirused on rakusisesed parasiidid ja vajavad paljunemiseks peremeesorganismi. Viiruseid leidub igas keskkonnas ning nende populatsioon ja mitmekesisus on väga suur. Laialdane levik meie planeedil võimaldab viirustel mängida "looduslike mootorite" rolli, juhtides ülemaailmset energia ja toitainete ringlust kontrollides nii bakterite, ainuraksete kui ka hulkraksete populatsioone. Viirused on ka inimese mikrobioomi lahutamatuks osaks.

Maal arvatakse olevat üle 100 miljoni erineva viiruse liigi. Osad nendest on patogeensed, põhjustades haigusi, mis võivad halvimal juhul lõppeda surmaga. Seetõttu on oluline nii uute viiruste uurimine kui ka viiruste tuvastamine erinevatest keskkondadest.

Viiruste tuvastamiseks ja uute viiruste avastamiseks on välja töötatud mitmeid meetodeid.

Üldiselt saab need jagada kolme rühma:

1. viiruse osakeste avastamine (nende kindlaks tegemine, nende kontsentratsiooni määramine, suurused ja muud füüsikalised ja keemilised omadused) või viiruse nakkavuse kindlaks tegemine (viroloogiline uuring)
2. otsene indikatsioon viiruse antigeenide proovide näidistele (immunoloogilised meetodid)
3. viiruse nukleiinhapete analüüs (molekulaar-geneetilised meetodid).

Antud töö keskendub peamiselt viimasele rühmale, sest sekveneerimisandmete analüüs võimaldab lisaks teadaolevate viiruste tuvastamisele ka uute viiruste avastamist erinevatest keskkondadest. Sekveneerimise andmete analüüsiks on loodud erinevaid bioinformaatilisi programme, millega saab viiruseid klassifitseerida ja identifitseerida. Programmid jaotatakse kõige üldisemalt kahte gruppi: kahe järjestuse võrdlusel (BLAST) põhinevad ja k-meeridel põhinevad. Mõlemal lähenemisel on nii positiivseid kui ka negatiivseid külgi.

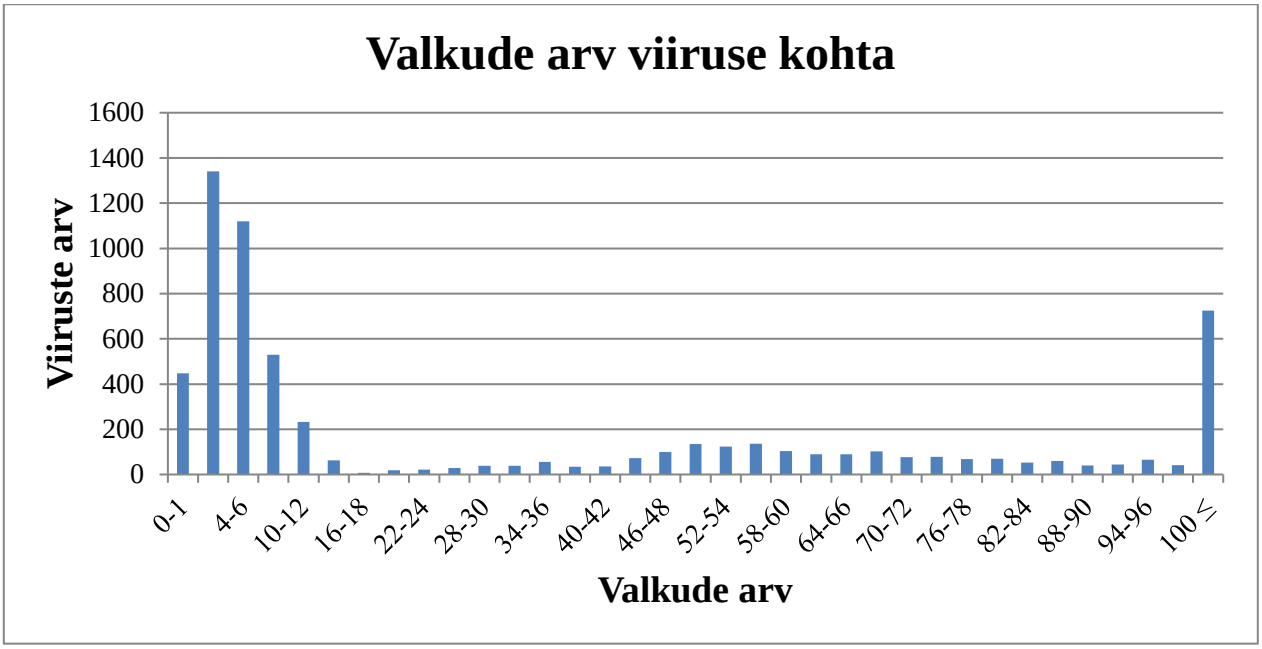
KIRJANDUSE ÜLEVAADE

1. VIIRUSTE ÜLEVAADE

Viirused koosnevad nukleiinhappest ja valkudest ning replitseeruvad ainuraksetes või hulkraksete organismide rakkudes (Koonin *et al*., 2006). Viirused koosnevad kahest või kolmest elemendist: (1) geneetiline materjal (DNA või RNA); (2) geneetilist materjali ümbritsev valgukate, mida nimetatakse kapsiidiks; (3) mõnel juhul ümbritseb kapsiidi veel peremeesraku päritoluga lipiidne membraan. Viiruslikud kapsiidid on sümmeetrilised nanokonteinerid, mille läbimõõt on tavaliselt vahemikus 20 kuni 120 nm. Paljudel juhtudel koosnevad kapsiidid ainult ühe valgu koopiatest (Hu *et al*., 2008). Erandina saab välja tuua suured viirused nagu pandooraviirus ja pithoviirus, mille suurus ületab ühte mikromeetrit (Legendre *et al*., 2014).

Ajalooliselt olid vene bioloog Dmitri Ivanovski ja Hollandi botaanik Martinus Willem Beijerinck esimesed, kes üksteisest sõltumata eraldasid esmakordselt tubaka mosaiikhaigust põhjustanud tubakaspetsiifilise patogeeni. Ivanovski tõestas, et nakatunud tubakalehtede ekstrakt jäi nakkuslikuks isegi pärast filtreerimist, mis tavaliselt eraldab bakterid uuritavast proovist. Esialgu arvati, et nakkust võib põhjustada bakteriaalne toksiin, kuid hiljem jõudis Beijerinck järeldusele, et tegemist on uue patogeeniga, mis vajab elusate rakkude replikatsiooni ja paljunemist ning hakkas nimetama seda viiruseks. (Bos, 1999)

Üldiselt kodeerivad viiruse genoomid suhteliselt vähe valke. Suurem osa uuritud viirustest kodeerib ligikaudu 1-12 valku (joonis 1). Vaatamata sellele petlikule lihtsusele on viiruste replikatsiooni ja biokeemilised mehhanismid väga mitmekesised (Zeigler Allen *et al*., 2017). Selline mitmekesisus võimaldab viirustel nakatada kõiki elusorganisme, nii baktereid, arhesid kui ka eukarüoote ning neid leidub kõigis ökoloogilistes niššides (Haynes and Rohwer, 2011). Laialdane levik meie planeedil võimaldab viirustel mängida "looduslike mootorite" rolli, mis juhivad ülemaailmset energia ja toitainete ringlust kontrollides nii bakterite, ainuraksete kui ka hulkraksete populatsioone (Suttle, 2007).



Joonis 1. Kodeeritavate valkude arv viiruse kohta. Viiruse genoome on Uniproti andmebaasis (*UniProt reference proteomes*) kokku 6279, mis on vastavalt nende poolt kodeeritavate valkude arvu põhjal jaotatud rühmadesse. (http://www.uniprot.org/proteomes/?query=*&fil=reference%3Ayes+AND+taxonomy%3A%22Viruses+%5B10239%5D%22, 16.05.2018)

1.1. Viiruste arvukus

Erinevate metagenoomide analüüs on näidanud, et viiruste mitmekesisus on veelgi suurem kui varem arvati (Simmonds *et al*., 2017). Ainuüksi selgroogsetel organismidel arvatakse olevat erineviad viiruseid ligikaudu 100 miljonit. Kaasates sinna bakterite, arhede ja teiste ainuraksete organismide viirused tõuseb eeldatav viiruste arv palju suuremaks (Raccaniello, 2013, http://www.virology.ws/2013/09/06/how-many-viruses-on-earth/). Viiruseid leidub kõikides keskkondades, näiteks maailma ookeanides on 1 liitris vees ligikaudu 10^{10} viiruse osakest (Fuhrman, 1999). Avatud ookeanides on kokku ligikaudu 1,2 x 10^{30} ja ookeanilises pinnases ligikaudu 3,5 x 10^{31} viiruslikku osakest (Mokili *et al*., 2012). Viirustest on kõige arvukamad baktereid nakatavad viirused ehk bakteriofaagid. Bakteriofaage on hinnanguliselt 4,8 x 10^{31} partiklit ja nendest 97% esinevad pinnases ja setetes, mis on aga kaks kõige vähem uuritud bioomi – umbes 2,5% avalikult kättesaadavatest viiruslikest metagenoomidest (tabel 1). Kaks kõige rohkem uuritud bioomi – inimese soolestik ja magevesi – annavad suhteliselt väikese panuse viiruste kogu arvukusele. 2016. andmete põhjal on tuvastatud vaid 257 698 viiruse

genotüüpi, mis näitab viiruslike metagenoomsete andmete vähesust. (Cobián Güemes *et al*., 2016)

Tabel 1. Viirusetaoliste osakeste ligikaudne arvukus erinevates bioomides. *(Cobián Güemes et al.*, 2016, osaline)

Bioom	**Prokarüootsete rakkude arv**	**Viirusetaoliste osakeste arv**	**Viirustaoliste osakeste protsent**
Merevesi	$1{,}01 \times 10^{29}$	$1{,}29 \times 10^{30}$	2,683%
Magevesi	$1{,}26 \times 10^{26}$	$1{,}76 \times 10^{27}$	0,004%
Muud veekogud	$2{,}44 \times 10^{27}$	$7{,}32 \times 10^{28}$	0,152%
Setted	$3{,}80 \times 10^{30}$	$4{,}18 \times 10^{31}$	87,013%
Pinnas	$2{,}50 \times 10^{29}$	$4{,}88 \times 10^{30}$	10,148%
Inimene	$2{,}80 \times 10^{23}$	$2{,}80 \times 10^{22}$	0,000%
Kokku	$\mathbf{4{,}15 \times 10^{30}}$	$\mathbf{4{,}80 \times 10^{31}}$	**100%**

1.2. Viiruste süstemaatika

Viiruste klassifikatsioon põhineb erinevate viirust kirjeldavate omaduste (näiteks geneetiline materjal, replikatsiooni mehhanism, kodeeritavad valgud, geeni järjestus, interaktsioonid peremeesorganismidega) kogumisel ja võrdlemisel (King, A. M. Q. *et al.*, 2012). Eelpool toodud omaduste kogumine ja analüüsimine on väga töömahukas. Seega soovitakse tulevikus viiruseid klassifitseerida ka üksnes geneetilise materjali alusel (Simmonds *et al*., 2017). Suure viiruste mitmekesisuse tõttu usuvad osad viroloogid, et viiruste geneetilise spekter on pidev ja ühtlane, milles pole eraldi seisvaid osi ja seega viiruste jagamine erinevatesse kategooriatesse pole mõttekas. (Siddell, 2018, https://microbiologysociety.org/publication/current-issue/imaging/article/why-virus-taxonomy-is-important.html)

1.2.1. Taksonoomiline klassifikatsioon

Organismide taksonoomiline klassifitseerimine põhineb eeldusel, et nende vahel on evolutsiooniline seos ning tänapäeval elavad organismid omavad ühtset päritolu. Kõik rakulised

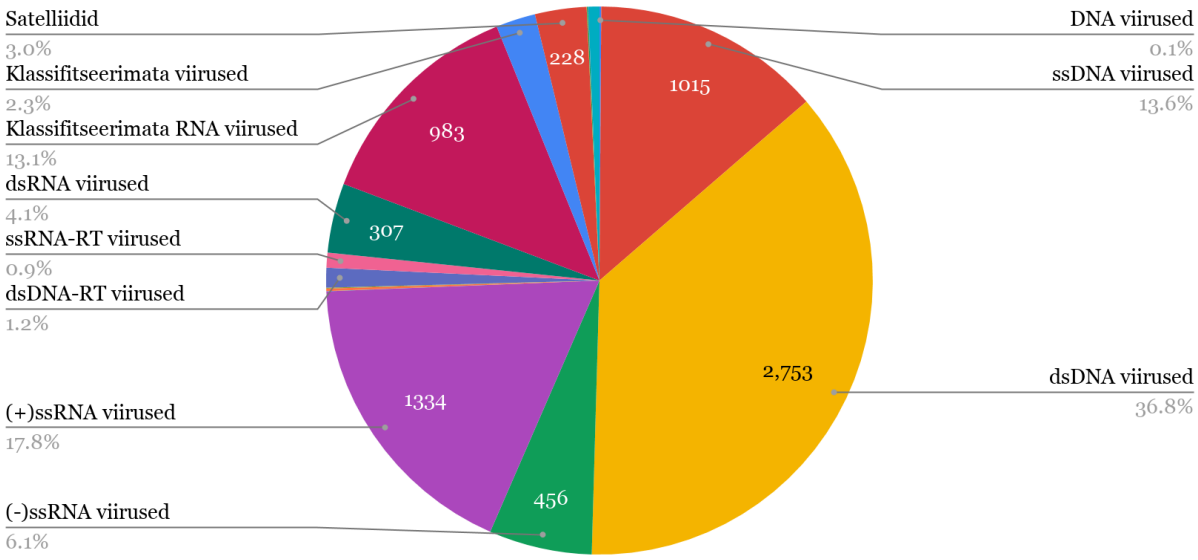
organismid pärinevad viimasest ühisest eellasest LUCA (*last universal common ancestor*) ning moodustavad monofüleetilise rühma. Viiruseid peetakse aga polüfüleetilisteks ehk neil on mitu evolutsioonilist alget ning viirused moodustavad mitu monofüleetilist rühma (Lefkowitz *et al*., 2018). Viiruste rühmitamist alustatakse liikidest. Seejärel kogutakse liigid perekondadesse ja edasi sugukondadesse ja kuni seltsideni, mis on praegu viiruste kõrgeim taksonoomiline tase (https://viralzone.expasy.org/). Viiruste taksonoomia korrashoiu ja täiustamise eest vastutab rahvusvaheline viiruse taksonoomia komitee (ICTV). ICTV ülesanne on ühendada omavahel sarnased viirused suhete hierarhias ja aidata meil mõista viiruste maailma (Lefkowitz *et al*., 2018). Praeguste andmete järgi on viiruste seltse 8 ja sugukondi 125, millest 68% pole veel seltsidesse jaotatud. Viiruste perekondi teatakse kokku 492 (https://viralzone.expasy.org/, https://talk.ictvonline.org/taxonomy/).

