TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOINFORMAATIKA ÕPPETOOL

Taksonoomiliste profiilide visualiseerimine

Bakalaureusetöö

12 EAP

Leiki-Maria Arroval

Juhendaja PhD Maido Remm

TARTU 2025

INFOLEHT

Taksonoomiliste profiilide visualiseerimine

Metagenoomika on oluline mitmes valdkonnas, nagu meditsiin ja keskkonnateadus, võimaldades mõista mikroobikoosluste geneetilist ja funktsionaalset mitmekesisust. Töö eesmärk oli võrrelda taksonoomiliste profiilide visualiseerimise tööriistu ning hinnata, kuidas andmestiku ülesehitus ja programmide funktsioonid mõjutavad diagrammide kujundust. Krona diagramm oli informatiivne ja interaktiivne, Paviani Sankey diagrammid selged ja ühetasandilised, kuid kumbki ei võimaldanud värvimist. Tableau Public mullidiagrammid olid lihtsad, ent infovaesed. Mullidiagrammides kujutati liike, lugemeid ja patogeensust; toiduandmestikus eristati taimi ja putukaid värvuse alusel.

Märksõnad: metagenoomika, Krona, Sankey, mullidiagramm, taksonoomia, liigilise koostise visualiseerimine, Pavian, taksonoomiline profiil
CERCS kood: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Visualization of taxonomic profiles

Metagenomics plays an important role in various fields, such as medicine and environmental science, enabling the understanding of the genetic and functional diversity of microbial communities. The aim of this study was to compare tools for visualizing taxonomic profiles and to assess how data structure and program functionality influence diagram design. The Krona diagram was informative and interactive, while Pavian's Sankey diagrams were clear and single-layered, but neither tool supported color customization. The Tableau Public bubble charts were simple but lacked detail. The bubble charts depicted species, read counts, and pathogenicity; in the food dataset, plants and insects were distinguished by color.

Keywords: metagenomics, Krona, Sankey, bubble chart, taxonomy, visualization of species composition, Pavian, taxonomic profile
CERCS code: B110 Bioinformatics, medical informatics, biomathematics, biometrics

SISUKORD

KASTUTATUD LÜHENDID

SISSEJUHATUS

1. KIRJANDUSE ÜLEVAADE
    - 1.1. Metagenoomsed uuringud
        - 1.1.1. Metagenoomika kliinilistes katsetes
            - 1.1.1.1. Soolestiku mikrobioomi metagenoomika
            - 1.1.1.2. Hammaste mikrobioomi metagenoomika
            - 1.1.1.3. Vere (rakuvaba DNA) metageoomika
        - 1.1.2. Toiduohutuse metagenoomika
        - 1.1.3. Keskkonnaanalüüside metagenoomika
    - 1.2. Taksonoomilise profiili määramine
        - 1.2.1. Kraken
        - 1.2.2. MetaPhlAn
        - 1.2.3. BLAST
    - 1.3. Taksonoomilise profiili visualiseerimine
        - 1.3.1. Krona diagramm
        - 1.3.2. Sankey diagramm
        - 1.3.3. Mullidiagramm
2. EKSPERIMENTAALOSA
    - 2.1. Töö eesmärgid
    - 2.2. Materjal ja metoodika
        - 2.2.1. Krona diagrammide koostamine KronaTools-i abil
        - 2.2.2. Sankey diagrammide koostamine Paviani abil
        - 2.2.3. Mullidiagrammide koostamine Tableau Publicu abil
    - 2.3. Tulemused
        - 2.3.1. Krona diagrammid
        - 2.3.2. Sankey diagrammid
        - 2.3.3. Mullidiagrammid
    - 2.4. Arutelu

KOKKUVÕTE

SUMMARY

KASUTATUD KIRJANDUSE LOETELU

KASUTATUD VEEBIAADRESSID

KASUTATUD LÜHENDID

BAM – binaarne järjestusjoondusfail (ingl. k. *Binary Alignment Map*)

D3 - veebipõhine visualiseerimisraamistik, mis seob andmed otse HTML-i elementidega, kasutades brauseri sisseehitatud dokumentide objekti mudelit (ingl. k. *data- driven documents*)

HCMV – inimese tsütomegaloviirus (ingl. k. *human cytomegalovirus*)

HSL – värvitoon, küllastus ja heledus (ingl. k. *hue, saturation and lightness*)

HTML5 – hüperteksti märgendikeel (ingl. k. *HyperText Markup Language, version 5*)

LCA – lähim ühine eellane (ingl. k. *lowest common ancestor*)

METAREP – metagenoomika aruandlusplatvorm (ingl. k. *JCVI Metagenomics Reports*)

NCBI – Riiklik Biotehnoloogia Infokeskus (ingl. k. *National Center for Biotechnology Information*)

ORF – avatud lugemisraamid (ingl. k. *Open Reading Frames*)

Perl – interpreteeritav programmimiskeel (ingl. k. *Practical Extraction and Report Language*)

RDP – ribosomaalse andmebaasi projekt (ingl. k. *Ribosomal Database Project*)

RSF – radiaal ruumitäitmine (ingl. k. *radial space-filling*)

XHTML – veebilehtede loomiseks kasutatav keel (ingl. k. *eXtensible HyperText Markup Language*)

XML – laiendatav märgistuskeel (ingl. k. *eXtensible Markup Language*)

SISSEJUHATUS

Metagenoomika on oluliselt täiendanud meie teadmisi mikroobikooslustest, pakkudes põhjalikku arusaama nende geneetilisest ja funktsionaalsest mitmekesisusest erinevates ökosüsteemides.

Mikroobikooslused ei ole pelgalt keskkonna osa, vaid nad tegutsevad ka sümbiootiliselt, mõjutades otseselt oma peremeeste tervist ja talitlust. Kuna need kooslused on keerukad ja esinevad väga erinevates keskkondades, tuleb metagenoomikauuringud läbi viia täpselt ja metoodiliselt, et saadud tulemused kajastaksid uuritavate organismide ja keskkondade mitmekesisust.

Metagenoomika kasutus on viimastel aastatel kiiresti kasvanud, eriti haiguste ja mikroobikoosluste uurimisel. Selliste suurte ja keerukate andmestike analüüsimisel on oluline kasutada interaktiivseid visualiseerimisvahendeid, mis võimaldavad eristada näiteks patogeene taustamikroobidest ning võrrelda tulemusi eri proovide vahel. Ilma sobivate visuaalsete tööriistadeta võib olulise teabe leidmine osutuda keeruliseks.

Käesoleva töö eesmärk on uurida erinevaid taksonoomiliste profiilide visualiseerimise tööriistu ning tuua esile nende erinevused, sarnasused, tugevused ja puudused. Lisaks vaadeldakse, kuidas andmestiku muutused ja tarkvarade funktsionaalsus mõjutavad loodud diagrammide kujundust ja tõlgendatavust.

1. KIRJANDUSE ÜLEVAADE

1.1. Metagenoomsed uuringud

Metagenoomika uurib erinevate organismide (enamasti mikroobide) geneetilist materjali, mis on eraldatud ja analüüsitud ühest keskkonnaproovist. Selle lähenemise abil saab hinnata nii organismide koosseisu kui ka nende funktsionaalset rolli. Metagenoomikat kasutatakse sageli mikroobikooslustega seotud uuringutes (National Human Genome Research Institute, 2025). See on kultuurist sõltumatu tehnika, mis võimaldab tuvastada ja iseloomustada organismide mitmekesisust erinevates proovides (Pérez-Cobas *et al.*, 2020).

Metagenoomilises uuringus kogutakse proove erinevatest keskkondadest, näiteks mullast, ookeaniveest, esemete pindadelt vatitatampooni abil või suuremate organismide mikroobikooslustest, nagu inimkeha mikrobioom (Bouchot *et al*., 2014).

Segre (2025) märgib, et metagenoomi võib kujutada kui kogumit kõigi koos elavate organismide DNA-st. Seda protsessi võib võrrelda pusle kokkupanekuga – mitte ühe, vaid mitme erineva puslega, mis on samasse karpi segamini pandud. Metagenoomika eesmärk on need pusled üheaegselt kokku panna, et mõista iga organismi genoomi ning nende omavahelisi seoseid.

Varem uuriti mikroobe peamiselt isoleeritult – teadlased eraldasid ja kasvatasid üksikuid mikroorganisme, et analüüsida nende genoomi ja geenifunktsioone. Tegelikkuses elavad mikroobid kogukondadena, kus nad omavahel kas teevad koostööd või konkureerivad üksteisega (Bouchot *et al*., 2014).

Metagenoomika võimaldab terviklikult uurida mikroobikooslusi, kasutades tänapäevaseid järjestustehnoloogiaid ja arvutuslikke meetodeid, ilma et oleks vaja üksikuid mikroorganisme laboritingimustes eraldada või kasvatada. Viimase kümnendi jooksul on see valdkond oluliselt kaasa aidanud mikroobide ökoloogia, evolutsiooni ja mitmekesisuse mõistmisele ning kujunenud uurimissuunaks paljudes teaduslaborites (Thomas *et al*., 2012).

Proovi töötlus on metagenoomika üks kriitilisemaid etappe. DNA peab peegeldama kogu proovi mitmekesisust ning olema piisava kvaliteediga, et tagada usaldusväärne sekveneerimine. Iga proovitüüp vajab spetsiifilist töötlusprotokolli ning olemas on mitmeid tõhusaid DNA

eraldamise meetodeid. Suurtes projektides eelistatakse kasutada ühtset protokolli, et tulemused oleksid paremini võrreldavad.

1.1.1. Metageoomika kliinilistes proovides

1.1.1.1. Soolestiku mikrobioomi metagenoomika

Soolestiku mikrobioota koosneb enam kui tuhandest mikroobiliigist, mis aitavad kaitsta organismi patogeenide eest, toetavad immuunsüsteemi ja ainevahetust ning toimivad osaliselt ka endokriinse organina. Inimese seedetraktis elab keerukas ja dünaamiline mikrobioota, kuhu kuuluvad arheed, bakterid, viirused ja eukarüoodid (Wang *et al*., 2015).