1.2.2. Baltimore klassifikatsioon

Üks kõige levinumaid ja enam kasutatavamaid viiruste klassifikatsioone taksonoomilise klassifikatsiooni kõrval on Baltimore klassifikatsioon. Tihti kasutatakse Baltimore klassifikatsiooni üheskoos taksonoomilse klassifikatsiooniga. Antud süsteemi töötas välja ameerika bioloog David Baltimore. Baltimore klassifikatsiooni puhul ei ole evolutsioonilise suguluse eeldust nagu seda on taksonoomilises klassifikatsioonis. Antud süsteem põhineb genoomi tüübil ja mRNA tootmise mehhanismil ning jaotab viirused seitsmesse rühma:

1. Üheahelalise genoomiga DNA viirused (ssDNA)
2. Kaheahelalise genoomiga DNA viirused (dsDNA)
3. Plussahelalised RNA viirused ((+)ssRNA)
4. Miinusahelalised RNA viirused ((-)ssRNA)
5. Kaheahelalise genoomiga RNA viirused (dsRNA)
6. Üheahelalised RNA retroviirused (ssRNA-RT)
7. Kaheahelalised DNA retroviirused (dsDNA-RT)

Retroviiruste eripäraks on pöördtranskriptaasi olemasolu ja nende alla kuuluvad ssRNA-RT ja dsDNA-RT viirused (https://viralzone.expasy.org/). Kõige suurema osa, 36,8% moodustavad üheahelalised DNA viirused (joonis 2). DNA viiruste rohkus on osaliselt tingitud andmete kallutatusest, mida põhjustab DNA-viirustega tehtud metagenoomiliste uuringute ülekaal (Popgeorgiev, Temmam, Raoult, Desnues, 2013).



Joonis 2. Viiruste rühmitumine erinevatesse Baltimore klassidesse. NCBI *Viral Genomes* andmebaasi andmetel on viiruste täielikke genoome 7484 (https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239, 12.03.2018). Mõned DNA viirused (näiteks *pleolipoviridae*) ei ole jagatud ssDNA või dsDNA alla, sest toimub konverteerimine üheahelaliselt genoomil kaheahelaliseks või ei ole täpselt teada, millise genoomiga on tegemist.

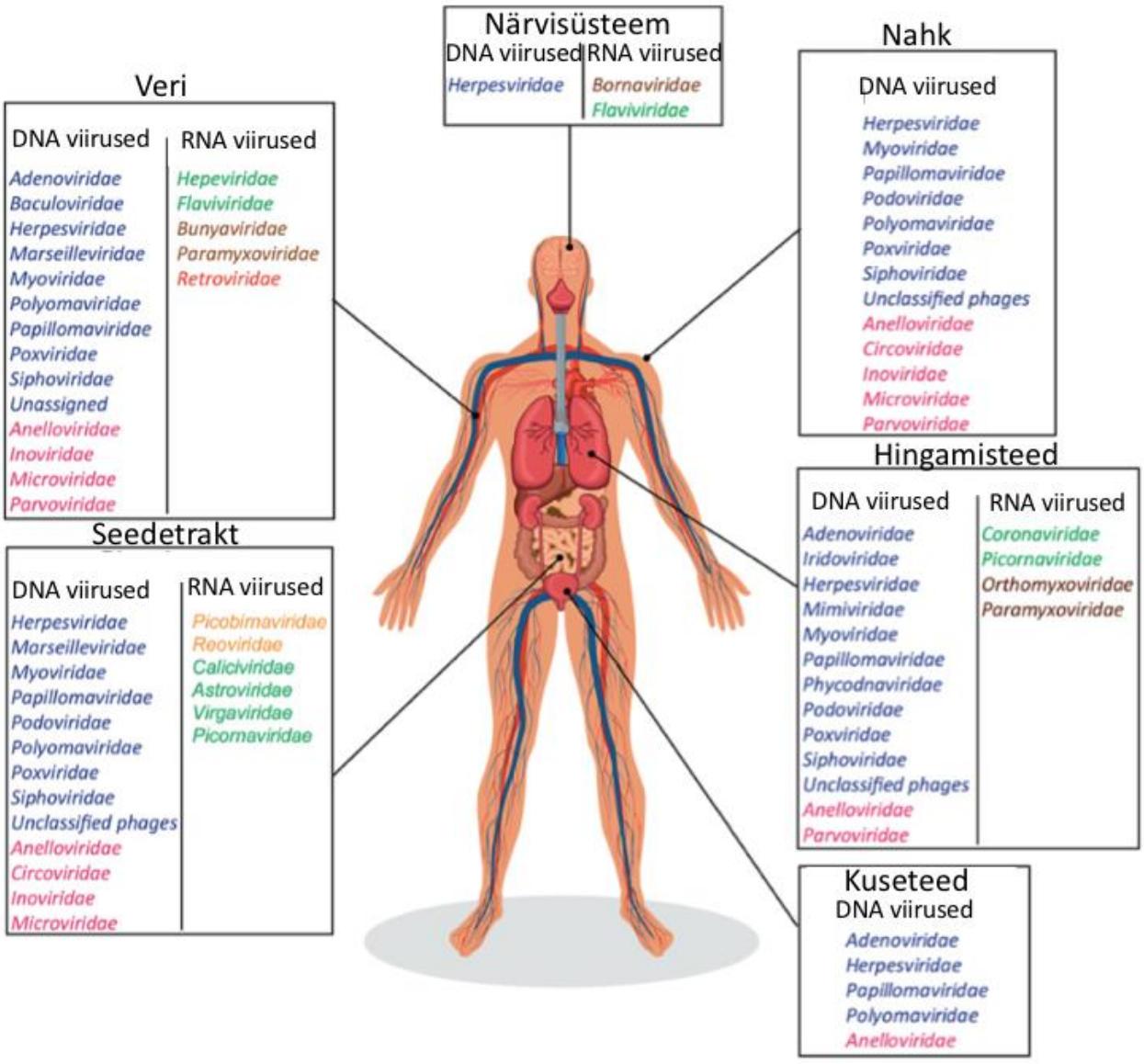
1.3. Inimene ja viirus

Inimese keha pinnal ja soolestikus leidub väga palju erinevaid viiruseid (joonis 3), mille kogumit nimetatakse viroomiks, mis on meie mikrobioomi lahutamatu osa (Ly *et al.*, 2016). Viiruste mitmekesisust eri keha piirkondades mõjutab nii kehaosa mikrokeskkond, kuid ka inimese vanus, toitumine, geograafiline paiknemine/elukoht ja mikrobioomi teiste komponentide olemasolu (Zárate *et al.*, 2017).

Juuksed, nahk ja küüned mängivad olulist rolli barjäärina, kaitstes inimkeha väljastpoolt. Antud mikrokeskkonnad esindavad ka kompleksset ökosüsteemi, mis sisaldab erinevaid bakteri-, seene- ja viiruse liike. Ühel ruutsentimeetril nahal on 10^6 viiruse osakest (Zárate *et al.*, 2017). Tervetest nahaproovidest on enim leitud eukarüootseid DNA viiruseid, näiteks *Circoviridae* sugukonna ssDNA viirused ja *Polyomaviridae* ja *Papillomaviridae* sugukonna dsDNA viirused (Popgeorgiev *et al.*, 2013).

Üha rohkem tõendeid näitavad, et ilmselt tervetel inimestel ei ole veri steriilne ja mõned veres leiduvatest viirustest võivad olla patogeensed. Samuti on ka veri oluline viiruste reservuaar,

täpsemalt 10^5 viirust ühes milliliitris (Zárate *et al*., 2017). Enamik "normaalsetest" vere viirusfloorast sisaldab *Anelloviridae* sugukonna ssDNA viiruseid, millest kõige sagedamini esineb Torque teno viiruseid (Popgeorgiev *et al*., 2013), mis on seotud näiteks hingamisteede haiguste, autoimmuunhaiguste ja hepatiidiga (Ssemadaali *et al*., 2016).



Joonis 3. Näited inimeses leiduvatest viirustest. DNA viirused moodustavad sinised dsDNA ja roosad ssDNA viirused. RNA viirused moodustavad oranžid dsRNA, rohelised (+)ssRNA ja pruunid (-)ssRNA viirused. Punased on retroviirused. (Popgeorgiev *et al.*, 2013, kohandatud)

Inimese viroomi uuringu väga oluline osa on ka viiruste ja teiste mikrobioomi komponentide vaheliste interaktsioonide kirjeldamine. Eriti tähtsad on need bakteritega, kus mõlemad osapooled võivad moduleerida üksteise infektsioonilisust. Üldiselt arvatakse, et viiruste kooslus on inimese mikroflooras väga muutuv (Ly *et al*., 2016). Samas on hiljutised uuringud näidanud, et suuõõne

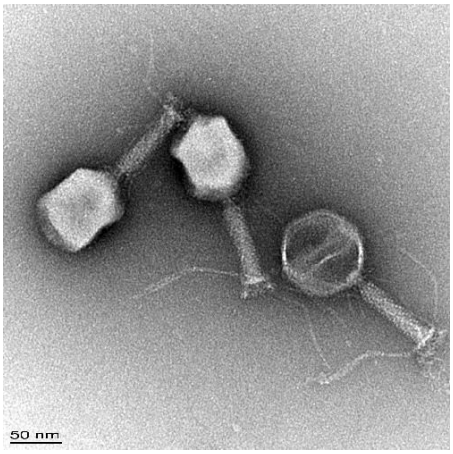
(Abeles et al., 2014) ja soolestiku (Minot et al., 2011) mikrobioomi bakteriofaagid võivad olla väga püsivad. (Popgeorgiev *et al.*, 2013)

Lisaks inimese viroomis leiduvatele viirustele on meie genoomis ligikaudu 100 000 endogeenset viiruslikku elementi (EVE), moodustades umbes 8% inimese genoomist (Belshaw *et al*., 2004). EVE-d on üldjuhul inaktiivsed või mittefunktsionaalsed, kuid teatud tingimustes võivad taasaktiveeruda ja viiruslikke transkripte ja valke toota. Mõned haigused, näiteks hulgiskleroos või amüotroofiline lateraalne skleroos, on seotud inimese endogeense retroviiruse fragmentide ekspressiooni suurenemisega. (Katzourakis ja Gifford, 2010)

Uute sekveneerimistehnoloogiate kasutuselevõtt on võimaldanud teadlastel inimeste viroome analüüsida üle maailma. Tänu sellele on avastatud uusi viiruseid tervetel ja haigetel indiviididel, võimaldades viiruste seostamist spetsiifiliste haigustega (Popgeorgiev *et al*., 2013). Praeguse seisuga on teada 219 viiruse liiki, mis suudavad inimesi nakatada. Samas ka tervetel inimestel leidub erinevaid viiruseid varjatud kujul, ilma haigusnähtudeta. Lisaks inimesed, kes on nakkusest taastunud, võivad veel kanda patogeenseid viiruseid (Zárate *et al*., 2017). Näiteks inimesed, kes on läbi põdenud noroviiruste poolt põhjustatud gastroenteriidi, võivad levitada patogeene veel kuni 8 nädalat (Atmar *et al*., 2008).

2. VIIRUSE OSAKESTE AVASTAMINE, ARVUKUSE MÄÄRAMINE JA ÜLDINE UURIMINE

Viiruste osakeste (virionide) tuvastamiseks või uute viiruste avastamiseks erinevatest keskkondadest on mitmeid meetodeid. Kõige lihtsam viis virione tuvastada on neid mikroskoobist vaadates. Viiruste väikse suuruse tõttu ei saa neid valgusmikroskoopia abil tuvastada, selle asemel kasutatakse elektronmikroskoopiat (EM) (Laue, 2010). EM on potentsiaalselt võimeline tuvastama kõiki proovis sisalduvad viiruse osakesed. Samuti on osutunud see oluliseks ka viiruste morfoloogiliste tunnuste iseloomustamisel, tänu millele saab viiruseid klassifitseerida taksonoomilistesse kategooriatesse (Roingeard, 2008). EM peamine meetod on negatiivne värvimine, mille puhul värvitakse vaid taust, tuues viiruse osakesed selle peal esile (joonis 4). Värvimiseks kasutatakse näiteks glutaaraldehüüdi, mis on väga efektiivne inaktiveeriv aine, kuid võib põhjustada agregaate ja maskeerida viiruse struktuure. Seejärel saab virionid kokku lugeda ning määrata nende hinnangulise arvukuse proovis. Antud meetod on suhteliselt kiire ja lihtne, ühe proovi valmistamiseks ja analüüsimiseks kulub üks tund. (Kunz *et al.*, 1970)



Joonis 4. *Enterobacteria* faag T4 pilt elektronmikroskoobist. Negatiivse värvimisega saadud elektronmikroskoobist pilt, kus on kujutatud faage T4, mis on umbes 90 nm laiused ja 200 nm pikkused. (http://www.scopem.ethz.ch/gallery/02.html)

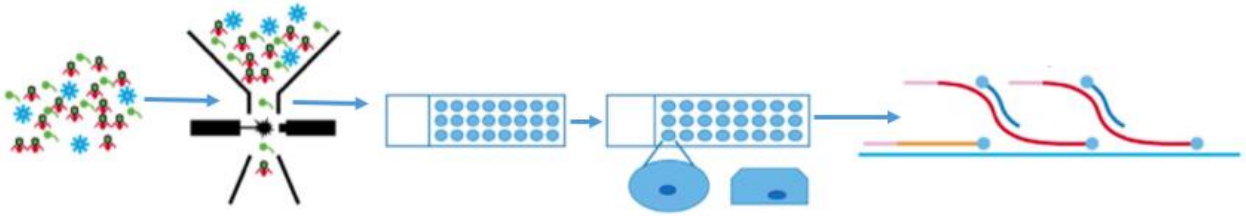
Lisaks EM-ile kasutatakse viiruste arvukuse määramiseks läbivoolutsütomeetriat. Läbivoolutsütomeetria kasutab ühte laserkiirt, mis võimaldab mitme viiruse omaduse, sealhulgas viiruste arvu ja suuruse kvantitatiivset mõõtmist. Viiruse osakesi värvitakse loendamiseks fluorestseeruvate nukleiinhappe värvidega (Kunz *et al.*, 1970). EM-ga sarnaselt piirdub arvukuse

määramine üldise hinnanguga, kuid erinevalt EM-st ei saa läbivoolutsütomeetriaga iseloomustada viiruste morfoloogilisi tunnuseid, näiteks kuju (Hayes *et al.*, 2017).

Viiruslike kapsiidide ja membraanivalkude olemasolu kindlaks tegemiseks ümbritsevatest viirustest kasutatakse massispektomeetriat (MS) (Pierson *et al*., 2014). Valgu identifitseerimine algab tavaliselt puhastamisest (üldiselt kasutatakse geelelektroforeesi), proteolüütilist lagundamist (viiruse kapsiid koosneb valgukestest, mida saab kaardistada piiratud ja selektiivse proteolüütilise lagundamise abil) ja massi analüüsi. Massi analüüs seisneb keemiliste ühendite ioniseerimises, et tekitada laetud molekulid või molekulide fragmendid. Nii saab määrata massi ja laengu suhte (Boggess, 2001). Erinevalt eelpool mainitud meetoditest sõltub MS olemasolevatest andmetest, et tuvastada valgu päritolu. (Trauger *et al*., 2003)

MS abil on avastatud gripiviiruseid otse kliinilistest proovidest, kuid selle teostamine on suhteliselt kulukas (Musaji *et al*., 2016). MS ei sobi mitte ainult viiruse alamtüübi kinnitamiseks, vaid on ka hea meetod viiruse epitoopide tuvastamiseks, mille alusel antikehad neid erisatavad (Chou *et al*., 2011).

Viiruste genoomide eraldamiseks ja iseloomustamiseks saab kasutada ka ühe viiruse genoomikat (SVG). See meetod kasutab virionide sorteerimiseks voolutsütomeetriat, millele järgneb genoomide amplifitseerimine mitmekordse nihke amplifikatsioon (MDA) (joonis 5). Võrreldes tavapärase polümeraasi ahelraktsiooniga (PCR), toodab MDA suurema suurusega genoome, mille sagedus on madalam. (Allen *et al*., 2011)



Joonis 5. SVG metoodika. Viiruse suspensioonid sorteeritakse läbivoolutsütomeetria abil üksikuteks viiruse osakesteks, mis seejärel pannakse agaroosidele. Viirus kaetakse veel täiendava kihi agaroosiga. Lõpuks tehakse kogu genoomi amplifitseerimine *in situ*. (Allen *et al*., 2011, kohandatud)

SVG on kasutatud ookeanitest ja meredest viiruste osakeste eraldamisel ja amplifitseerimisel. Atlandi ookeanist ja Vahemerest sorteeriti läbivoolutsütomeetria abil 2324 viiruse osakest. Ühe virioni kogu genoomi amplifitseerimisel saadi 392 merelise viiruse genoomi (vSAG), millest valiti juhuslikult 44 sekveneerimiseks. Analüüsid kinnitasid, et vSAG-id olid viirused, mitte

muud tüüpi bioloogilised osakesed. See näitab, et SVG meetod suudab avastada mõned tõenäoliselt kõige rikkalikumad ja ökoloogiliselt olulised viiruse liigid, mis teised metoodikad jätavad leidmata. (Martinez-Hernandez *et al*., 2017)

3. VIIRUSTE TUVASTAMISE MEETODID

Viiruste taksonoomia kindlaks tegemiseks on välja töötatud suur hulk mikrobioloogilisi, biokeemilisi, molekulaar-bioloogilisi, immunoloogilisi ja füüsikalisi meetodeid.

Viiruste uurimiseks ja identifitseerimiseks kasutatavad uurimismeetod jagatakse põhiliselt kahte rühma:

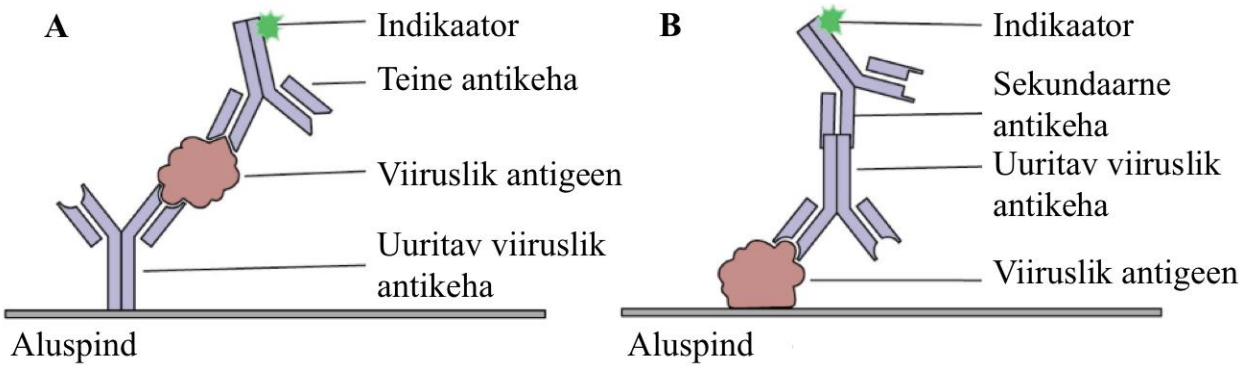
1. Viiruse antigeenide otsene indikatsioon proovidest ehk immunoindikatsioon (immunoloogilised meetodid);
2. viiruse nukleiinhapete analüüs (molekulaar-geneetilised meetodid).

(Carter ja Saunders, 2007)

3.1. Immunoloogilised meetodid

Immunoloogilisi meetodeid võib lugeda kõige hiljutisemate ja kõige levinumate meetodite hulka viiruste tuvastamisel. Nende meetodite tundlikkus sõltub tugevalt eelnevalt saadud antikehade kvaliteedist. Immunoloogilised meetodid on väga head viiruste määramisel, kuna need võimaldavad tuvastada viiruste esinemist uuritavates proovides ka siis, kui proovist nende eraldamine ebaõnnestub. (Guliy *et al*., 2018)

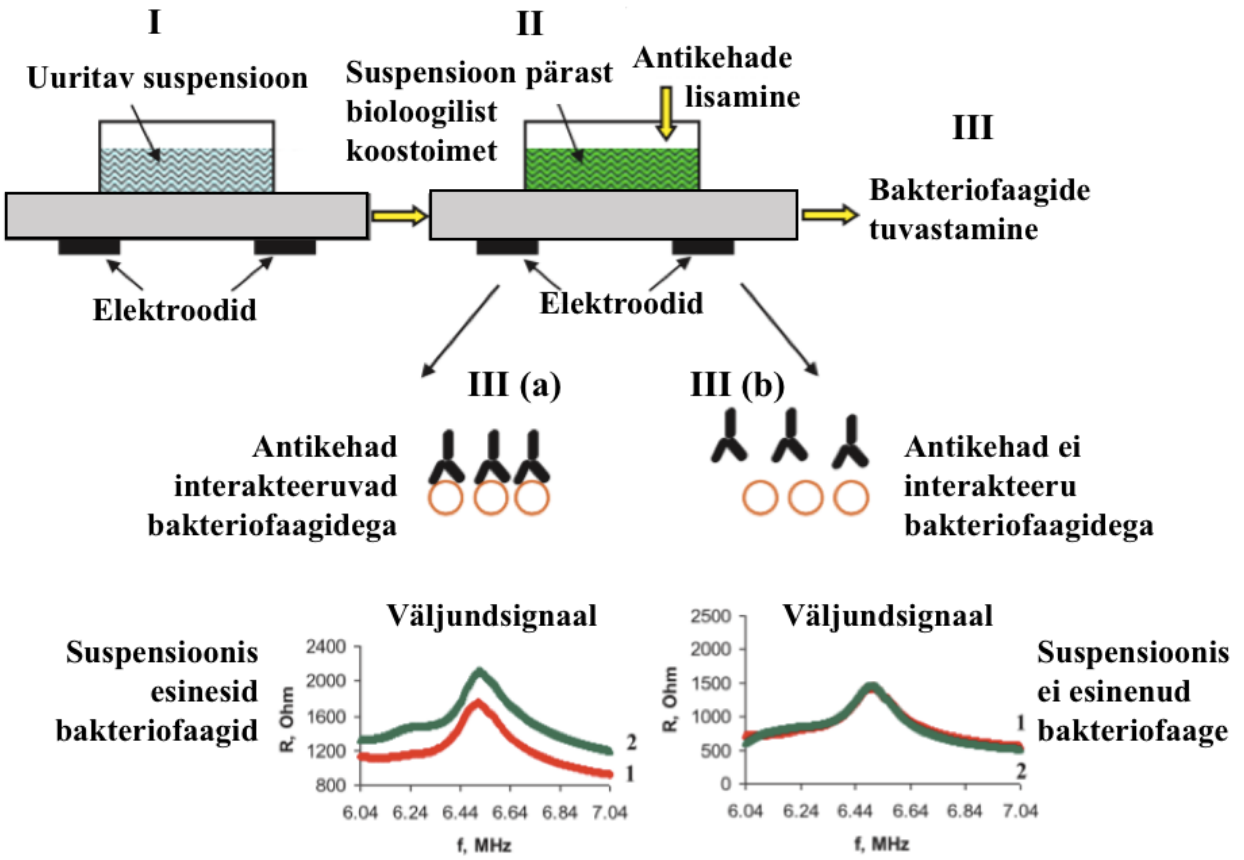
Kõige levinumaks immunoloogiliseks meetodiks on immunoensüümmeetod (ELISA) (Pavlov *et al.*, 1986), mis põhineb valkude omadusel kergesti seonduda aluspinnaga (joonis 6).



Joonis 6. ELISA skeem. (A) Esimeses variandis on viirusliku valgu vastu suunatud antikeha seotud mingi aluspinnaga, näiteks plastikust mikrotiiterplaat. Proovi lisamisel seondub viiruslik antigeen antud antikehaga ja seejärel tuvastatakse seotud viiruse antigeen teise markeeritud antikehaga. (B) Teise variandina seotakse viiruse osake või antigeen mingi aluspinnaga ning seejärel lisatakse antikehad. Kui proovis esineb viirusevastaseid antikehi, siis need seonduvad immobiliseeritud antigeeniga. Seotud antikehad tuvastatakse seejärel, kasutades sekundaarset

antikeha, mis seondub esimese antikehaga. (Raccaniello, 2010, http://www.virology.ws/2010/07/16/detection-of-antigens-or-antibodies-by-elisa/, kohandatud)

ELISA meetod on väga täpne ja tundlik ning samuti lihtsasti teostatav, kuid ELISA reaktiive on raske toota ja mõnede viiruste kvaliteetseid reaktiive pole saadaval, kuna ei tunta täpselt sihtmärki (https://www.eurofinsus.com/media/161936/detecting-virus-on-your-vines.pdf). Samuti on ELISA aeganõudev ja tulemuste saavutamiseks kulub umbes 12 tundi (Chou *et al.*, 2011). Bakteriofaagide tuvastamiseks nende spetsiifiliste antikehadega saab kasutada ka elektroakustilist meetodit (joonis 7). See meetod põhineb spetsiifilistest bioloogilistest vastasmõjudest tuleneva signaali kõikumiste tuvastamisel (Guliy *et al.*, 2016).