MetaHITi konsortsium uuris 124 Euroopa inimese väljaheiteproove metagenoomika abil ning tuvastas esmakordselt inimese soolestiku mikrobioomis 3,3 miljonit mitte-korduvat geeni – see ületas inimese genoomi enda geenide arvu ligikaudu 150-kordselt. Üllatuslikult olid üle 99% tuvastatud geenidest bakteriaalsed, viidates enam kui tuhandele bakteriliigi esinemisele kogu uuritud rühmas.

Kuigi andmetöötluse võimalused on viimastel aastatel oluliselt arenenud, on metagenoomide taksonoomiline ja funktsionaalne tõlgendamine endiselt keeruline ülesanne. Genoomide täielik rekonstrueerimine võimaldab avastada seni teadmata mikroobiliike, kuid keerukamate bioloogiliste funktsioonide automaatne määratlemine nõuab siiani inimese ekspertiisi (Yen ja Johnson, 2021).

1.1.1.2. Hammaste mikrobioomi metagenoomika

Parodontiit ja hambakaaries on bakteriaalsest päritolust haigused, mille täpsemaks mõistmiseks ja diagnoosimiseks kasutatakse üha enam metagenoomikat. See meetod võimaldab analüüsida suu mikrobioomi koosseisu ja funktsioone, aidates tuvastada haigust põhjustavaid mikroorganisme ning mõista nende rolli biofilmi tekkes ja hapetootmises (Xu ja Gunsolley, 2014).

Sülg on suuõõne oluline bioloogiline vedelik, mis aitab säilitada nii suupiirkonna kui ka kogu organismi tervist. Seetõttu on seda uuritud kui potentsiaalset biomarkerite allikat, mis seonduvad suuhaiguste ning üldise tervisliku seisundiga. Kuigi sülg on näärmetest väljudes steriilne, sisaldavad süljeproovid mitmekesist mikroobide kogumit, mille koosseis erineb hammaste ja igemete pinnal paiknevatest mikroobsetest biofilmidest (Belstrøm *et al*., 2017).

1.1.1.3. Vere (rakuvaba DNA) metageoomika

Vereülekandeid kasutatakse sageli eluohtlike seisundite raviks, kuid nende kaudu võivad edasi kanduda ka nakkushaigused. Isegi tervete doonorite veres võivad esineda patogeenid, mis võivad üle kanduda vereülekande või plasmaderivaatide kaudu. Inimese tsütomegaloviirus (HCMV), tuntud ka kui inimese herpesviirus 5, on laialt levinud β-herpesviirus, mis põhjustab mitmesuguseid infektsioone. Uuringud on näidanud, et HCMV võib levida vereülekandega, põhjustades esmase või korduva infektsiooni (Xu *et al*., 2020).

Febriilne neutropeenia on ägeda leukeemia patsientidel sage tüsistus, mille puhul nakkusetekitajate kiire ja täpne tuvastamine on keeruline. Traditsioonilised meetodid, nagu verekultuurid, on sageli ebatäpsed, eriti antibiootikumravi ajal. Metagenoomiline analüüs pakub siin tõhusat alternatiivi, võimaldades infektsioonide põhjuseid tuvastada täpsemalt ja kiiremini, mis aitab parandada raviotsuseid ning vähendada antibiootikumide ebaotstarbekat kasutamist (Qi *et al*., 2023).

1.1.2. Toiduohutuse metagenoomika

Toit sisaldab mitmesuguseid mikroobe, mis võivad mõjutada inimese tervist. Enamasti on uuritud, kuidas toidu koostis (näiteks toitainete koostis) mõjutab organismi, kuid viimastel aastatel on selgunud, et ka toidus leiduvad mikroobid võivad jõuda inimese kehasse ning saada osaks tema mikrobioomist. Seetõttu on oluline uurida põhjalikumalt, kuidas toidust pärit mikroobid mõjutavad inimese mikrobioomi ja tervist — nii indiviidi kui ka mikroobitüvede tasandil (Carlino *et al*., 2024).

Hoolimata teaduse ja tehnoloogia arengust toiduvaldkonnas, on toidupatogeenidega seotud terviseprobleemid globaalselt endiselt väga tõsised: igal aastal registreeritakse sadu miljoneid haigestumisi ning surmajuhtumite arv ulatub sadadesse tuhandetesse. Toiduga seotud haigusriske aitab paremini ennetada see, kui tootjad ja järelevalveasutused suudavad kiiresti ja täpselt tuvastada proovis esinevaid patogeene. Ainult nii on võimalik eri allikatest pärinevaid mikroorganisme omavahel võrrelda ja mõista, kuidas saastumine toimub või nakkus levib (Billington *et al*., 2022).

Metagenoomika võimaldab toidutööstuses uurida nii toodete mikroobikoosseisu kui ka kontrollida nende autentsust ja päritolu. Samuti saab seda meetodit kasutada toiduohutuse seisukohalt oluliste riskide tuvastamiseks — näiteks võib see aidata avastada allergeene,

toksiine või soovimatuid mikroobe, sealhulgas potentsiaalseid patogeene, mis võivad esineda nii tooraines kui ka tootmisprotsessi käigus.

1.1.3. Keskkonnaanalüüside metagenoomika

Ohutu joogivee kättesaadavus on rahvusvaheliselt tunnustaud põhiõigus, kuid üle kahe miljardi inimese kogu maailmas ei oma endiselt ligipääsu turvalisele joogiveevarustusele (Brumfield *et al*., 2020).

Veekogudes esineb mitmesuguseid mikroorganisme, sealhulgas fekaalset päritolu saaste ja antibiootikumiresistentsust kandvaid liike, mis võivad kujutada tõsist ohtu inimeste tervisele. Näiteks püsivad veeproovides sageli *Giardia* ja *Cryptosporidium*, mis viitavad fekaalsete ainete esinemisele ja on seotud vee kaudu levivate (Urrea *et al*., 2025).

Laiemalt aitavad veekogude mikroobikoosluste analüüsid paremini mõista ökosüsteemide toimimise põhimõtteid ning hinnata sellega seotud tervise- ja keskkonnaohte. Füüsikaliste ja bioloogiliste andmete koosvaatlus on oluline, kuna veekeskkonnad võivad kujuneda saaste leviku kanaliks ning seetõttu vajavad need pidevat jälgimist ja hindamist.

1. 2. Taksonoomiliste profiilide määramise tarkvara

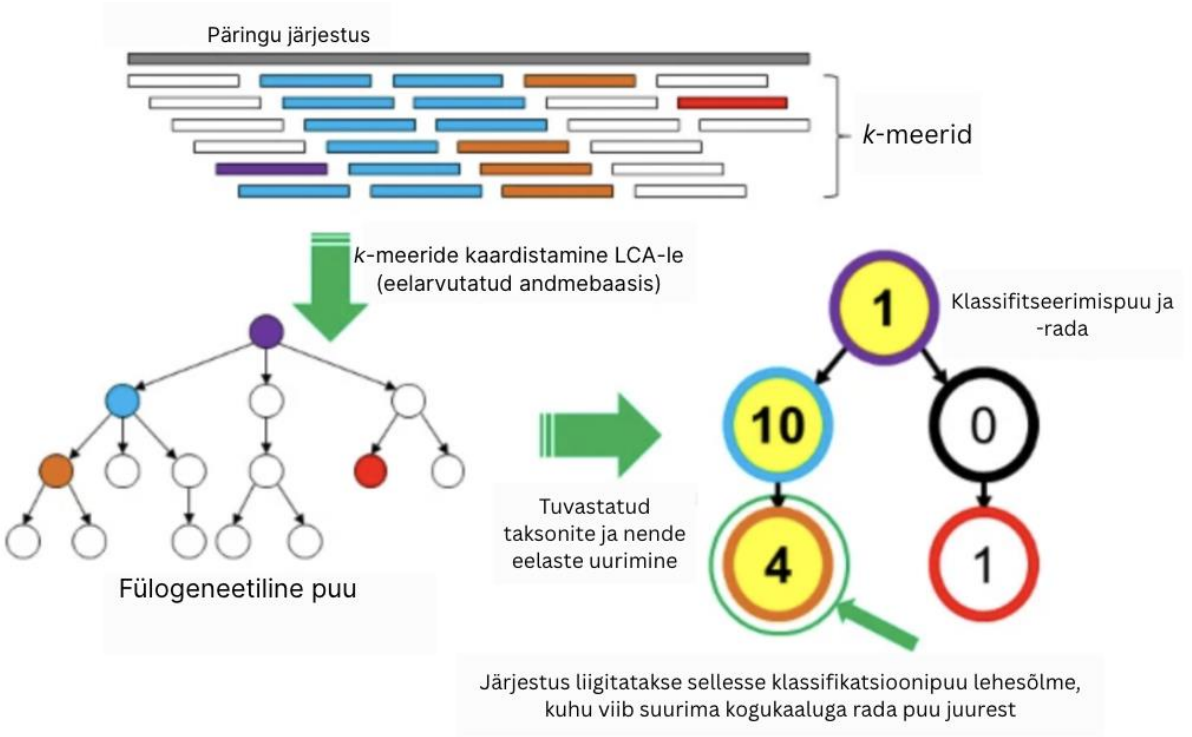
1. 2. 1. Kraken

Kindla pikkusega järjestikused nukleotiidi- või aminohappejärjestused, mida nimetatakse *k*-meerideks, on olulised tööriistad paljudes bioinformaatilistes analüüsides. *k*-meeridel põhinevaid algoritme kasutatakse laialdaselt nii genoomika kui ka proteoomika valdkonnas, kuna need võimaldavad tõhusat järjestusandmete otsingut ja analüüsi (Moeckel *et al.*, 2024).

Kraken on taksonoomilise profiili määramise tarkvara, mis kasutab *k*-meeride põhist andmebaasi, kus iga *k*-meer on seotud selle esinemise alusel taksonoomilise puu lähima ühise esivanemaga (Wood ja Salzberg, 2014). Iga järjestuses esinev *k*-meer võrreldakse andmebaasiga ning saadud taksonite alusel määratakse võimalik kuuluvus taksonoomilises hierarhias. Kui järjestuses ei leidu ühtegi sobivat *k*-meeri, jääb see klassifitseerimata.