Joonis 7. Elektroakustilise meetodi üldine skeem. Mõõtmisprotsess koosneb järgmistest etappidest: I faagi suspensioon paigutatakse vedeliku mahutisse ja mõõdetakse anduri analüütilist signaali; II antikehad lisatakse vedelale anumale ja mõõdetakse analüütilist signaali; III saadud tulemusi analüüsitakse ja see võimaldab teha järelduse uuritavate bakteriofaagide olemasolu kohta suspensioonis: III (a) kui antikehad interakteeruvad bakteriofaagidega, registreeritakse analüütiline signaal bakteriofaagide suspensioonile antikehadega (kõver 1) ja ilma nendeta (kõver 2), kõverad on oluliselt erinevad; III (b) kui antikehad ei interakteeru bakteriofaagidega, registreeritakse analüütiline signaal bakteriofaagide suspensioonile antikehadega (kõver 1) ja ilma nendeta (kõver 2), kõverad on praktiliselt ühesugused. (Guliy *et al*., 2018, kohandatud)

Elektroakustiline meetod bakteriofaagide tuvastamiseks erineb teadaolevatest meetoditest immobiliseeritud antikehadega lihtsuse, suhteliselt suure tundlikkuse ja kiiruse tõttu. Antud meetod võimaldab tuvastada bakteriofaagid vahetult uuritava proovi vedelas faasis, registreerides bakteriofaag-spetsiifilise antikeha interaktsiooni, mille läbiviimiseks kulub umbes viis minutit. (Guliy *et al.*, 2018)

3.2. Molekulaar-geneetilised meetodid

Viiruste ja viroidide tuvastamiseks ja ka avastamiseks saab kasutada meetodeid, mis põhinevad nukleiinhapete analüüsil. Võrreldes immunoloogiliste meetoditega on molekulaar-geneetilised meetodid palju töömahukamad, kallimad ja aeglasemad. Samas nukleiinhappe analüüs võimaldab avastada rakukultuuris raskesti paljundatavaid viiruseid, eristada antigeenselt sarnaseid viiruseid ja tuvastada madala arvukusega viiruseid. Molekulaar-geneetiliste meetodite hulka kuulub nii traditsiooniline PCR analüüs kui ka fluorestsents *in situ* hübridisatsioon. (M, K., Hasamy *et al.*, 2008)

3.2.1 Geeni markeerimisel põhinevad meetodid

Erinevalt rakulistest organismidest ei oma viirused ühte ühist geeni. Seega saab markergeenide abil eristada vaid konkreetseid viiruste sugukondi, perekondi või rühmi. PCR on üks laialt levinumaid ja lihtsamaid meetodeid kindla nukleiinhappe järjestuse paljundamiseks ja tuvastamiseks proovist (Mullis ja Faloona, 1987). PCR-põhine viiruste tuvastamine ja identifitseerimine eeldab viiruse genoomi või geenide olemasolu andmebaasides. PCR on äärmiselt tundlik meetod, olles võimeline tuvastama viiruste olemasolu, mille osakaal proovis on väga madal, piisab ka ühest viirusest (M et al., 2008). Suure tundlikkuse tõttu võib isegi vähene saastatus tekitada valepositiivseid tulemusi. Lisaks võib probleeme tekitada ka viiruste kõrge mutatsioonimäär, põhjustades viiruslike nukleiinhappejärjestuste ulatuslikke muutusi, mis muudavad spetsiifilise PCR praimeri kasutamise problemaatiliseks (Clem *et al.*, 2007). Kuna aga PCR on kiire, odav ja suhteliselt lihtne ning see oli ka esimene laialdaselt kasutatav sihtmärgi amplifitseerimise tehnoloogia, siis on kaubanduslikult saadaval arvukalt viiruslikke praimerite komplekte (M, K., Hasamy *et al.*, 2008).

Teine laialt kasutusel olev detekteerimismeetod on fluorestsents *in situ* hübridisatsioon (FISH), mida kasutatakse rakkude ja kudede nukleiinhapete (RNA ja DNA) markeerimiseks. FISH seisneb DNA sondi hübridiseerimises selle komplementaarse järjestusega kromosoomsetest preparaatidest. Sondid on värvitud fluorestseeruvate nukleiinhapete värvidega, et neid saaks vaadelda fluorestsentsmikroskoobiga (Rudkin ja Stollar, 1977). FISH on väga täpne ja suudab ühendada omavahel mitu oligonukleotiidi proovi. See eelis annab võimaluse avastada ühest proovist mitu erinevat mikroorganismi (Volpi ja Bridger, 2008). Lisaks saab FISH-i ühendada teiste tuvastamismeetoditega, näiteks immunoloogiliste meetoditega. Kombineerimine võib vajalik olla siis, kui antud organismile pole kättesaadavaid antikehi või need on ebaefektiivsed (Raquin *et al*., 2012). FISH-i on kasutatud näiteks B-hepatiiti põhjustavate viiruste (HBV) tuvastamiseks inimese HepAD38 rakkudest, kus toimub HBV replikatsioon (Zhang *et al.*, 2017).

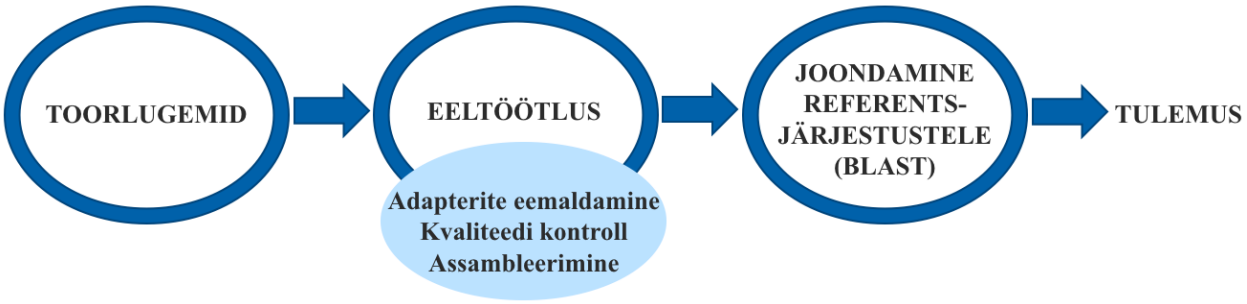
4. SEKVENEERIMISE ANDMETE ANALÜÜSIMINE

Teise põlvkonna sekveneerimise (NGS) tehnoloogia võimaldab viiruseid tuvastada ning uusi viiruseid avastada kliinilistest ja keskkonnaproovidest peremeesorganismidest sõltumatult. Sekveneerimisandmete analüüs nõuab aga bioinformaatilisi programme või töövooge, mis võimaldaksid suurte andmemahtude kiiret ja täpset töötlemist (Lin *et al*., 2017). Selleks kavandatud töövood on tavaliselt jagatud kahte kategooriasse. Esimese kategooria moodustavad kahe järjestuse võrdlemisel põhinevad töövood, mille alla kuuluvad näiteks VirusSeeker, Vipie, SURPI ja VirSorter. Teise kategooriasse kuuluvad programmid kasutavad k-meeride hulkade võrdlemist. Nende hulka kuuluvad näiteks NBC, Kraken, CLARK, Centrifuge ja VirFinder. Eelpool mainitud programmid või töövood on mõeldud teadaolevate viiruste tuvastamiseks, kuid, näiteks SURPI, VirFinder ja VirSorter võimaldavad ka uute viiruste avastamist.

4.1. Kahe järjestuse võrdlemisel põhinevad meetodid

Esimesse kategooriasse kuuluvad töövood, mis põhinevad kahe järjestuse võrdlemisel, näiteks lugemite võrdlemisel andmebaasis olevate järjestustega. Antud töövoo võtmeprogrammiks on tihti BLAST (*The Basic Local Alignment Search Tool*), mis on üks laialt levinumaid programme kahe järjestuse võrdlemisel. Tegemist on lokaalse joonduse otsingu programmiga, mis võrdleb nukleotiidide või valgu järjestusi andmebaasides leiduvate järjestustega. Seejärel arvutab BLAST sarnasusskoori ning E-väärtuse, mis näitab kui mitu vähemalt sama skooriga või suurema skooriga joondust oleks võinud leida antud suurusega andmebaasist juhuslikult. Seega sõltub E-väärtus andmebaasi suurusest (Kerfeld ja Scott, 2011). BLAST-i võib kasutada järjestuste funktsionaalsete ja evolutsiooniliste suhete tuvastamiseks, samuti aitab see tuvastada geenipere liikmeid. Sekveneerimise andmete analüüsimisel on BLAST sageli esimene samm, et saada esialgseid tulemusi. (Altschul *et al.*, 1990)

Mitmed viiruste tuvastamiseks kavandatud töövood baseeruvad BLAST-il, millele eelneb toorandmete eeltöötlus (joonis 7). Eeltöötluse alla kuulub tavaliselt adapterite eemaldamine, kvaliteedi kontroll ning mõnel juhul ka assambleerimine kontiigideks ehk ühendlugemiks. Kontiigi puhul on tegemist ülekattes olevate lugemite ühendamisel saadav pikem vahedeta järjestus.



Joonis 8. Kahe järjestuse võrdlemisel põhinevate meetodite üldine skeem. Esimene samm on toorandmete eeltöötlus ja kontroll, mille alla kuulub tavaliselt adapterite eemaldamine, kvaliteedi kontroll ning mõnel juhul ka assambleerimine kontiigideks. Saadud kontiigid või lugemid joondatakse BLAST-i või mõne muu joonduse otsingu programmiga. Tulemuseks on kõige sarnasem järjestus, mis leidus otsingus kasutatud andmebaasis.

Sarnast skeemi (joonis 8) kasutab VirusSeeker, mis on mõeldud nii uute eukarüootsete viiruste avastamiseks kui ka viroomi koostise analüüsiks. VirusSeekeri põhiline tugevus on valepositiivsete tulemuste tuvastamine eukarüootsete viiruste detekteerimisel, kasutades järjestikuseid BLAST-i töövooge ja kureeritud viiruste andmebaasi. Viroomi koostise analüüsiga tegeleb VirusSeekeri alamtöövoog VirusSeeker-Virome (VS-Virome), mis on mõeldud nii tuntud kui ka uute viiruslike järjestuste ja arvukuse määratlemiseks metagenoomidest. VS-Virome'i eeltöötluse alla kuulub adapterite eemaldamine, kattuvate paarislugemite ühendamine, lugemite kvaliteedi kontrollimine ning madala keerukusega järjestuste ja peremeesjärjestuste tuvastamine. Uute viiruste avastamiseks kasutatakse VirusSeeker-Discovery (VS-Discovery) töövoogu, mis hõlbustab ka kontiigidepõhist viroomi koostise analüüsi. VS-Discovery sisaldab ka *de novo* assembleerimist, et luua kontiige, mis suurendab referentsandmetest väga erinevate viiruste avastamise tõenäosust. (Zhao *et al*., 2017)

VirusSeeker-i kõrval on ka teisi BLASTI-il baseeruvaid programme viiruste tuvastamiseks. Vipie on veebipõhine viiruse populatsiooni mitmekesisuse määramise programm, mis erinevalt teistest bioinformaatilistest töövoogudest võimaldab erinevate proovide samaaegset analüüsimist. Kõigi toorlugemite jaoks tehakse paralleelselt järgmised sammud: kvaliteedikontroll, assembleerimine, kontiigide taksonoomiline klassifikatsioon, joondamine BLAST-iga ja lõpuks saadakse taksonoomilise klassifikatsiooniga identifitseeritud võrdlusjärjestused (joonis 8). Vipie töötleb kõiki proove identse protokolliga ja need on kvantifitseeritud ühiste võrdlusjärjestuste suhtes, mis on väga oluline uute viiruste avastamisel. (Lin *et al*., 2017)

Kahe järjestuse võrdlemise meetodit kasutab veel ka arvutuslik töövoog SURPI (*Sequence-based Ultrarapid Pathogen Identification*), mis on ette nähtud patogeenide identifitseerimiseks kliinilistest proovidest. SURPI on võimeline analüüsima keerulisi metagenoomseid proove, mis sisaldavad enam kui 1,1 miljardit järjestust. Need andmekogud hõlmavad mitmesuguseid tuvastatud patogeene, proovitüüpe ja katvuse sügavusi. SURPI-l on kaks režiimi – kiire ja terviklik. Kiires režiimis tuvastab SURPI viirused ja bakterid 11 minuti kuni 5 tunniga, mis on lineaarselt võrdeline lugemit arvuga. Terviklikus režiimis identifitseeritakse kõik teadaolevad mikroorganismid lugemite põhjal, millele järgneb assembleerimine ja valgu homoloogiline otsing erinevates viirustes. Selle peale kulub 50 minutit kuni 16 tundi. Mõlema režiimi puhul on BLAST asendatud SNAP-ga. Tervikliku režiimi puhul kasutatakse ka RAPSearch-i transleeritud valgujärjestuste võrdlemiseks. Need kaks programmi võimaldavad NGS metagenoomika andmete analüüsi teostada kuni kümme korda kiiremini, kuid sama täpselt kui BLAST. (Naccache *et al*., 2014)