Kraken klassifitseerib järjestusi, kaardistades iga *k*-meeri taksonoomilise puu sõlmele, kus see esineb. Kõik vastavad taksonid ja nende eelastmed moodustavad kärbitud alampuu üldisest fülogeneetilisest puust. Iga harule määratakse kaal, mis vastab selle taksoniga seotud *k*-meeride

arvule. Juurest lehesõlmeni kulgeva haru skoor arvutatakse sõlmede kaalude summana. Suurima skooriga haru määrab järjestuse taksonoomilise kuuluvuse, kus lehesõlm tähistab kõige spetsiifilisemat võimalikku taksonit (vt joonis 1).



Joonis 1. Krakeni järjestuse klassifitseerimine protsess. Kohandatud allikast: Wood ja Salzberg, 2014.

Bracken (ehk *Bayesian Reestimation of Abundance after Classification with KrakEN)* on tööriist, mis täiendab Krakeni lugemispõhist klassifikatsiooni, pakkudes statistilist täpsustust mikroobide esinemissageduse hindamisel (Lu *et al*., 2017). See kasutab lisaks genoomi struktuurset infot, et paremini eristada üksteisele geneetiliselt sarnaseid liike ja arvutada nende suhtelist esinemust liigi-, perekonna- või kõrgemal taksonoomilisel tasemel. Selline lähenemine muudab Bracken'i sobivaks keeruliste mikroobikoosluste kvantitatiivseks analüüsiks.

1. 2. 1. MetaPhlAn

MetaPhlAn (*Metagenomic Phylogenetic Analysis*) on tööriist, mis võimaldab määrata mikroobikoosluse koosseisu ja hinnata nende suhtelisi osakaale otse metagenoomsetest järjestusandmetest (Segata *et al*., 2012). Erinevalt üldisematest lähenemistest keskendub MetaPhlAn kindlalt määratletud markerijärjestustele – geenilõikudele, mis on omased konkreetsetele mikroobirühmadele. Kuna markerite valik on sihipärane ja piiratud, võimaldab see meetod kiiret ja täpset liigitasemel või kõrgemal taksonoomilisel tasemel klassifitseerimist, ka suurte andmehulkade korral. Markerid hõlmavad ka erinevaid funktsionaalseid kategooriaid, võimaldades teha järeldusi mikroobide potentsiaalsete rollide kohta ökosüsteemis.

MetaPhlAn ei keskendu üksnes järjestuste loendamisele, vaid püüab anda realistlikuma ülevaate mikroobikoosluse struktuurist, hinnates ka mikroobi rakkude arvu proovis. Selleks võetakse arvesse iga mikroobirühma tüüpilise genoomi suurust. Kui tööriista testiti simuleeritud andmestikel, osutus MetaPhlAn täpsemaks ja jõudluselt paremaks kui mitmed teised olemasolevad tööriistad, sealhulgas PhymmBL, mis oli täpne, kuid oluliselt aeglasem.

MetaPhlAn on loodud töötama tõhusalt ka piiratud arvutusressursside korral – näiteks suudab see töödelda sadu järjestusi sekundis ka tavapärases arvutis. Selline efektiivsus on võimaldanud esmakordselt viia läbi suuremahulisi liikide tasemel analüüse ka pärismaailma andmestikega.

1. 2. 1. BLAST

BLAST (*Basic Local Alignment Search Tool*) on järjestuste võrdlemise tööriist, mida saab kasutada nii veebipõhiselt kui ka lokaalselt käsurealt (Altschul *et al*., 1997). Tööriistal on mitu versiooni, mis võimaldavad võrrelda DNA, RNA või valgu järjestusi sobiva tüüpi andmebaasidega. BLAST kasutab heuristilist algoritmi, mis otsib esmalt lühikesi vastavusi (nn *hits*) päringu ja andmebaasi järjestuste vahel ning laiendab neid, et leida suurema kattuvusega piirkonnad. BLAST on loodud homoloogsete järjestuste otsimiseks, kuid seda on võimalik kasutada ka taksonoomilise profiili määramiseks.

Üheks BLAST-i tugevuseks on selle võime tuvastada sarnasusi järjestuste vahel isegi siis, kui need on evolutsiooniliselt oluliselt erinevad. Samas on selle täpsuse puuduseks aeglane töötluskiirus (Shah *et al*., 2018).

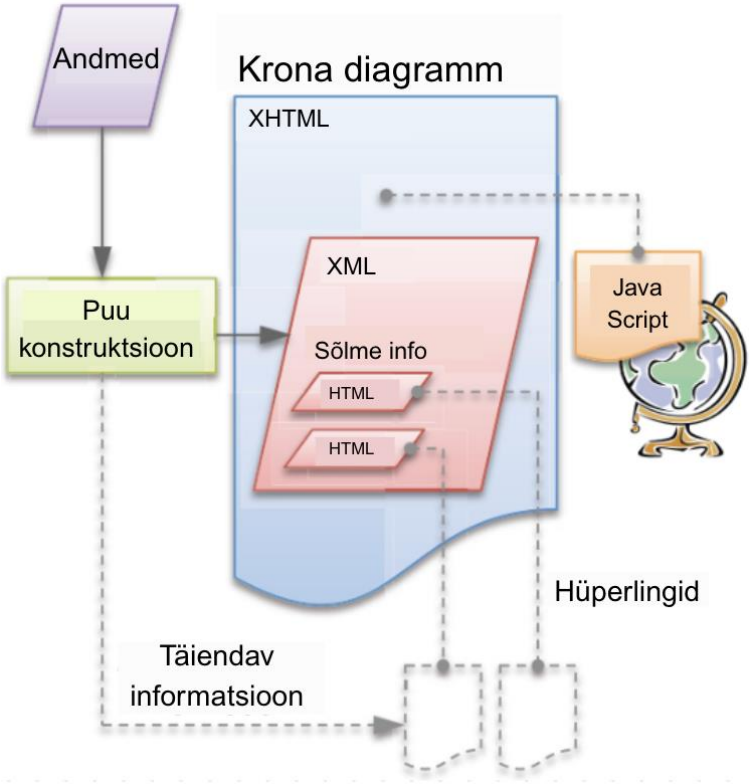
1.3. Taksonoomilise profiili visualiseerimine

1.3.1. Krona diagramm

Krona diagramm (tuntud ka kui *sunburst chart*, *ring chart* ja *radial treechart*) sobib hästi hierarhiliste andmete kujutamiseks, sest see koosneb mitmest kontsentrilisest rõngast, kus iga rõngas tähistab üht hierarhiataset (Aplakidou *et al*., 2024).

Selleks, et tagada tööriista laialdasem kasutatavus ning kohandatavus kiiresti areneva metagenoomika valdkonna vajadustega, on Krona visualiseerimissüsteem välja töötatud tuginedes HTML5 standardile (Ondov *et al*., 2011).

Krona töötab veebibrauseris, kasutades HTML5 ja JavaScripti, mis võimaldab visualiseeringuid kuvada lokaalselt ilma täiendava tarkvara paigalduseta. Diagramm salvestatakse ühe failina, mis sisaldab nii andmeid kui ka vajalikku koodi, tagades kohese kuvamise brauseris. Paindlik ülesehitus saavutatakse, kui diagrammiandmed salvestatakse XML-vormingus ja lisatakse XHTML-dokumenti, mis viitab veebis asuvale JavaScripti liidesele. Kui veebibrauser avab selle XHTML-faili, loeb JavaScripti kood XML-andmed sisse ja loob nende põhjal Krona diagrammi (vt joonis 2).



Joonis 2. Krona ülesehitus. Kohandatud allikast: Ondov *et al*., 2011.

Krona töötab paindlikult, pakkudes sujuvat värskendamist ja täiendusi ning sisaldab tööriistu nagu KronaTools, mis võimaldavad andmete importimist erinevatest bioinformaatika tööriistadest. KronaTools-pakett sisaldab Perlis kirjutatud skripte, millega saab diagramme koostada ja visualiseerida. Taksonoomilised klassifikatsioonid, sealhulgas hierarhilised, on saadaval erinevates tööriistades, nagu RDP Classifier, Phymm/PhymmBL, MG-RAST ja Galaxy, mis võimaldavad nende importimist. Lisaks saab kasutada BLAST-tulemusi NCBI andmebaasist või METAREP metagenoomika andmekogust.

BLASTI-i toorandmete puhul põhineb klassifikatsioon parimate vastepunktide vähimal ühisel esivanemal. Andmed kaardistatakse taksonoomiapuule, mis laaditakse automaatselt NCBI taksonoomiabaasist ja indekseeritakse.

Kui andmeid imporditakse RDP-st ja PhymmBL-ist, saab iga sõlme määrdumisvastase omaduse visualiseerimiseks kasutada värvigradiendi abil esitatud infot. MG-RAST-i, METAREP-i ja BLAST-i andmete sõlmed värvitakse näiteks keskmise e-väärtuse logaritmi või kattuvuse protsendi põhjal.

Krona diagramm meenutab tavalist sektordiagrammi, kuid lisab sellele hierarhilise struktuuri, kus iga takson on esindatud omaette sektorina. Nende sektorite suurus ei ole ainult visuaalselt esindatud, vaid määratakse nurkade järgi. Taksoni sektori suurus peegeldab sellele omistatud järjestuste hulka, väiksemad sektorid selle sees näitavad, kuidas andmed edasi alamtasanditeks jagunevad. Selline struktuur võimaldab meil kiiresti mõista, kuidas erinevad tasemed on omavahel seotud ja kuidas need omavahel suuruse järgi jagunevad.

Kuna taksonoomilised tasemed on sageli üsna erineva suurusega, siis sorteeritakse need diagrammil suuruse järgi, et neid oleks lihtsam võrrelda. Diagrammides kasutatakse HSL-mudelit, kus värvid jaotuvad taksonite vahel lähtudes nende hierarhilisest positsioonist ja kvantitatiivsest osakaalust. Nii luuakse visuaalne järjepidevus, mis aitab sarnaseid harusid hõlpsamini eristada, vastavalt nende suurusele. See tagab visuaalse järjepidevuse, eriti kui osa harusid on andmestikus kvantitatiivselt domineerivad. See muudab erinevate tasemete ja rühmade võrdlemise lihtsamaks.

Visuaalse selguse säilitamiseks võimaldab Krona kuvada järjestikuseid taksonitasemeid kompaktsemalt. Vajadusel saab hierarhiat hõlpsalt laiendada või kokku tõmmata, et vältida liigset detailirohkust ja säilitada ülevaatlikkus. Vaadet on võimalik veelgi lihtsustada, piirates taksonoomilise sügavuse näiteks kuue tasemeni. Samas rõhutatakse sügavamaid, olulisemaid tasemeid (näiteks liik), vähendades vahepealsete tasemete märgistuspinda. Teksti paigutus on optimeeritud nii, et lehtsõlmede märgistused asuvad radiaalselt, sisemiste sõlmede omad tangentsiaalselt, võimaldades tihedat ja loetavat esitust ka siis, kui sügavamate tasemete sektorid hakkavad kuju poolest lähenema ristkülikutele.

Radiaalsetes kujutistes jagatakse ringi sektoriteks järjestikuste taksonoomiliste tasemete alusel, mille tulemusel muutuvad sügavamad sõlmed järjest kitsamaks ja kuju poolest lähemale ristkülikutele. Kui kogu diagrammi lihtsalt suumida tavapärases koordinaatruumis, kaoks nende sõlmede nurgaline jaotus, mis on RSF-kujutiste ruumitõhususe ja intuitiivsuse oluline eeldus.

Krona kasutab suumimise puhul polaarset koordinaatruumi, et vältida kalduvust muuta diagrammi struktuuri. Suumimisel tõstetakse esile keskne takson, muutes selle visuaalselt silmatorkavamaks, samas kui teised osad jäävad taustale. Selline visuaalne lähenemine säilitab ülevaate kogu hierarhiast, võimaldades samal ajal keskenduda konkreetsele alamhulgale. Lõplik suumimisvaade säilitab kogu informatsioonitiheduse, võimaldades uurida ka sügavamaid taksonoomilisi tasemeid.

Hierarhias ülespoole liikumine toimub sarnaselt: klõpsates kas kesksel ülemisel sõlmel või kõrvalasuvatel koonddiagrammidel, taastub eelnev ülevaade pöördanimatsiooni abil.

Krona võimaldab lisaks taksonite arvulisele jaotusele visualiseerida ka muid tunnuseid, näiteks määramiskindlust või kategoorilisi tunnuseid. Sõlmi saab värvida kas käsitsi määratud kategooriate alusel või kasutada arvuliste muutujate korral värvigradiente, näiteks punase-rohelise skaala alusel, mis näitab klassifitseerimise usaldusväärsust.

Selline taksonoomia visuaalne esitlus on väärtuslik metagenoomi analüüsimisel. Radiaalne struktuur toetab taksonitevaheliste seoste intuitiivset tajumist ja aitab esile tuua mikroobikoosluste koostise ja esinemismustrid (Aplakidou *et al*., 2024).

1.3.2. Sankey diagramm

Pavian on interaktiivne visualiseerimistööriist, mis võimaldab uurida metagenoomika klassifitseerimistulemusi, mis on saadud näiteks tööriistadega nagu Kraken, KrakenUniq, Centrifuge ja MetaPhlAn (Breitwieser *et al*., 2019). See sobib hästi nii üksikute proovide kui ka mitme proovi tulemuste võrdlemiseks, aidates tuvastada mikroobikoosluste erinevusi. Paviani eripäradeks on Sankey diagramm ja genoomi joondusvaatur, mis toetavad tulemuste paremat mõistmist ka keerukates andmestikes.

Paviani Sankey diagrammid pakuvad visuaalselt selget ja tõhusat liidest keerukate metagenoomsete andmestike analüüsimiseks ja võrdlemiseks. Ühe proovi vaates kasutatakse vaikimisi Sankey diagrammi, kus järjestuste vool liigub fülogeneetilise puu juuretasemelt spetsiifilisemate taksonoomiliste tasemeteni. Voolu laius on proportsionaalne järjestuste arvuga, tuues visuaalselt esile kõige arvukamad taksonoomilised rühmad (Breitwieser *et al*., 2019). Sõlmed tähistavad taksoneid või muid kategooriaid, mida ühendavad vood ehk

järjestuste arvuga proportsionaalsed jooned; mida laiem joon, seda suurem on vastava rühma esindatus proovis (Aplakidou *et al*., 2024).

Võrdlusvaade võimaldab mitme proovi tulemusi tabelina võrrelda, kus järjestused esindavad taksoneid ja veerud proove. Tabelit saab filtreerida näiteks liigitasemele ning tulemusi sorteerida lisamõõdikute (näiteks protsent, z-skoor või järjestus) põhjal (Breitwieser *et al*., 2019).

Lisaks sisaldab Pavian joondusvaaturit, mis kuvab BAM-failide põhjal, kuidas järjestused genoomil paiknevad. See aitab hinnata, kas suur järjestuste arv viitab tegelikule kohalolule või võib olla näiteks saastest tingitud. Samuti on võimalus alla laadida viidatud RefSeq genoomikogumid.

Pavian on üles ehitatud R-is Shiny rakendusena, mis võimaldab interaktiivset andmeanalüüsi veebibrauseri kaudu. Visualiseerimiseks kasutatakse D3-põhiseid jooniseid, näiteks Sankey diagramme, ning JavaScripti teeke nagu networkD3 (Breitwieser *et al*., 2019). D3 on veebipõhine visualiseerimisraamistik, mis seob andmed otse HTML-i elementidega, kasutades brauseri sisseehitatud dokumentide objekti mudelit (DOM). Selle lähenemine võimaldab visualiseerida andmeid paindlikult ja selgelt, pakkudes dünaamilist sisu muutmist ning tugevat tuge animatsioonide ja interaktsioonide loomiseks (Bostock *et al*., 2011). Joondusvaates BAM-failide laadimiseks kasutatakse Rsamtools paketti. Rakendust saab käivitada otse R-keskkonnast või paigaldada serverisse (Breitwieser *et al*., 2019).

1.3.3. Mullidiagramm

Mullidiagramm on visuaalne tööriist, mis võimaldab esitada kolme mõõtmega andmeid kahemõõtmelisel tasapinnal (Aplakidou *et al*., 2024). Iga ring ehk mull tähistab ühte andmepunkti: selle horisontaalne ja vertikaalne asukoht kajastavad kahe muutuja väärtusi, samas kui ringi suurus peegeldab kolmanda muutuja suurust.

Bioloogias kasutatakse mullidiagramme näiteks liikide arvukuse võrdlemiseks eri elupaikades: mullide asetus võib kajastada keskkonnategureid, samas kui mullide suurus peegeldab liigi levimust või esinemissagedust.

Selline kujutus aitab esile tuua mustreid, seoseid ja võimalikke ökoloogilisi trende. Näiteks võib mulli suurus olla logaritmiliselt skaleeritud, et kajastada teatud geenide — näiteks avatud lugemisraamide (ORFide) — arvu, mis on seotud konkreetse taksoniga.

2. EKSPERIMENTAALOSA

2.1. Töö eesmärgid

Töö eesmärgid on järgmised:

a) visualiseerida reaalsetest andmestikest (sepsise andmestik ja toidu metagenoomika andmestik) saadud liikide klassifitseerimise tulemused kolme erineva programmiga;
b) võrrelda visualiseerimise tulemusi, hinnates kui hästi toovad nad esile ühe või teise andmestiku olulisi aspekte;
c) hinnata iga tööriista funktsionaalsuust ja kasutusmugavust.

2.2. Materjal ja metoodika

Visualiseerimiste aluseks kasutati 16 sepsisega patsiendi proovi ning üht simuleeritud toiduproovi. Sepsisega seotud andmestik sisaldas kaheksat cfDNA-proovi (vereseerumis või plasmas vabalt leiduvad DNA fragmendid) ning kaheksat pelletiproovi, mis sisaldasid tsentrifuugimisel saadud tahket setet. Putuka- ja taimseid komponente sisaldav simuleeritud toiduproov loodi Marten Rikbergi bakalaureusetöö raames.

2.2.1. Krona diagrammide koostamine KronaTools-i abil

Krona diagrammide loomiseks kasutati GitHubist kättesaadavat juhendit ja KronaTools paketti. Tarkvara paigaldati Linuxi terminali kaudu bioinformaatika töögrupi serveris mamba.ebc.ee kasutaja kodukataloogi. Sisendfailides olid vajalikud NCBI ID-d, taksonoomilised määrangud ja vastavad lugemite arvud.

2.2.2. Sankey diagrammi koostamine Paviani abil

Sankey diagrammi genereerimiseks kasutati Docker'i platvormi. Docker võimaldab rakendusi konteinerites isoleeritult käivitada, mis lihtsustab nende haldamist ja kasutuselevõttu (Docker, n.d.). Pärast Docker'i paigaldamist otsiti üles sobiv Paviani konteiner (florianbw/pavian) ja käivitati see. „Advanced options" alt määrati pordiks 0, mis võimaldas süsteemil automaatselt sobiv port leida (antud juhul 55001:80). Paviani veebiliidesesse laaditi sisendiks Bracken'i tulemused, mille põhjal sai andmeid edasi filtreerida ja visualiseerida.

2.2.3. Mullidiagrammide koostamine Tableau Publicu abil

Mullidiagrammide loomiseks kasutati Tableau Public tarkvara, mis võimaldab erinevat tüüpi graafilisi visualiseeringuid. Sisendina kasutati liigitasemel taksoneid ja nendele vastavaid

lugemite arve. Andmete kujutamiseks valiti sobiv mullidiagrammi formaat, mis võimaldas taksonite kvantitatiivseid suhteid visuaalselt võrrelda.

2.3. Tulemused

2.3.1. Krona diagrammid

Krona diagrammi oluliseks eripäraks ja tugevuseks on selle interaktiivne kasutajaliides, mis võimaldab liikuda süvitsi igasse taksonoomilisse tasandisse. Diagramm kujutab kõiki taksoneid protsentidena, arvutatuna lugemite koguarvu ja iga konkreetse taksoni osakaalu põhjal. Kuigi Krona diagramm on infoküllane ja kasutajasõbralik, võib see suurte andmemahtude korral muutuda keerukaks – väikese osakaaluga taksoneid võib olla raske märgata.

Oluliseks puuduseks on kohandamisvõimaluste piiratus, näiteks ei saa muuta värviskeemi vastavalt kasutaja soovile.