4.2. K-meeridel põhinevad viiruste tuvastamiseks kavandatud töövood

Metagenoomist saab tuvastada viiruslikke lugemeid ka ilma eeltöötluse ja joondamiseta, kasutades järjestuste k-meere. K-meer on fikseeritud pikkusega (k) oligomeer, mis koosneb nukleotiididest (joonis 9). Kõikide võimalikke k-meeride arv on 4^k. K-meeride kasutamine pakub suuremat spetsiifilisust, kuid madalamat sensitiivsust. (Ounit *et al*., 2015)

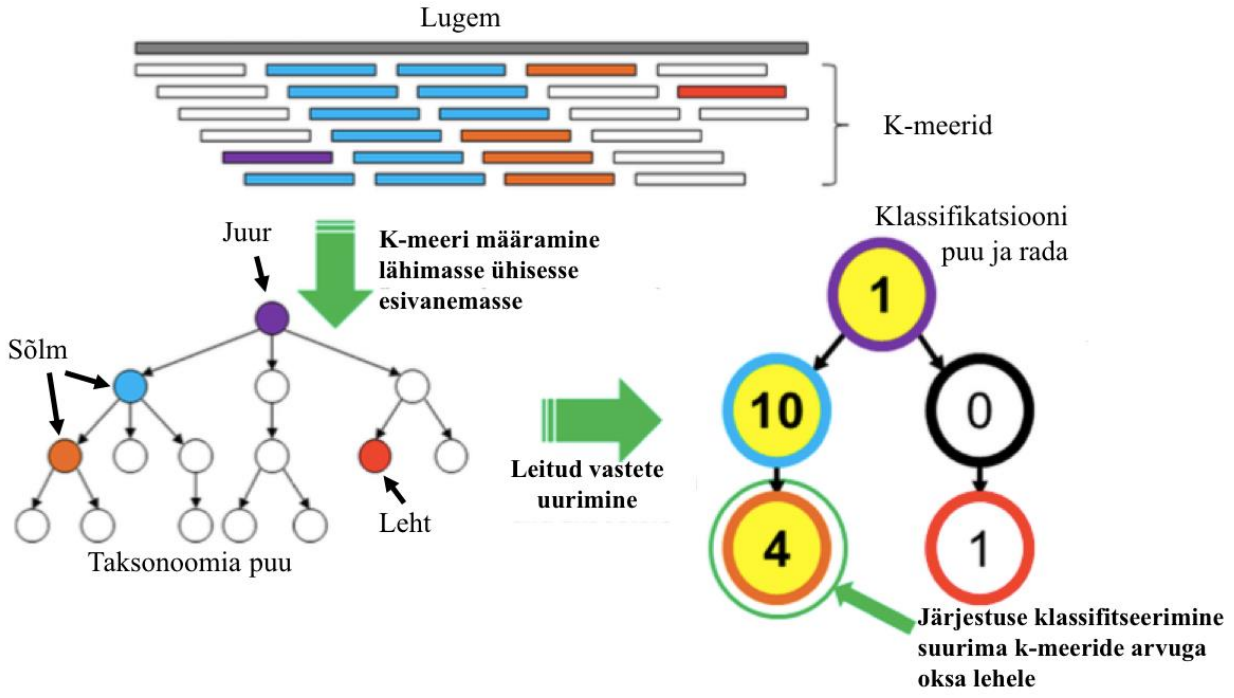
Järjestus: **CCGTAA**
5-meerid: **CCGTA, CGTAA**
4-meerid: **CCGT, CGTA, GTAA**
3-meerid: **CCG, CGT, GTA, TAA**

Joonis 9. Näide järjestusest ja sellest saadud k-meeridest. Joonisel on välja toodud ühe 6 nukleotiidi pikkuse järjestuse põhjal erinevad 5-, 4- ja 3-meerid.

Välja on töötatud mitmeid k-meeridel põhinevaid programme, et taksonoomiliselt klassifitseerida lugemid metagenoomsetest andmetest. Sellised programmid on näiteks CLARK (*CLAssifier based on Reduced K-mers*) (Ounit *et al*., 2015), Kraken (Wood ja Salzberg, 2014) ja NBC (*Naive Bayes Classifier*) (Rosen *et al*., 2011), mis tuvastavad eelkõige baktereid, kuid saab kasutada ka

viiruste puhul. Eelpool nimetatud programmid hindavad mikroobide liikide suhtelist arvukust, võrreldes lugemitest tehtud k-meeride hulka referentsgenoomide k-meeridega. Kuna need programmid pole geenikesksed, siis saab neid kasutada nii kodeerivate kui ka mittekodeerivate järjestuste klassifitseerimiseks (Bazinet ja Cummings, 2012). Kuna klassifitseerimine tugineb teadaolevate genoomide võrdlemisele, siis ei sobi need algoritmid uut viiruste avastamiseks. (Hurwitz *et al*., 2018)

NBC otsib vasteid valgu k-meeridele andmebaasidest, mis sisaldavad viiruseid ja klassifitseerib <100 lugemit minutis (lm) (Rosen *et al*., 2011). Kuna see on miljonite lugemit sisaldavate andmekogumite jaoks liiga aeglane, loodi 2014. aastal kiirem programm Kraken, mis klassifitseerib üle 1,5 miljoni lugemi minutis (Wood ja Salzberg, 2014). Krakeni kiiruse eelis tuleneb suurel määral k-meeride täpse vaste andmebaasi päringute kasutamisest, mille teeb võimalikuks järjestatud mikroobide genoomide väga suur ja üha kasvav arv. Krakeni algoritm (joonis 10) võimaldab saavutada suurt tundlikkust ja täpsust perekondlikul tasemel.

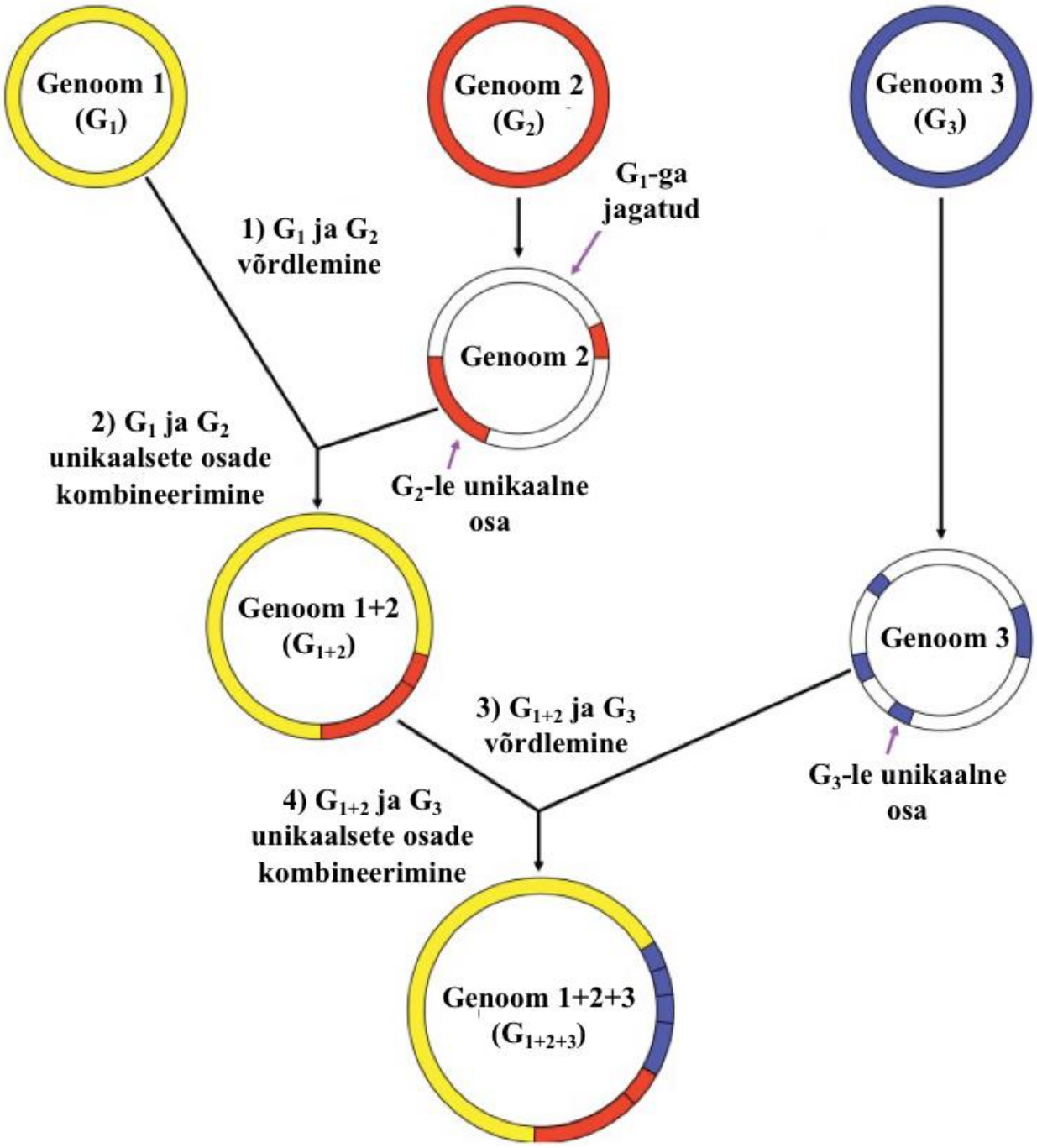


Joonis 10. Krakeni klassifitseerimisalgoritm. Järjestuse klassifitseerimiseks kaardistatakse iga järjestuses olev k-meer andmebaasi k-meeri sisaldavate genoomide madalaima ühise eelasega sõlme. Taksonoomia puus on igal sõlmel kaal, mis on võrdne k-meeride arvuga sõlme taksoniga seotud järjestuses. Seejärel leitakse suurima kaaluga oks ning järjestus klassifitseeritakse kõige tipmisele sõlmele. (Wood ja Salzberg, 2014, kohandatud)

Erinevalt NBC-st võib Kraken jätta mõned järjestused klassifitseerimata, kui puuduvad piisavad tõendid. Kuna NBC klassifitseerib kõik järjestused, kipub see väljastama rohkem valepositiivseid tulemusi kui Kraken (Wood ja Salzberg, 2014). Samas on täheldatud, et NBC näitab kõrgemat täpsust ja tundlikkust perekonna tasandil (Ounit *et al*., 2015).

Lisaks Krakenile ja NBC-le on olemas ka teisi k-meeridel põhinevad programme. CLARK võimaldab täpselt ja tõhusalt klassifitseerida metagenoomilisi lugemeid liigi või perekonna tasemel, mis põhinevad k-meeride väiksematel komplektidel. Esmalt loob CLARK sihtjärjestustest indeksi, milleks on räsi tabel. Seejärel eemaldatakse tabelist k-meerid, mis ilmuvad rohkem kui ühele sihtmärgile. Sellised ühised k-meerid võivad põhjustada vigu, kuna andmebaasides on palju ühiste k-meeridega järjestusi. Lugem määratakse sellesse lahtrisse, mille k-meere see kõige rohkem sisaldab. Teised metagenoomilised klassifikaatorid, nagu ka Kraken, ei paku sellist müra kaitset, mis on väga kasulik sihtjärjestuste lugemisel. (Ounit *et al*., 2015)

Töövoog Centrifuge on lugemite klassifitseerimise programm, mis võimaldab kiiret, täpset ja tundlikku lugemite märgistamist ja liikide kvantifitseerimist. Süsteem kasutab indekseerimiskava (joonis 11), mis on optimeeritud spetsiaalselt metagenoomse klassifitseerimise probleemi jaoks. Centrifuge, erinevalt Krakenist, kasutab palju vähem mälu (4,2 GB 4078 bakteriaalse ja 200 arhe genoomiga) ja klassifitseerib suurel kiirusel järjestused, mis võimaldavad tal mõne minuti jooksul töödelda miljoneid lugemeid. Krakeni andmebaas vajab märkimisväärselt rohkem mälu (93 GB 4278 prokarüootse genoomiga), kui tänapäeva lauaarvutites on. Centrifuge võib määrata ühe järjestuse kuni viite taksonoomia kategooriasse. See strateegia erineb Krakenist, mis valib alati ühe taksonoomilise kategooria, kasutades kõigi liikide madalaimat ühist eellast. (Kim *et al*., 2016)



Joonis 11. Centrifuge'i klassifitseerimisalgoritm. Kõiki genoome võrreldakse ja valitakse kaks kõige sarnasemat genoomi (G_1 ja G_2), suurima arvu k-meeride põhjal. G_2 järjestused, mis on G_1-ga identsed (≥99%), jäetakse kõrvale ja ülejäänud unikaalsed järjestused G_2-st lisatakse genoomi G_1, luues ühendatud genoomi G_{1+2}. Seejärel lisatakse ühendatud genoomi järjestused, mis on <99% identsed genoomis G_3, luues genoomi G_{1+2+3}. See protsess kordub kogu Centrifuge'i andmebaasi jaoks, kuni iga ühendatud genoomi järjestused ei ole ≥99% identsed mis tahes muu genoomi suhtes. (Kim *et al*., 2016, kohandatud)

Kokkuvõttes on ülaltoodud programmidest väikseima ajakuluga CLARK, mille kiirus ületab Krakenit kuni viis korda (Ounit *et al*., 2015) ja Centrifuge'i kuni üheksa korda (Kim *et al*., 2016). CLARK on ka kõige täpsem, kuid kõige tundlikum on Centrifuge. Kõigi kolme programmi omavahelised erinevused täpsuse ja tundlikkuse mõõtmisel osutusid väga väikesteks (tabel 2).