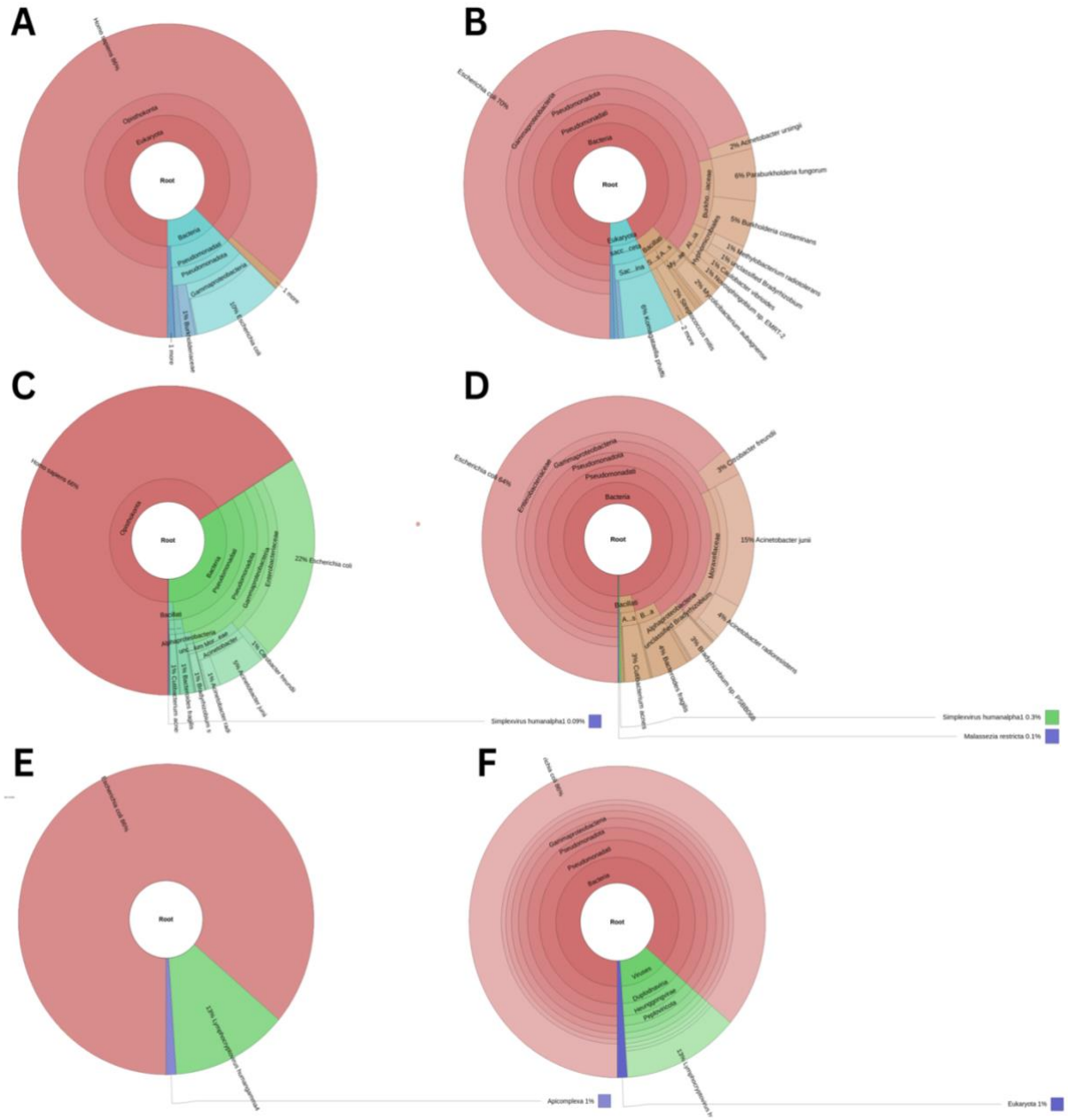
Krona diagrammi visuaalne ülesehitus tõstab automaatselt esile suurima osakaaluga taksonid. Lisaks võimaldab „collapse“ funktsioon liigselt detailset infot peita, säilitades üldvaate loetavuse. Diagrammi interaktiivsus võimaldab ka kuvada kõrvaldiagramme, mis näitavad valitud taksoni osakaalu kogu andmestikus.

Joonis 3 sisaldab Krona diagrammid kolmest sepsiseproovist, mis põhinevad cfDNA või pelletiproovide andmestikel. Diagrammide valikuga demonstreeritakse Krona funktsionaalsust ja illustreeritakse, kuidas inimese järjestuste eemaldamine mõjutab tulemuste visuaalset loetavust.

- Diagrammid A ja B pärinevad samast sepsise cfDNA proovist. Diagrammis A moodustab inimese DNA 86% kogutulemusest. Inimese järjestuste eemaldamise järel (diagramm B) muutub taksonoomiline jaotus märgatavalt: näiteks suureneb *Escherichia coli* osakaal 10% → 70%. Samuti muutuvad nähtavaks taksonid, mis varem jäid varju, näiteks *Komagataella phaffii* (pärmseen, kuulub *Ascomycota* hõimkonda ja *Saccharomycetes* klassi) ning erinevad *Alphaproteobacteria* klassi esindajad.
- Diagrammid C ja D esindavad ühte sepsise pelletiproovi. Diagrammis C moodustab inimese DNA 66% andmestikust. Lisaks on nähtav *Simplexvirus humanalpha1* (Herpes simplex virus 1), mis esineb 0,09% ulatuses. Inimese eemaldamisega (diagramm D) tõuseb *E. coli* osakaal 22% → 64% ning viirus suureneb 0,3%-ni. Lisaks ilmub

nähtavale *Malassezia restricta*, mis oli enne vaid 0,04% (nüüd 0,1%). Kuigi erinevus on vaid 0,06 protsendipunkti, tähistab see suhtelist kasvu 60% – see rõhutab, kui olulist mõju võib inimese järjestuste eemaldamine avaldada väikeste taksonite puhul.

- Diagrammid E ja F andmestikud pärinevad samast cfDNA proovist. Erinevus seisneb selles, et diagramm E-l on kasutatud "collapse" funktsiooni, mis võimaldab kuvada vähem taksoneid. Näiteks tõstetakse esile *Apicomplexa* hõimkond (1% osakaal). Diagrammil F on sama piirkond esitatud üldisemal tasemel *Eukaryota* domeenina, rõhutades klassifikatsioonipuu ülemist tasandit.



Joonis 3. Krona diagrammid sepsiseproovide võrdlemiseks. Diagrammipaarid A–B ja C–D põhinevad vastavalt samal cfDNA- ja pelletiproovil; A ja C sisaldavad inimese järjestusi, B ja

D neist puhastatud versioone. Diagrammid E ja F on samast proovist – E-s on kasutatud „collapse“ funktsiooni, F-s kuvatakse kogu taksonoomiline hierarhia.

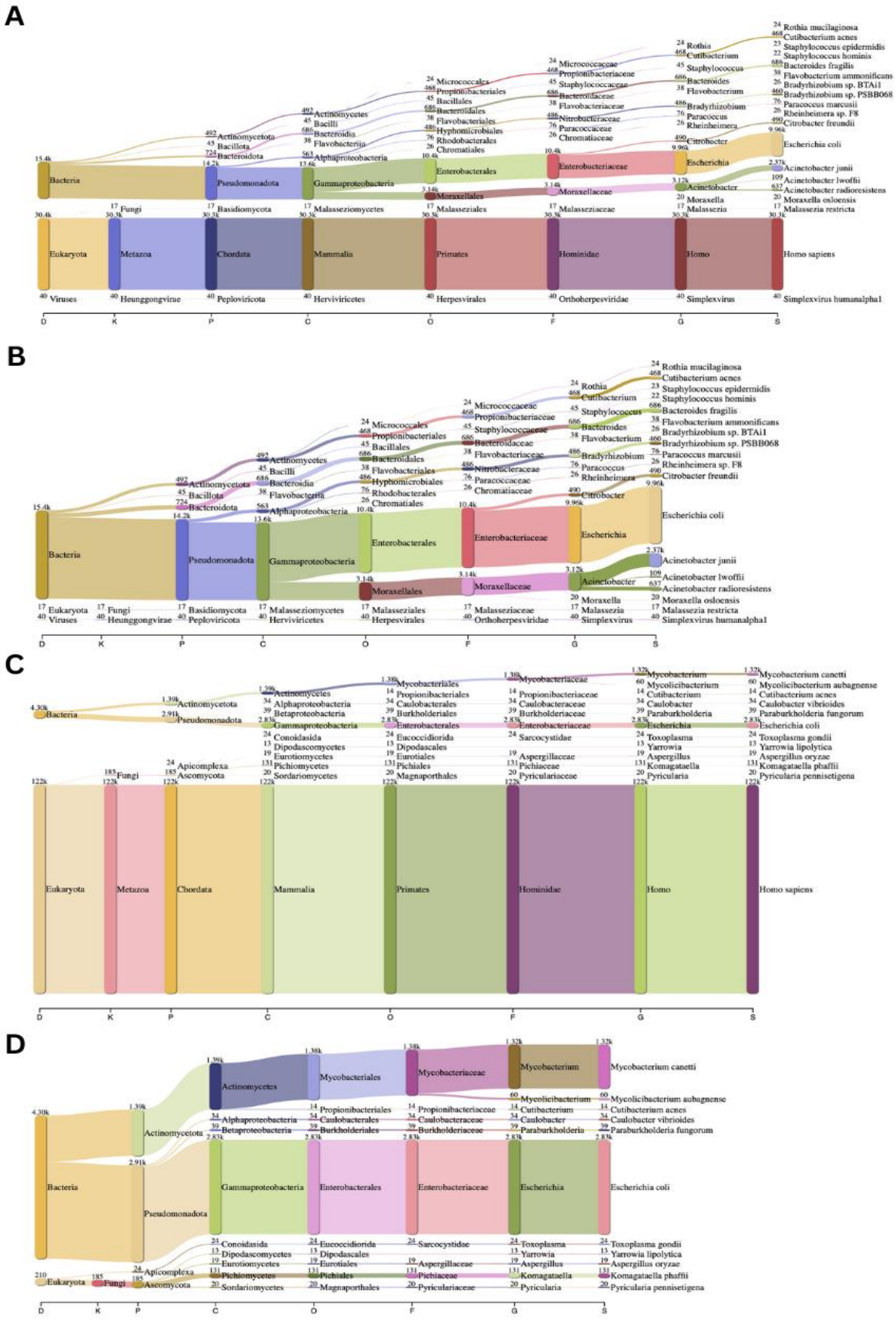
2.3.2. Sankey diagrammid

Sankey diagramm võimaldab visualiseerida kõiki taksoneid ja nende lugemite arve ühel joonisel. Taksonid on esitatud "vooludena" (lainetena), mille laius on proportsionaalne vastava taksoni lugemite arvuga – mida laiem voog, seda suurem esindatus andmestikus. Selline esitlus muudab tulemuste ülevaatamise visuaalselt intuitiivseks, kuid väikeste erinevuste hindamine võib olla keeruline, kui lugemite arvud on omavahel sarnased.