Tabel 2. Krakeni, CLARK-i ja Centrifuge'i võrdlus.

Tunnus	**Kraken ja CLARK[1]**		**Kraken ja Centrifuge[2]**	
Kiirus	4,1 miljonit lugemit minutis (lm)[#]	32 miljonit lm (lühikest)	Keskmiselt 1 061 947 lm	Keskmiselt 563 380 lm
Mälu nõuded (RAM)	93 GB	4 GB	93 GB	4,2 GB
Tundlikkus*	82,67%	82,26%	77,94%	78,09%
Täpsus*	99,26%	99,31%	87,18%	87,08%

* Tundlikkust ja täpsust võrreldi perekonna tasemel, genoomid sekveneeriti Illumina HiSeq-ga.
[#] Lugemid olid 100 aluspaari pikkused
1. Võrdlus baseerub (Ounit *et al*., 2015)
2. Võrdlus baseerub (Kim *et al*., 2016)

4.3. Uute viiruste avastamiseks kavandatud töövood

Meie arusaam viiruse mitmekesisusest ja mõjudest on endiselt piiratud liiga väheste mudelsüsteemide ja võrdlusgenoomide tõttu. Üks võimalus täita need lüngad, on tuvastades viiruse signaali metagenoomsetest andmetest. Töövoog VirSorter on kavandatud viirusliku signaali tuvastamiseks genoomsetest andmetest. VirSorteri etapid toore nukleotiidijärjestuse eeltöötlemisest hõlmavad assambleerimist, rõngakujuliste järjestuste tuvastamist, geenide ennustamist ja kõikide järjestuste valimist, milles on ennustatud geene rohkem kui kaks. Järjestuse eeltöötlemise järel tuvastatakse viiruse spetsiifilised piirkonnad, kasutades mitmeid meetodeid (kodeerivate geenide olemasolu, viiruse-sarnaste geenide rikastamine, kodeeriva ahela vahetus kahe järjestikuse geeni vahel). Lõpuks, kui 80% kõigist kontiigi ennustatud geenidest on ennustatud viiruslikeks, peetakse kogu kontiigi viiruslikuks. (Roux *et al*., 2015)

VirSorter ei sobi lühikeste (<3kbp) kontiigide tuvastamiseks, sest see eeldab ennustuse tegemiseks kolme kodeeriva geeni olemasolu. Lisaks on teada, et mõnedel viirustel on ligikaudu 11-14% mittekodeerivat piirkonda, mis vähendab tõenäosust, et kontiigi sisse satub 3 kodeerivat geeni (Brown TA, 2002). VirSorteri tulemused ei ole omavahel võrreldavad, sest töövoog kasutab esmalt kõiki kontiige, et hinnata tausta ja viiruse ennustused tehakse, kui võrreldakse individuaalseid kontiige nende taustaga. (Roux *et al*., 2015)

Sarnaselt VirSorterile on töövoog VirFinder loodud uute viiruste avastamiseks, kuid kasutab geenipõhiste sarnasuste otsimise asemel k-meere (tabel 3). Viirustel esinevad spetsiifilised k-meerid, mida peremeesorganismides ei esine ja see võimaldab avastada tundmatuid viirused. VirFinder ennustused on stabiilsed igale kontiigile, olenemata teistest kontiigidest, mida testitakse samal ajal. Erinevalt VirSorterist on VirFinder paremate tulemustega lühikeste järjestuste korral, kuna geeni leidmist või geeniga sarnasuse leidmist ei ole vaja. VirSorter ja VirFinder mõlemad ei ole väga edukad eukarüootide viiruste tuvastamisel, kuna nende kohta on vähe andmeid. (Ren *et al*., 2017)

Tabel 3. VirFinderi ja VirSorteri võrdlus. Antud tabelis on välja toodud põhilised erinevused VirFinderi ja VirSorteri vahel.

VirFinder	**VirSorter**
K-meeri põhine	Geenipõhine
Tuvastab paremini lühikesi (<3kbp) viiruslikke kontiige	Tuvastab lühikesed (<3kbp) kontiigid ainult siis, kui need sisaldavad 3 kodeerivat geeni
Tulemused ei sõltu teistest kontiigidest, mida testitakse samal ajal	Tulemused sõltuvad sellest, millised muud vead on päringu andmekogumisse kaasatud

Kahe järjestuse võrdlemisel põhinevad meetodid järgivad sama üldist skeemi – viiruste lugemeid kõigepealt töödeldakse ja tehakse kvaliteedikontroll, siis joondatakse referentsjärjestustega ning sarnasuse skoori põhjal saadakse sobiv vaste. K-meeridel põhinevate programmide jaoks sellist üldist skeemi pole. Näiteks Kraken kasutab taksonoomia puud ja Centrifuge ühendatud genome, kui CLARK vaatab ainult genoomispetsiifilisi k-meere. Küll aga leiavad kõik k-meeri põhised programmid lugemi vaste suurima arvu ühiste k-meeride põhjal referentsjärjestuse ja lugemi vahel.

5. VIIRUSTE SEIRE

Inimese viroomi kirjeldus on suhteliselt uus ja üsna piiratud. Alates viiruste avastamisest enam kui 100 aastat tagasi peeti inimese viiruseid üksnes patogeenideks. Esimesena avastati kollapalaviku viirus 1901. aastal ja jätkuvalt leitakse veel kolm kuni neli uut liiki aastas (Woolhouse *et al.*, 2012). Viirustel võib olla kas kasulik või kahjulik mõju inimeste tervisele, sõltuvalt nende interaktsioonidest peremeesorganismi, teiste viiruste ja bakteritega. Eelpool nimetatud meetodeid ja programme on võimalik kasutada viiruste seireks. Viiruste ja ka teiste patogeenide seireprogrammid on tervisekaitse ja pandeemiate ärahoidmise seisukohalt hädavajalikud. Viiruste seire kätkeb endas nii linnade heitvee analüüsi, viiruslike infektsioonide leviku jälgimist kui ka viirusliku leviku alguskoha tuvastamist. (Pride *et al.*, 2012)

Viiruste seire toob tihti kaasa märkimisväärseid väljakutseid. Üks olulisemaid tähelepanekuid haiguste seirel on see, et kõik viirused ei arene edukalt uutes peremeestes, mis on üle kandunud teistelt peremeesorganismidelt. Paljud sellised viirused kujutavad endast mööduvat nakkust ja surevad varsti, isegi infektsioonide puudumisel. Näiteks, hoolimata korduvatest levikutest lindudelt inimestele, pole H5N1 linnugripi viirus olnud suuteline arenema inimeselt inimesele ülekanduvaks (Lam *et al.*, 2015). Teised viirused on olnud edukamad ja põhjustanud olulisemaid puhanguid uutes peremeestes. Näiteks Ebola viirus on tekitanud mitmeid epideemiaid, mille levikut on takistatud piiride sulgemisega. Kuigi on tõendeid Ebola viiruse kiire kohanemisega inimestes hiljutise Lääne-Aafrika epideemia ajal, peavad viirusel olema ka sellised tunnused, mis on vajalikud viiruse edasiseks ülekandumiseks uues peremehes (Dudas *et al.*, 2017). Leidub veel ka endeemilisi inimese patogeene, mis suudavad edasi kanduda pikka aega paljudele uutele inimestele ja ei vaja looma, et uuesti inimesele üle kanduda. Kõige tuntum näide on HIV, mis põhjustab AIDS-i. (Geoghegan ja Holmes, 2017)

5.1. Viiruste seire reoveest

Linnaheited koosnevad inimese uriinist, väljaheidetest ja naha ketetest, mistõttu sisaldab kanalisatsioon tuhandetest elanikest suure hulga patogeenseteid ja kommensaalseid viiruseid, bakteriteid ja algloomi. Lisaks sellele kulgeb läbi inimese soolestiku suur hulk taime viirusi. Viiruste eemaldamiseks kanalisatsioonist kasutatakse reoveepuhasteid (STP), mis ei ole nii efektiivne kui fekaalsete indikaatorbakterite kasutamine (Payment *et al.*, 1982). Kõrgem viiruse ellujäämine võib kujutada endast ohtu tarbijatele, tekitades infektsiooniohtu läbi vee ja toidu.

Viiruste taksonoomilist mitmekesisust on proovitud määrata otse kanalisatsiooni proovidest ning on leitud näiteks hepatiiti põhjustavaid patogeene (Fernandez-Cassi *et al*., 2018). Kogu maailmas on peaaegu 3% inimestest nakatunud C-hepatiidi viirusega (HCV) (Mohd Hanafiah et al., 2013). Nagu HIV puhul, saadakse HCV peamiselt vanematelt. Edukad HCV seireprogrammid on inimeste tervise kaitseks hädavajalikud, et katkestada HCV levik. HCV eksisteerib arvukate variantide populatsioonina igas nakatunud indiviidis (Martell *et al.*, 1992). Ülemaailmne hepatiidi puhangu ja seire tehnoloogia (GHOST) on pilvepõhine süsteem, mis võimaldab kõigil kasutajatel sõltumata nende arvutusalastest teadmistest täpset HCV molekulaarset jälgimist. GHOST-i HCV järjestamise protokoll kasutab uut amplikonipõhist järjestuse määramise meetodit, mis on suunatud HCV genoomi HVR1-le. See piirkond on valitud suhteliselt suure varieeruvuse tõttu, mis võimaldab geeniülekannetest tulenevate evolutsiooniliste vahemaade täpset hindamist. GHOST-ist saadud HCV geneetiliste andmete abil saadakse asjakohane rahvatervist käsitlev teave tõhusate sekkumismeetmete juhtimiseks. (Longmire *et al*., 2017)

Kanalisatsioonis on ka suurel kogusel inimeste uriini, mis sisaldab samuti viiruseid. Uriini kaudu erituvad viirused on halvasti uuritud (Santiago-Rodriguez *et al*., 2015), mis võib olla tingitud sellest, et uriini peetakse tavaliselt steriilseks keskkonnaks. Uriiniga eritatavaid viiruste uurimiseks puhastamata kanalisatsioonist koguti 14 erineva vanuse ja päritoluga tervetelt vabatahtlikelt 100 ml uriini (7 meest ja 7 naissoost 25 kuni 63-aastaselt), kuigi enamik neist elasid Barcelonas. Uriini viiruse kontsentraat sisaldas järgmisi DNA viiruse sugukondi, mis nakatavad inimesi: *Papillomaviridae*, *Polyomaviridae* ja järjestused, mis on kaugelt seotud rõngaskujuliste ssDNA perekondadega *Circoviridae* ja *Anelloviridae*. Need tulemused toovad esile, et uriin aitab kaasa linnaheitvee viiruslikule mitmekesisusele, tuues esile peamiselt DNA viirused. (Fernandez-Cassi *et al*., 2018)

5.2. Ülemaailmne viroomi projekt

Patogeensete viiruste kiire muteerumine ja omavaheline kombineerumine (näiteks seagripp ja linnugripp) on tekitanud maailmas uusi pandeemiaid, mis on avaldanud suurt mõju inimeste tervisele ja majandusele. Vähesed teadmised viiruslike ohtude mitmekesisusest ning nende esilekerkimise juhtudest takistavad meil leevendada haiguste levikut, kuna ei osata välja töötada sobivaid ravimeid.

2018. aastal käivitub ülemaailmse viroomi projekt (GVP), mille eesmärgiks on kindlaks teha suurem osa viiruslikest ohtudest ning anda asjakohaseid andmeid rahvatervise sekkumiste kohta tulevaste pandeemiate korral. Inimesi nakatavaid viiruse sugukondi teatakse 25, kus on kuni 827000 teadmata viirust, mis omavad zoonootilist potentsiaali. See on loomade nakkushaiguste potentsiaal kanduda üle inimestele (Carroll *et al*., 23.02.2018). GVP nõuab suuri investeeringuid ja isegi juhul, kui avastatakse suur hulk potentsiaalseid zoonoose, võib ainult väike osa tõenäoliselt põhjustada suuri haiguspuhanguid ja suremuse inimestel. Arvestades üksikute epideemiatega seotud kulude suurt maksumust, võivad GVP poolt toodetud andmed pakkuda olulist investeeringutasuvust, suurendades diagnostilist suutlikkust uue haiguspuhangu varajastes staadiumides. Zoonootiliste haiguste esinemissageduse suurenemisega kaasnevate hüppeliselt kasvavate majanduslike kahjude hiljutised analüüsid (Jones *et al*., 2008) näitavad, et pandeemiate leevendamise strateegiad annavad 10:1 investeeringutasuvuse. GVP eesmärk on parandada võimet tuvastada, diagnoosida ja avastada viirusi haavatavas elanikkonnas. Nagu inimgenoomi projekt, pakub GVP avalikkusele kättesaadavaid andmeid, mis võivad tuua avastusi, mida on raske ennustada, näiteks viirused, mis põhjustavad vähki või käitumishäireid. (Carroll *et al*., 2018)

ARUTELU

Paljud viirused on võimelised põhjustama mitmesuguseid haigusi. Osadel viirustel on potentsiaali põhjustada ka pandeemiaid. Seega on väga oluline uurida ja tuvastada viiruseid erinevatest keskkondadest. Metagenoomsete andmete arvutusanalüüs patogeenide identifitseerimiseks on mitmel põhjusel keeruline. Esiteks, analüüsides kasutatavad andmemahud on väga suured, NGS-tehnoloogia suudab toota päevas lugeda üle 100 GB toorlugemeid (Loman *et al*., 2012). Nende lugemite joondamise/klassifitseerimise algoritmid peavad olema võimelised need massiivsed järjestusandmed läbi töötama võimalikult väikese ajakuluga. Teiseks peavad programmid olema võimelised tuvastama patogeeni ka väga väikesest osahulgast, sest sageli on huvipakkuvate lugemite (patogeeni) osakaal proovis väga madal (Kostic *et al*., 2012). Lisaks ei võimalda madal lugemite arv ka *de novo* assambleerimist kontiigideks (Kostic *et al*., 2011). Seega, peavad programmid võimaldama lühikeste järjestuste, tavaliselt pikkusega vaid 100-300 nukleotiidi, täpset klassifitseerimist. K-meeridel põhinevad programmid tulevad lühikeste järjestuste tuvastamisega paremini toime kui geenipõhised, sest need ei eelda kodeerivate geenide olemasolu järjestuses. Kolmandaks probleemiks viiruste tuvastamisel on võrdlusandmete vähesus andmebaasides (Xu *et al*., 2011). Praegusel hetkel on enamus viiruste tuvastamise meetodid suunatud bakteriofaagide leidmisele, kuna nende järjestuste osakaal on andmebaasides suurem (Hurwitz *et al*., 2018).

Ajalooliselt baseeruvad esimesed programmid viiruslike järjestuste leidmiseks homoloogide otsingutel, mis on limiteeritud referentsgenoomide (geenide) arvuga (Hurwitz *et al*., 2018). Homoloogide otsingu programmidest on kõige laialdasemalt kasutusel BLAST. Hiljem väljatöötatud k-meeridel põhinevate meetodite eeliseks joonduspõhiste meetodite ees on lühemad arvutusajad (Chan ja Ragan, 2013). Näiteks k-meeridel põhinev Centrifuge on peaaegu 2000 korda kiirem kui MegaBLAST, mis on omakorda 10 korda kiirem klassikalisest BLAST-ist (Kim *et al*., 2016). K-meeri põhised programmid võimaldavad küll kiiremaid arvutusi, aga toetuvad taas referentsandmetele nagu ka joondamisel põhinevad lahendused.

Joondamismeetodi tulemused sisaldadavad üldjuhul lugemi täpset positsiooni referentsjärjestustel, identsust ja sarnasuse skoori. Seevastu k-meeri analüüsi üldjuhul vaatab, kas antud k-meer leiti proovist või mitte. Tulemuseks on enamasti üks arv, mis näitab, kui suur osa k-meere leiti kõigist referentsi k-meeridest. K-meeri meetodite tulemuste parandamiseks tundub seega mõistlik lisada analüüsi positsiooniline info, säilitades samal ajal kiirema arvutamise aja

eelise (Sievers *et al*., 2017). Sellisel juhul saaks teada, kas k-meerid leiti ühest piirkonnast ehk ainult ühest geenist või k-meere leiti üle kogu viiruse genoomi. See eemaldaks potentsiaalsed valepositiivsed tulemused, mida võib põhjustada üksiku geeni ülekanne peremeesorganismi või peremeesorganismis olevad endogeensed viiruslikud elemendid.

Viiruslikud järjestused võivad mõningal juhul olla väga sarnased teistest organismidest pärit järjestustega. Näiteks, on DNA polümeraasid laialdaselt konserveerunud ja ei pruugi olla klassifitseerimiseks piisavalt unikaalsed. Probleemiks on ka eukarüootsete viiruslike järjestuste suhteline vähesus andmebaasides. Samuti võivad kordusterikkad piirkonnad inimese genoomis jäljendada AT-rikkaid viiruseid (Hurwitz *et al*., 2018). See tekitab suuri probleeme järjestuste päritolu kindlaks määramisel, eriti järjestuste puhul, mille E-väärtus on lävendi lähedal. Antud probleemile pööras tähelepanu ka Zhao *et al*. oma teadustöös, kus lugemitel oli aminohappeline sarnasus nii *Phycodnaviridae* viirustega (E-väärtus 3 x 10^{-5} kuni 10^{-3}), kui ka bakterite ja bakteriofaagide järjestustega. Samas kuna E-väärtus oli täpselt üle lävendi (10^{-3}), liigitati järjestused viiruslikeks, kuid tegu võis olla ka bakteriofaagiga. *(Zhao et al.*, 2017)

Uurides olemasolevate või uute viiruslike järjestuste seost mõne tunnuse või seisundiga (nagu haigus või ökoloogiline muutuja), on hädavajalik tulemuste võrreldavus. Seda saab ohutult saavutada vaid juhul, kui kõik proovid on töödeldud identse protokolliga ja kui need on kvantifitseeritud ühiste referentsjärjestuste komplekti vastu. Sellist kvantifitseerimist teostab näiteks Vipie. (Lin *et al.*, 2017)

Käesoleva töö eesmärgiks oli anda hinnang töövoogude edukusele viiruste tuvastamisel. Eelneva analüüsi põhjal on k-meeri põhistel programmidel suuremad šansid viiruste tuvastamisel, sest need on arvutuslikult kiiremad ja tundlikumad võrreldes geenipõhiste programmidega. K-meeridel põhinevad programmid vajavad mitmeid täiendusi: positsioonilise info lisamine, mida hetkel pakuvad vaid joondamispõhised meetodid; k-meeri arvukuse kaasamine analüüsi. Ma arvan, et mõlemad täiustused võiksid parandada k-meeridel põhinevate programmide täpsust.

KOKKUVÕTE

Viiruseid leidub kõigis ökoloogilistes niššides, need juhivad ülemaailmset energia ja toitainete ringlust kontrollides nii bakterite, ainuraksete kui ka hulkraksete arvukust. Mõned viirustest võivad olla patogeensed, mistõttu on nende tuvastamine erinevatest keskondadest väga oluline. Viiruste tuvastamiseks ja avastamiseks on loodud mitmeid erinevaid töövooge, mis laiendavad meie teadmisi nende mitmekesisusest. Need töövood, mis tegelevad viiruste järjestuste andmete analüüsiga, saab jagada kahte rühma – kahe järjestuse sarnasusel põhinev ja k-meeridel hulkade võrdlemisel põhinev tuvastamine.

Suurimateks probleemideks viiruste tuvastamisel on andmebaasides olevate võrdlusandmete vähesus, järjestusandmete läbi töötamine võimalikult väikese ajakuluga ja viiruste tuvastamine proovidest, kus nende osakaal on väga madal. K-meeridel põhinevate programmide eeliseks võrrelduna joonduspõhiste töövoogudega, on suurem kiirus, samas kui täpsus ja tundlikkuson võrreldavad. K-meeridel põhinevad töövood töötavad hästi ka siis, kui lugemite madal osakaal ei võimalda nende assambleerimist kontiigideks, mistõttu peavad programmid suutma klassifitseerida ka lühikesi järjestusi. See on suur eelis k-meeridel põhinevatele meetoditele, mis ei eelda järjestuses kodeerivate geenide olemasolu, mida geenipõhised programmid teevad.

Viiruste tuvastamise töövooge kasutatakse viiruste seireks, mis hõlmab nii viiruslike infektsioonide leviku jälgimist kui ka viirusliku leviku alguskoha tuvastamist. Viiruste ja ka teiste patogeenide seireprogrammid on tervisekaitse ja pandeemiate ärahoidmise seisukohalt hädavajalikud. Arvestades üksikute epideemiatega seotud kulude suurt maksumust, pakuvad seireprogrammid ka olulist investeerigutasuvust.

RESÜMEE / SUMMARY

Various pipeline analysis for virus detection

Brigitta-Robin Raudne

Summary

Viruses are found in all ecological niches, they lead the global energy and nutrient circulation by controlling both unicellular and multicellular populations. Some of the viruses are pathogenic and cause diseases, which is why their identification from various environments is important. Several different pipelines have been developed to detect and discover viruses, which extend our knowledge of their diversity as well as abundance. These pipelines can be broadly divided into two groups – detection based on two-sequence comparison and detection based on k-mers.

There are many drawbacks in detecting viruses from samples. The greatest problem is the detection of viruses from a small fraction in case of metagenomics, but also, the relatively low amount of reference data in databases. In addition, there is usually a need for processing a huge amount of sequence data quickly. K-mer based programs allow faster results compared to alignment-based programs. However, they both depend on reference data. In many cases, the low proportion of reads that are of interest does not allow their assembly to contigs, therefore programs must also be able to classify short sequences. This again gives the advantage to k-mer based detection, which does not need sequence assembly. Also, k-mer based programs do not require the presence of encoding genes in sequence reads like gene-based programs do.

On a larger scale, virus detection pipelines can be used for monitoring, which includes observing both the spread of viral infections and the detection of the onset of viral spread. Monitoring projects for viruses and other pathogens are essential for health protection and pandemic prevention. Given the high cost of single epidemic events, data produced by the virus monitoring projects may provide a substantial return on investment.

TÄNUSÕNAD

Tänan oma juhendajaid Mikk Puustusmaad ja Mihkel Vaherit, kes olid abivalmis ja panustasid käesoleva töö valmimisse oma energiat ja aega. Tänan ka Maido Remmi, kes andis mulle võimaluse töötada Bioinformaatika õppetoolis.

KASUTATUD KIRJANDUSE LOETELU

Abeles, S.R., Robles-Sikisaka, R., Ly, M., Lum, A.G., Salzman, J., Boehm, T.K., and Pride, D.T. (2014). Human oral viruses are personal, persistent and gender-consistent. ISME J. *8*, 1753–1767.

Allen, L.Z., Ishoey, T., Novotny, M.A., McLean, J.S., Lasken, R.S., and Williamson, S.J. (2011). Single virus genomics: a new tool for virus discovery. PloS One *6*, e17722.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Atmar, R.L., Opekun, A.R., Gilger, M.A., Estes, M.K., Crawford, S.E., Neill, F.H., and Graham, D.Y. (2008). Norwalk Virus Shedding after Experimental Human Infection. Emerg. Infect. Dis. *14*, 1553–1557.

Bazinet, A.L., and Cummings, M.P. (2012). A comparative evaluation of sequence classification programs. BMC Bioinformatics *13*, 92.

Belshaw, R., Pereira, V., Katzourakis, A., Talbot, G., Paces, J., Burt, A., and Tristem, M. (2004). Long-term reinfection of the human genome by endogenous retroviruses. Proc. Natl. Acad. Sci. U. S. A. *101*, 4894–4899.

Boggess, B. (2001). Mass Spectrometry Desk Reference (Sparkman, O. David). J. Chem. Educ. *78*, 168.

Bos, L. (1999). Beijerinck's work on tobacco mosaic virus: historical context and legacy. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *354*, 675–685.

Brown TA. Understanding a genome sequence. In: Brown TA, editor. Genomes. 2. Oxford: Wiley-Liss (2002)

Carroll, D., Daszak, P., Wolfe, N.D., Gao, G.F., Morel, C.M., Morzaria, S., Pablos-Méndez, A., Tomori, O., and Mazet, J.A.K. (2018). The Global Virome Project. Science *359*, 872–874.

Carter, J.B., and Saunders, V.A. (2007). Virology: principles and applications (Chichester, England ; Hoboken, NJ: John Wiley & Sons).

Chan, C.X., and Ragan, M.A. (2013). Next-generation phylogenomics. Biol. Direct *8*.

Chou, T.-C., Hsu, W., Wang, C.-H., Chen, Y.-J., and Fang, J.-M. (2011). Rapid and specific influenza virus detection by functionalized magnetic nanoparticles and mass spectrometry. J. Nanobiotechnology *9*, 52.

Clem, A.L., Sims, J., Telang, S., Eaton, J.W., and Chesney, J. (2007). Virus detection and identification using random multiplex (RT)-PCR with 3'-locked random primers. Virol. J. *4*, 65.

Cobián Güemes, A.G., Youle, M., Cantú, V.A., Felts, B., Nulton, J., and Rohwer, F. (2016). Viruses as Winners in the Game of Life. Annu. Rev. Virol. *3*, 197–214.