Paviani tööriist pakub mitmeid konfigureerimisvõimalusi, sealhulgas taksonoomiliste tasemete filtreerimist (näiteks liik, perekond, hõimkond), sõlmede vahemaa ja suuruse muutmist, joonise üldkõrguse ja läbipaistvuse reguleerimist ning siltide asukoha kohandamist. Klõpsates taksonile, saab muuta ka selle positsiooni diagrammil. See muudab diagrammi paindlikuks ja kohandatavaks vastavalt kasutaja vajadustele. Üheks piiranguks on see, et taksonite värviskeemi ei ole võimalik kasutaja poolt muuta.

Joonis 4 sisaldab nelja Sankey diagrammi, mis põhinevad sepsisega seotud cfDNA ja pelletiproovidel. Diagrammid on valitud Paviani funktsioonide testimiseks ja võrdlemiseks. Lisaks analüüsiti, kuidas mõjutab inimese järjestuste eemaldamine diagrammi loetavust ja üldist visuaalset selgust. Kõigis kujutatud diagrammides on säilitatud kogu taksonoomiline teave.

- Diagrammid A ja B põhinevad samal pelletiproovil. Diagramm A sisaldab inimese järjestusi, mis moodustavad suure osa andmestikust. Diagrammis B on inimese järjestused eemaldatud, mille tulemusel muutuvad nähtavamaks teised taksonid. Näiteks saab paremini esile tõsta *Malassezia restricta* (*Fungi* esindaja), ja laiemate "lainete" kaudu on visuaalselt kergem võrrelda erinevate liikide esinemist.
- Diagrammid C ja D esindavad sama cfDNA-proovi. Diagramm C sisaldab suures mahus inimese järjestusi, mille tõttu teiste taksonite vood jäävad peaaegu nähtamatuks. Inimese eemaldamisel (diagramm D) muutub bakterite, näiteks *Escherichia coli* ja *Mycobacterium canetti*, osakaal selgelt nähtavaks. Nende taksonite laineid saab nüüd hõlpsasti omavahel võrrelda, mis võimaldab teha visuaalseid järeldusi nende kvantitatiivse esindatuse kohta andmestikus.

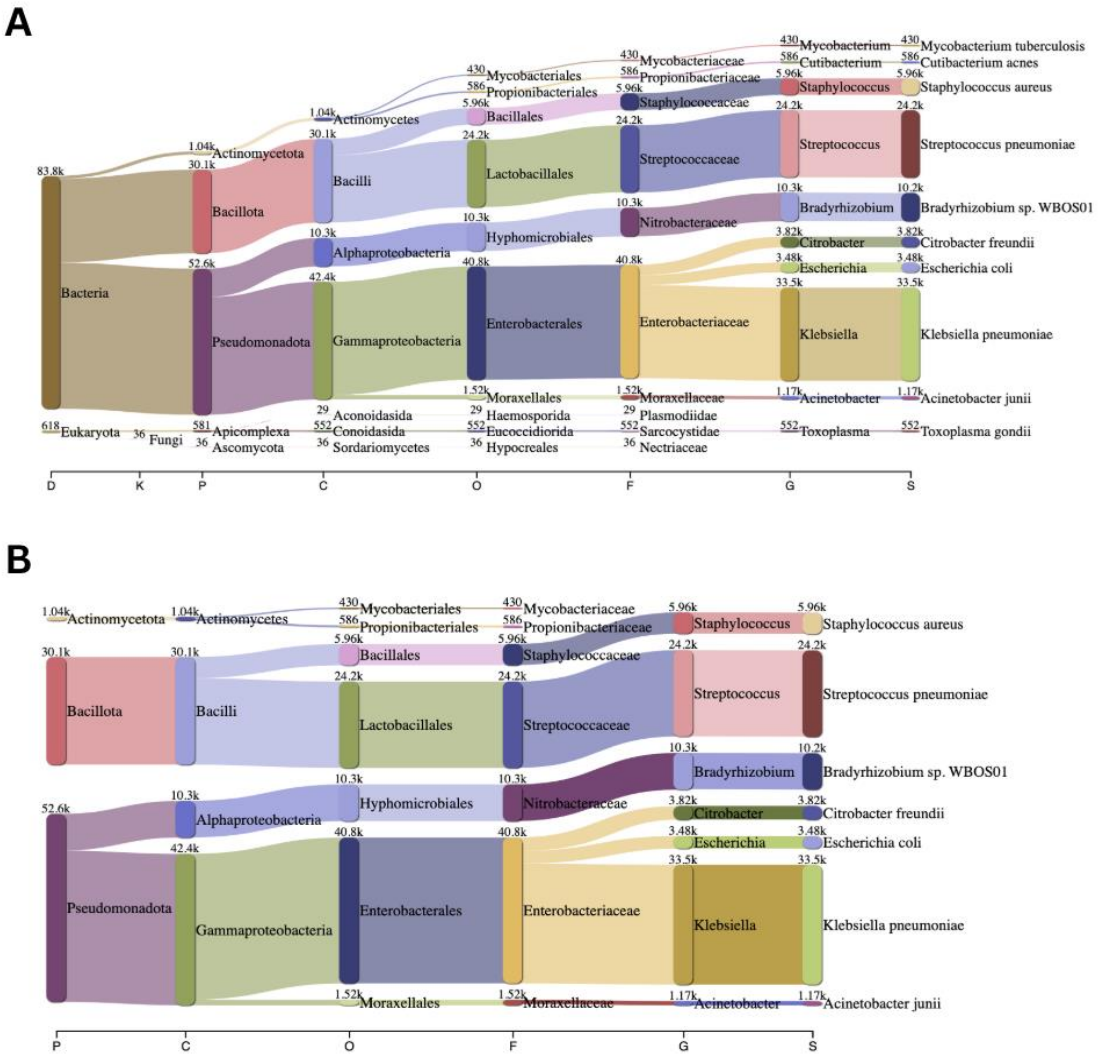


Joonis 4. Inimese järjestuste mõju Sankey diagrammidele sepsisega seotud proovides. Diagrammid A ja B põhinevad samal pelletiproovil – A-s on inimene alles, B-s eemaldatud, et

rõhutada teiste taksonite jaotust. Diagrammid C ja D pärinevad samast cfDNA-proovist. Diagrammis C varjutavad inimese järjestused muu info, D-s võimaldab nende eemaldamine paremat nähtavust bakterite ja seente osas.

Joonis 5, mis sisaldab diagramme A ja B, põhineb samal sepsisega seotud pellet-andmestikul, millest inimese järjestused on eelnevalt eemaldatud. Diagrammide erinevus seisneb Paviani tööriista kasutamises: B diagrammil on rakendatud täiendavaid filtreid, mille abil jäeti alles vaid *Eukaryota* välistavad taksonid ning bakterid, mille lugemite arv ületas 1000. Kuna kujutatakse ainult baktereid, eemaldati ka domeenitase, et visualiseering oleks selgem ja keskenduks ainult uuritavale rühmal.

Filtreerimine võimaldab keskenduda bioloogiliselt olulisematele üksustele, eemaldades müraks peetava info. Diagramm B rõhutab visuaalselt bakterirühmi, mis olid kõige arvukamalt esindatud, näiteks *Enterobacteriaceae* ja *Streptococcaceae*. Nende rühmade kõrge lugemite arv võib viidata võimalikule rollile sepsise kujunemises või lihtsalt tugevale esindatusele proovis. Selline lähenemine aitab tuvastada domineerivaid mikroobe ja võimalikke patogeene, keda võiks edasiselt uurida. Lisaks parandab fookustatud visualiseerimine üldist loetavust ja aitab vältida liigset taksonoomilist „müra“, mis võib analüüsi tulemusi varjutada.



Joonis 5. Filtreerimata ja filtreeritud Sankey diagrammid samast sepsisega seotud pelletiproovist. Diagrammid A ja B pärinevad samast proovist, kust inimese järjestused on eelnevalt eemaldatud. A-kujutis sisaldab kõiki ülejäänud taksoneid (sh bakterid, seened ja muud eukarüoodid). Diagrammil B on kasutatud Paviani filtreid – visualiseeritud on vaid bakterid, mille lugemite arv ületab 1000 ning klassifikatsiooni kõrgeim tase (domeen) on eemaldatud.

2.3.3. Mullidiagrammid

Mullidiagrammid on visuaalselt kõige lihtsamini arusaadavad, kuna need keskenduvad olulisematele andmetele ega ole nii infotihedad nagu Krona või Sankey diagrammid. Kuna kasutati Tableau Public keskkonda, oli suur osa tööst manuaalne. Tableau Public keskkonna kasutamine tähendas, et suurem osa tööst tehti käsitsi. See muutis protsessi küll ajamahukamaks, kuid andis samas ka märkimisväärse paindlikkuse – diagramme sai vabalt kohandada vastavalt uurimuse vajadustele.