Dudas, G., Carvalho, L.M., Bedford, T., Tatem, A.J., Baele, G., Faria, N.R., Park, D.J., Ladner, J.T., Arias, A., Asogun, D., et al. (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature *544*, 309–315.

Fernandez-Cassi, X., Timoneda, N., Martínez-Puchol, S., Rusiñol, M., Rodriguez-Manzano, J., Figuerola, N., Bofill-Mas, S., Abril, J.F., and Girones, R. (2018). Metagenomics for the study of viruses in urban sewage as a tool for public health surveillance. Sci. Total Environ. *618*, 870–880.

Fuhrman, J.A. (1999). Marine viruses and their biogeochemical and ecological effects. Nature *399*, 541–548.

Geoghegan, J.L., and Holmes, E.C. (2017). Predicting virus emergence amid evolutionary noise. Open Biol. *7*, 170189.

Guliy, O.I., Zaitsev, B.D., Kuznetsova, I.E., Shikhabudinov, A.M., Balko, A.B., Teplykh, A.A., Staroverov, S.A., Dykman, L.A., Makarikhina, S.S., and Ignatov, O.V. (2016). Application of the method of electro-acoustical analysis for the detection of bacteriophages in a liquid phase. Biophysics *61*, 52–58.

Guliy, O.I., Zaitsev, B.D., Borodina, I.A., Shikhabudinov, A.M., Staroverov, S.A., Dykman, L.A., and Fomin, A.S. (2018). Electro-acoustic sensor for the real-time identification of the bacteriophages. Talanta *178*, 743–750.

Hayes, S., Mahony, J., Nauta, A., and van Sinderen, D. (2017). Metagenomic Approaches to Assess Bacteriophages in Various Environmental Niches. Viruses *9*, 127.

Haynes, M., and Rohwer, F. (2011). The Human Virome. In Metagenomics of the Human Body, K.E. Nelson, ed. (New York, NY: Springer New York), pp. 63–77.

Hu, Y., Zandi, R., Anavitarte, A., Knobler, C.M., and Gelbart, W.M. (2008). Packaging of a polymer by a viral capsid: the interplay between polymer length and capsid size. Biophys. J. *94*, 1428–1436.

Hurwitz, B.L., Ponsero, A., Thornton, J., and U'Ren, J.M. (2018). Phage hunters: Computational strategies for finding phages in large-scale 'omics datasets. Virus Res. *244*, 110–115.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., and Daszak, P. (2008). Global trends in emerging infectious diseases. Nature *451*, 990–993.

Katzourakis, A., and Gifford, R.J. (2010). Endogenous viral elements in animal genomes. PLoS Genet. *6*, e1001191.

Kerfeld, C.A., and Scott, K.M. (2011). Using BLAST to teach "E-value-tionary" concepts. PLoS Biol. *9*, e1001014.

Kim, D., Song, L., Breitwieser, F.P., and Salzberg, S.L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. *26*, 1721–1729.

King, A. M. Q., Adams, M. J., Cartens, E. B, ja Lefkowitz, E. J. (2012). Virus taxonomy: classification and nomenclature of viruses: ninth report of the International Committee on Taxonomy of Viruses (London ; Waltham, MA: Academic Press).

Koonin, E.V., Senkevich, T.G., and Dolja, V.V. (2006). The ancient Virus World and evolution of cells. Biol. Direct *1*, 29.

Kostic, A.D., Ojesina, A.I., Pedamallu, C.S., Jung, J., Verhaak, R.G.W., Getz, G., and Meyerson, M. (2011). PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat. Biotechnol. *29*, 393–396.

Kostic, A.D., Gevers, D., Pedamallu, C.S., Michaud, M., Duke, F., Earl, A.M., Ojesina, A.I., Jung, J., Bass, A.J., Tabernero, J., et al. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. *22*, 292–298.

Kunz, F., Matt, G., and Hackl, H. (1970). Plasma phospholipids in type IV hyperlipoproteinemia. Atherosclerosis *11*, 265–278.

Lam, T.T.-Y., Zhou, B., Wang, J., Chai, Y., Shen, Y., Chen, X., Ma, C., Hong, W., Chen, Y., Zhang, Y., et al. (2015). Dissemination, divergence and establishment of H7N9 influenza viruses in China. Nature *522*, 102–105.

Laue, M. (2010). Electron Microscopy of Viruses. In Methods in Cell Biology, (Elsevier), pp. 1–20.

Lefkowitz, E.J., Dempsey, D.M., Hendrickson, R.C., Orton, R.J., Siddell, S.G., and Smith, D.B. (2018). Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res. *46*, D708–D717.

Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., Lescot, M., Poirot, O., Bertaux, L., Bruley, C., et al. (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proc. Natl. Acad. Sci. U. S. A. *111*, 4274–4279.

Lin, J., Kramna, L., Autio, R., Hyöty, H., Nykter, M., and Cinek, O. (2017). Vipie: web pipeline for parallel characterization of viral populations from multiple NGS samples. BMC Genomics *18*, 378.

Loman, N.J., Constantinidou, C., Chan, J.Z.M., Halachev, M., Sergeant, M., Penn, C.W., Robinson, E.R., and Pallen, M.J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat. Rev. Microbiol. *10*, 599–606.

Longmire, A.G., Sims, S., Rytsareva, I., Campo, D.S., Skums, P., Dimitrova, Z., Ramachandran, S., Medrzycki, M., Thai, H., Ganova-Raeva, L., et al. (2017). GHOST: global hepatitis outbreak and surveillance technology. BMC Genomics *18*, 916.

Ly, M., Jones, M.B., Abeles, S.R., Santiago-Rodriguez, T.M., Gao, J., Chan, I.C., Ghose, C., and Pride, D.T. (2016). Transmission of viruses via our microbiomes. Microbiome *4*.

M, K., hasamy, and K, D.A. (2008). Evaluation of in vitro antibacterial property of seaweeds of southeast coast of India. Afr. J. Biotechnol. *7*, 1958–1961.

Martell, M., Esteban, J.I., Quer, J., Genescà, J., Weiner, A., Esteban, R., Guardia, J., and Gómez, J. (1992). Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. J. Virol. *66*, 3225–3229.

Martinez-Hernandez, F., Fornas, O., Lluesma Gomez, M., Bolduc, B., de la Cruz Peña, M.J., Martínez, J.M., Anton, J., Gasol, J.M., Rosselli, R., Rodriguez-Valera, F., et al. (2017). Single-virus genomics reveals hidden cosmopolitan and abundant viruses. Nat. Commun. *8*, 15892.

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2011). The human gut virome: inter-individual variation and dynamic response to diet. Genome Res. *21*, 1616–1625.

Mohd Hanafiah, K., Groeger, J., Flaxman, A.D., and Wiersma, S.T. (2013). Global epidemiology of hepatitis C virus infection: New estimates of age-specific antibody to HCV seroprevalence. Hepatology *57*, 1333–1342.

Mokili, J.L., Rohwer, F., and Dutilh, B.E. (2012). Metagenomics and future perspectives in virus discovery. Curr. Opin. Virol. *2*, 63–77.

Mullis, K.B., and Faloona, F.A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol. *155*, 335–350.

Musaji, A., Fahlman, R., and Charlton, C. (2016). Mass spectrometry of influenza virus using clinically available MALDI-TOF platform. J. Clin. Virol. *82*, S45.

Naccache, S.N., Federman, S., Veeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., Bouquet, J., Greninger, A.L., Luk, K.-C., Enge, B., et al. (2014). A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res. *24*, 1180–1192.

Ounit, R., Wanamaker, S., Close, T.J., and Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics *16*, 236.

Pavlov, R.M., Mikhaĭlov, L.G., and Vlasov, V.I. (1986). [Advantages and disadvantages of fibrobronchoscopy and rigid bronchoscopy in the diagnosis of lung tumors]. Probl. Tuberk. 45–46.

Payment, P., Lemieux, M., and Trudel, M. (1982). Bacteriological and virological analysis of water from four fresh water beaches. Water Res. *16*, 939–943.

Pierson, E.E., Keifer, D.Z., Selzer, L., Lee, L.S., Contino, N.C., Wang, J.C.-Y., Zlotnick, A., and Jarrold, M.F. (2014). Detection of late intermediates in virus capsid assembly by charge detection mass spectrometry. J. Am. Chem. Soc. *136*, 3536–3541.

Popgeorgiev, N., Temmam, S., Raoult, D., and Desnues, C. (2013). Describing the silent human virome with an emphasis on giant viruses. Intervirology *56*, 395–412.

Pride, D.T., Salzman, J., Haynes, M., Rohwer, F., Davis-Long, C., White, R.A., Loomer, P., Armitage, G.C., and Relman, D.A. (2012). Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. ISME J. *6*, 915–926.

Raquin, V., Wannagat, M., Zouache, K., Legras-Lachuer, C., Moro, C.V., and Mavingui, P. (2012). Detection of dengue group viruses by fluorescence in situ hybridization. Parasit. Vectors *5*, 243.

Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome *5*, 69.

Roingeard, P. (2008). Viral detection by electron microscopy: past, present and future. Biol. Cell *100*, 491–501.

Rosen, G.L., Reichenberger, E.R., and Rosenfeld, A.M. (2011). NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. Bioinforma. Oxf. Engl. *27*, 127–129.

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

Rudkin, G.T., and Stollar, B.D. (1977). High resolution detection of DNA-RNA hybrids in situ by indirect immunofluorescence. Nature *265*, 472–473.

Sievers, A., Bosiek, K., Bisch, M., Dreessen, C., Riedel, J., Froß, P., Hausmann, M., and Hildenbrand, G. (2017). K-mer Content, Correlation, and Position Analysis of Genome DNA Sequences for the Identification of Function and Evolutionary Features. Genes *8*, 122.

Simmonds, P., Adams, M.J., Benkő, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., et al. (2017). Consensus statement: Virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. *15*, 161–168.

Ssemadaali, M.A., Effertz, K., Singh, P., Kolyvushko, O., and Ramamoorthy, S. (2016). Identification of heterologous Torque Teno Viruses in humans and swine. Sci. Rep. *6*.

Suttle, C.A. (2007). Marine viruses--major players in the global ecosystem. Nat. Rev. Microbiol. *5*, 801–812.

Trauger, S.A., Junker, T., and Siuzdak, G. (2003). Investigating Viral Proteins and Intact Viruses with Mass Spectrometry. In Modern Mass Spectrometry, C.A. Schalley, ed. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 265–282.

Volpi, E.V., and Bridger, J.M. (2008). FISH glossary: an overview of the fluorescence in situ hybridization technique. BioTechniques *45*, 385–409.

Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. *15*, R46.

Woolhouse, M., Scott, F., Hudson, Z., Howey, R., and Chase-Topping, M. (2012). Human viruses: discovery and emergence. Philos. Trans. R. Soc. B Biol. Sci. *367*, 2864–2871.

Xu, Y., Falk, I.N., Hallen, M.A., and Fitzgerald, M.C. (2011). Mass spectrometry- and lysine amidination-based protocol for thermodynamic analysis of protein folding and ligand binding interactions. Anal. Chem. *83*, 3555–3562.

Zárate, S., Taboada, B., Yocupicio-Monroy, M., and Arias, C.F. (2017). Human Virome. Arch. Med. Res. *48*, 701–716.

Zeigler Allen, L., McCrow, J.P., Ininbergs, K., Dupont, C.L., Badger, J.H., Hoffman, J.M., Ekman, M., Allen, A.E., Bergman, B., and Venter, J.C. (2017). The Baltic Sea Virome: Diversity and Transcriptional Activity of DNA and RNA Viruses. MSystems *2*.

Zhang, X., Yue, L., Zhang, Z., and Yuan, Z. (2017). Establishment of a fluorescent in situ hybridization assay for imaging hepatitis B virus nucleic acids in cell culture models. Emerg. Microbes Infect. *6*, e98.

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

KASUTATUD VEEBIAADRESSID

Uniproti andmebaas (*UniProt reference proteomes*), kasutatud 16.05.2018, http://www.uniprot.org/proteomes/?query=*&fil=reference%3Ayes+AND+taxonomy%3A%22Viruses+%5B10239%5D%22

Raccaniello, 2010, http://www.virology.ws/2010/07/16/detection-of-antigens-or-antibodies-by-elisa/

Racaniello, 2013, http://www.virology.ws/2013/09/06/how-many-viruses-on-earth/

Siddell, 2018, https://microbiologysociety.org/publication/past-issues/imaging/article/why-virus-taxonomy-is-important.html

SIB Swiss Institute of Bioinformatics, https://viralzone.expasy.org

International Committee on Taxonomy of Viruses (ICTV), https://talk.ictvonline.org/taxonomy/

NCBI *Viral Genomes* andmebaas, kasutatud 12.03.2018, https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239

Scientific Center for Optical and Electron Microscopy, http://www.scopem.ethz.ch/gallery/02.html

Monis, 2012, https://www.eurofinsus.com/media/161936/detecting-virus-on-your-vines.pdf

LIHTLITSENTS

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Brigitta-Robin Raudne (30.10.1995)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
Erinevate töövoogude analüüs viiruste tuvastamisel,
mille juhendajad on Mikk Puustusmaa ja Mihkel Vaher,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-islisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseksTartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 28.05.2018