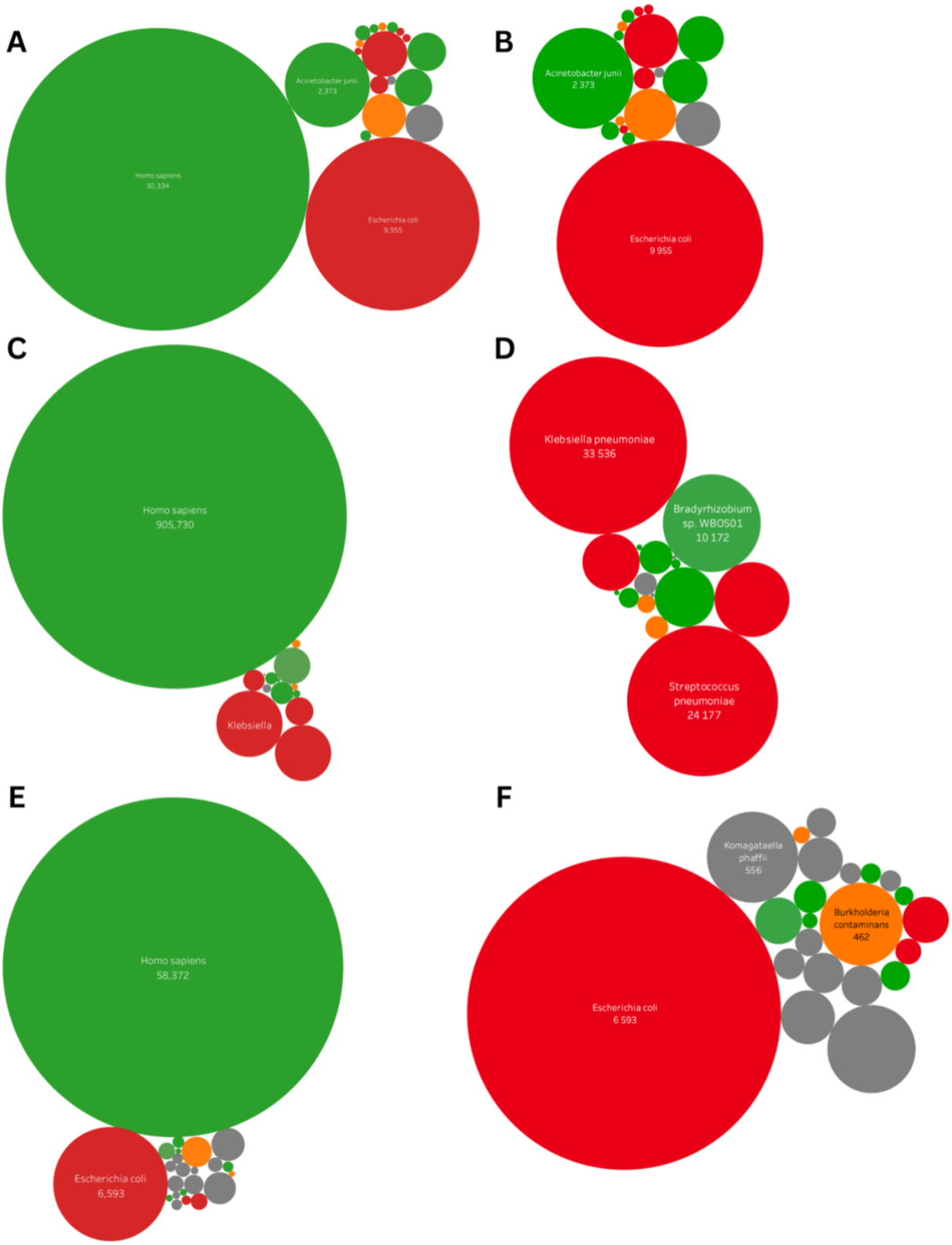
Näiteks võimaldas Tableau Public määrata mullide värvid vastavalt mikroorganismide patogeensusele sepsise kontekstis:

- Punane – patogeenne,
- Oranž – potentsiaalselt patogeenne,
- Roheline – mittepatogeenne,
- Hall – teadmata patogeensusega.

Lisaks oli võimalik hõlpsasti eemaldada või lisada andmepunkte (näiteks inimese andmed), mis võimaldas diagrammi selgust ja fookust paremini kontrollida.

Joonis 6 kujutab erinevate sepsiseproovide mullidiagramme. Mõnes diagrammis eemaldati inimene, et uurida, kuidas see mõjutab teiste taksonite nähtavust ja paigutust.

- Diagrammid A ja B pärinevad samast pelletiproovist. Kuna esialgne andmestik oli suure mahuga, võrreldi, kuidas inimese eemaldamine mõjutab teiste taksonite nähtavust. Pärast eemaldamist muutus mullide paigutus: näiteks domineerivate liikide *Escherichia coli* ja *Actinobacter junii* tõttu jäid väiksema osakaaluga taksonite nimed ja arvud esialgu varju. Selle probleemi saab lahendada interaktiivselt (näiteks suumimise või andmete filtreerimise abil), kuid liigne detail võib muuta diagrammi visuaalselt segaseks.
- Diagrammid C ja D on samuti samast pelletiproovist. C diagrammis domineerib inimese DNA (905 703 lugemit), mille tõttu kuvatakse vaid *Klebsiella* takson. Inimese eemaldamisel (D diagramm) tulid esile muud taksonid: *Klebsiella pneumoniae* (33 536 lugemit), *Bradyrhizobium sp. WBOS01* (10 172 lugemit) ja *Streptococcus pneumoniae* (24 177 lugemit).
- Diagrammid E ja F põhinevad samal cfDNA-proovil. Diagrammis E olid teised taksonid inimese kõrval nähtamatud, kuna nende lugemite arv oli väga väike. Inimese eemaldamisel (diagramm F) suurenes ülejäänud taksonite osakaal, mis võimaldas esile tuua näiteks *Komagataella phaffii* (556 lugemit) ja *Burkholderia contaminans* (462 lugemit).

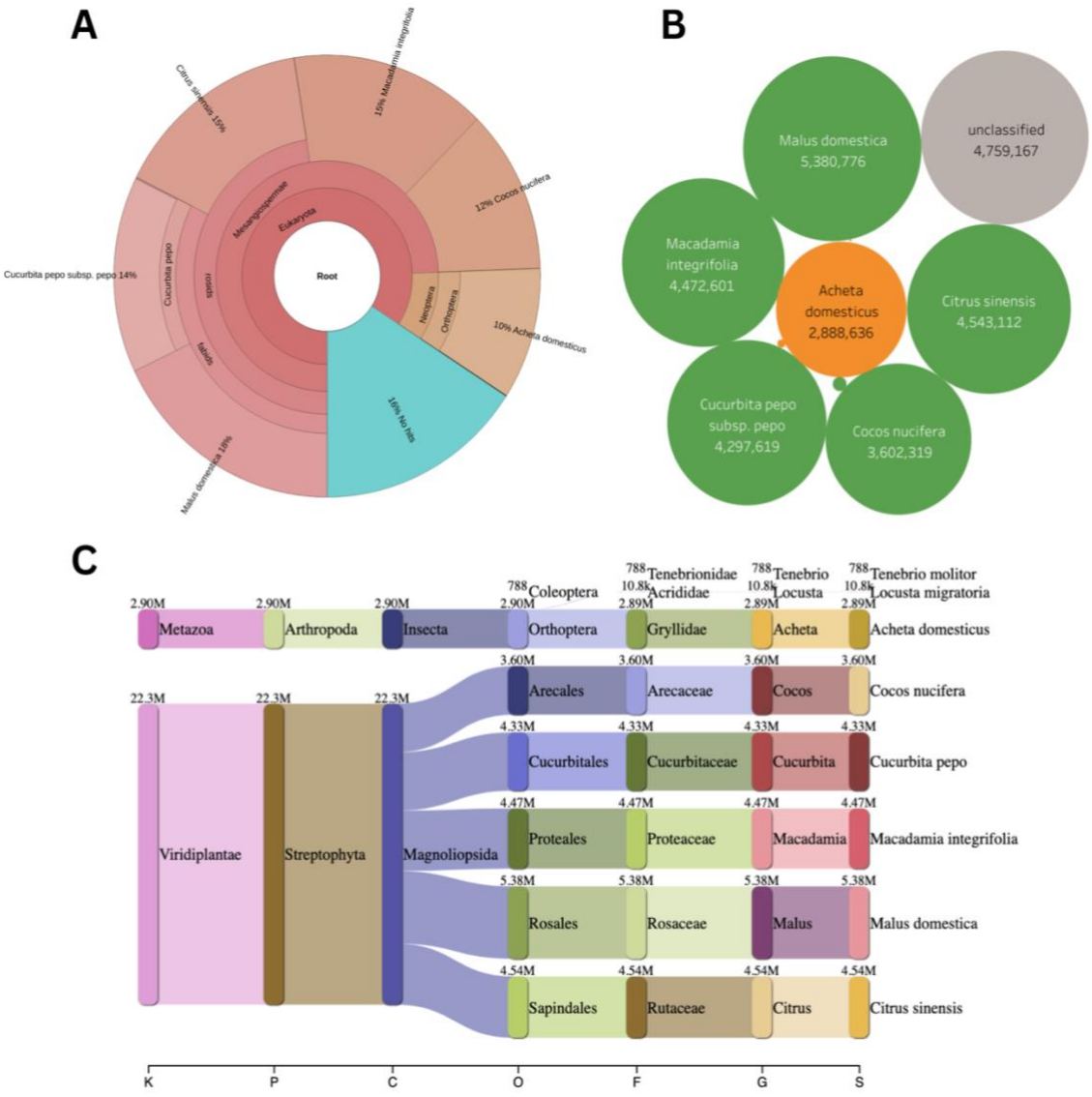


Joonis 6. Mullidiagrammid erinevatest sepsiseproovidest. Diagrammipaarid A–B, C–D ja E–F pärinevad vastavalt samadest prooviallikatest. A, C ja E sisaldavad inimese järjestusi, mis visuaalselt domineerivad. B, D ja F on vastavad diagrammid pärast inimese eemaldamist, mille tulemusel muutuvad paremini nähtavaks teised taksonid (näiteks *Escherichia coli*, *Klebsiella pneumoniae*, *Komagataella phaffii*).

2.3.4. Toiduproov

Joonis 7 sisaldab kolme diagrammi, mis on koostatud ühest simuleeritud toiduproovist, mis sisaldab lisaks toakilgi järjestustele ka erinevate taimede lugemeid. Kõik diagrammid põhinevad samal andmestikul ning andmeid ei ole filtreeritud – kõik liigimäärangud, sealhulgas madala esinemissagedusega liigid, on alles jäetud.

- Diagramm A on Krona diagramm, milles on kasutatud „collapse“ funktsiooni. Märkimisväärne on 16% osakaaluga „no hits“ sektor, mis vastab 4 759 167 järjestusele, mis jäid määramata (*unclassified*). Diagrammilt on näha, et enamik järjestustest kuulub taimedele, väiksem osa putukatele. Kuna *Tenebrio molitor* ja *Locusta migratoria* lugemite arv on andmestikus nii väike, siis nad jäävad diagrammil vaevu märgatavaks ning asuvad *Acheta domesticus*'e kõrval väga kitsas sektoris.
- Diagramm B on mullidiagramm, kus liigid on värvikoodiga eristatud: taimed rohelise ja putukad oranži värviga. „*Unclassified*“ lugemite ring on hall. Ringide suurus vastab lugemite arvule ning suuremate liikide puhul on arvud otse diagrammil nähtavad. Väiksemate ringide väärtusi saab vaadata interaktiivselt, viies kursor nende kohale.
- Diagramm C on Sankey diagramm, mis kujutab järjestusi lainetena taksonoomilises hierarhias. Kuna laine laius vastab lugemite arvule, on peamiste liikide hulga visuaalne võrdlemine lihtne. Taimede seas võib aga ilma täpsete numbriteta olla keeruline liike eristada, sest nende lainepaksused on sarnased. Putukate ja taimede eristamine on lihtne, kuna nende harud algavad erinevatest taksonoomilistest algpunktidest (*Metazoa* ja *Viridiplantae*). „*Unclassified*“ järjestusi Sankey diagramm ei kajasta.



Joonis 7. Simuleeritud toiduproovi visualiseerimine kolme erineva diagrammitüübiga: Krona, mullidiagramm ja Sankey Kõik kolm diagrammi põhinevad samal simuleeritud andmestikul. Krona diagrammil (A) on rakendatud *collapse*-funktsioon, mis lihtsustab hierarhilist vaadet. Mullidiagrammis (B) on ringid värvitud vastavalt organismitüübile: rohelised tähistavad taimi, oranžid putukaid, hallid määramata järjestusi. Sankey diagrammil (C) on kujutatud kõik taksonid; selgelt eristuvad taimede (*Viridiplantae*) ja loomade (*Metazoa*) domeenid.

2.4. ARUTELU

Kasutatud visualiseerimistööriistade ja tarkvarade paigaldamine ning kasutuselevõtt osutus erineva keerukusastmega ülesandeks. KronaTools-i allalaadimine võttis kõige kauem aega, sest installimine toimus Linuxis. Paigaldusprotsess eeldas ka NCBI andmebaasi lähtefailide eraldi allalaadimist, et diagrammid oleksid korrektsed, st sisaldaksid õigeid andmeid ja oleksid visualiseeritavad.

Paviani veebirakenduse kasutuselevõtuks tuli esmalt paigaldada Docker Desktop. Rakenduse allalaadimine toimus Docker'i kaudu, kasutades konteinerit florianbw/pavian. Käivitamisel määrati konteinerile pordiks väärtus 0, mis võimaldas süsteemil automaatselt määrata vaba pordi. Seejärel oli võimalik Pavian käivitada veebilehitsejas otse lokaalsest konteinerist, ilma täiendava seadistamiseta. Tööriist oli kiiresti töökorras ja kasutatav.

Tableau Public oli kolmest tööriistast kõige paremini ligipääsetav. See ei vajanud keerulist paigaldusprotsessi, vaid oli kättesaadav veebikeskkonnana. Kasutaja pidi looma konto ja seejärel oli võimalik koheselt alustada andmete laadimist ja diagrammide loomist. Rakendus oli koheselt kasutusvalmis ilma sõltuvusteta välistest komponentidest või andmebaasidest.

Loodud diagrammid erinesid märgatavalt nii sisulise sügavuse, interaktiivsuse kui ka visuaalse kujunduse osas. Krona diagrammid võimaldasid taksonoomilist andmestikku visualiseerida hierarhilises vormis. Kasutaja sai liikuda järjest sügavamale taksonoomia tasanditele, kus diagramm kohandus automaatselt vastavalt valitud taksonile. Igas vaates esitati taksonite osakaalud kogu andmestikus protsentidena, mis võimaldas kiiret kvantitatiivset hinnangut. Samas olid kujundusvõimalused piiratud – diagrammi taksoneid ei saanud värvida ning lugemite arvud olid nähtavad ainult siis, kui kasutaja taksonile klõpsas.

Paviani loodud Sankey diagrammid olid selged ja lihtsalt jälgitavad, kuna kõik taksonid esitati ühel tasandil. See tegi kogu andmestiku struktuuri mõistetavaks ka juhul, kui kasutaja polnud taksonoomilise süsteemiga põhjalikult tuttav. Diagrammide kõrval olid kättesaadavad täiendavad tabelvaated, kus taksonite andmeid sai sorteerida lugemite arvu, osakaalude või z-skooride alusel. Pavian võimaldas ka visualiseeringu kohandamist, näiteks taksonite eemaldamist, sõlmede vahede muutmist ja muid kohandusi. See andis kasutajale rohkem kontrolli diagrammi loetavuse üle. Nagu Krona puhul, puudus ka siin võimalus taksonite värvikodeerimiseks.

Tableau Publicus loodud mullidiagrammid pakkusid seevastu visuaalselt atraktiivset ja kiiresti mõistetavat kujutist, kus iga mull esindas ühte liiki ning selle suurus vastas lugemite arvule. Diagrammide tugevuseks oli paindlik kujundamine – näiteks sai mulli värvi määrata soovitud tunnuse alusel, nagu patogeensuse tase. See andis võimaluse tuua esile just need taksonid, mis olid analüüsi eesmärkide seisukohalt olulised. Samas jäi Tableau Publicu diagrammides esitatav teave sisuliselt vähem detailseks. Taksonoomiline hierarhia ei olnud kujutatud ning kõik liigitamised, näiteks taimede ja putukate eristamine, tuli teha käsitsi enne diagrammi loomist.

Kokkuvõtlikult võib öelda, et Krona sobib hästi taksonoomiliste seoste süvauurimiseks, ent selle paigaldus ja piiratud kujundusvõimalused muudavad kasutuskogemuse mõnevõrra keerukamaks. Pavian on funktsionaalne ja tasakaalustatud lahendus, mis pakub nii detailsust kui ka visualiseerimise paindlikkust, olles samas kasutusmugav. Tableau Public on eelkõige sobilik kiireks ja visuaalselt selgeks ülevaateks, kus analüüsi sügavus ei ole esmatähtis, kuid kujunduse kohandamine on oluline. Kõige intuitiivsemaks ja kasutajasõbralikumaks osutus Paviani Sankey diagramm, mis pakkus head kompromissi detailsuse, selguse ja interaktiivsuse vahel.

Kõik kasutatud tööriistad olid sobilikud toiduandmestike visualiseerimiseks, ent nende kasutusvõimalused erinesid sõltuvalt analüüsi eesmärgist. Krona ja Pavian suutsid automaatselt eristada putukate ja taimede taksonoomilisi rühmi, mis võimaldas kiiremat andmetöötlust ning vähendas vajadust käsitsi sekkumiseks. Tableau Publicu puhul tuli sellised eristused eelnevalt andmestikus käsitsi määratleda. Mullidiagrammide lihtsus tegi need sobilikuks üldise ülevaate andmiseks, kuid põhjalikum taksonoomiline eristus jäi tagasihoidlikumaks.

Toiduohutuse ja koostise analüüsi kontekstis oli kõige arusaadavam ja tõhusam visualiseerimisviis Paviani Sankey diagramm, mis koondas kogu andmestiku selgelt ühele tasandile ning võimaldas hõlpsalt jälgida erinevate organismide päritolu ja osakaalu. Tänu sellele sobis Pavian kõige paremini olukordadesse, kus oli oluline korraga jälgida suure hulga proovide või organismirühmade jaotumist ning nende suhtelist olulisust toidu koostises.

KOKKUVÕTE

Töös uuriti ja võrreldi erinevaid metagenoomika andmete taksonoomiliste profiilide visualiseerimise tööriistu ja nende abil loodud diagramme: KronaTools-i abil koostatud Krona diagrammid, Paviani Sankey diagramme ja Tableau Publicu mullidiagramme. Analüüsiti 16 sepsisega patsientide proovi (cfDNA ja pellet) ning üht simuleeritud toiduproovi. Eesmärgiks oli hinnata, kuidas erinevad programmid esitlevad andmeid, millised on nende eelised ja puudused ning kuidas andmestiku muutused (näiteks inimese järjestuste eemaldamine) mõjutavad diagrammide loetavust ja visuaalset arusaadavust.

Krona diagrammid osutusid interaktiivseteks ja infoküllasteks, kuid neil puudus võimalus värvikohanduseks. Sankey diagrammid olid visuaalselt lihtsamad ja võimaldasid filtreerimist, ent samuti ilma värvimisvõimaluseta. Mullidiagrammid olid visuaalselt paindlikud ning võimaldasid näiteks patogeensuse alusel värvimist, kuid edastasid vähem hierarhilist teavet.

Töö tulemusena leiti, et ükski tööriist ei ole kõiges ülim – nende valik sõltub kasutusotstarbest: kas vajadus on süvaanalüüsi, kiire võrdluse või visuaalse lihtsuse järele. Töö järeldusena saab öelda, et erinevad tööriistad sobivad erinevateks eesmärkideks: Krona sobib detailseks interaktiivseks analüüsiks, Pavian struktuurseks võrdlemiseks ning Tableau Public visuaalseks lihtsustamiseks ja kujunduslikuks kohandamiseks. Kõigi tööriistade puhul paranes diagrammide loetavus märgatavalt, kui kliinilistes andmestikes eemaldati inimese järjestused – see võimaldas paremini esile tõsta ülejäänud mikroobikoosluse, sealhulgas potentsiaalsed patogeenid.

Toiduproovide puhul sobisid kõik kolm visualiseerimisvahendit, pakkudes erineva rõhuasetusega, kuid täiendavaid võimalusi andmete esitamiseks. Kokkuvõttes kinnitas töö, et visualiseerimisvahendite valik tuleks teha vastavalt uurimisküsimusele, andmestiku eripärale ning analüüsi eesmärgile.

Visualization of taxonomic profiles

Leiki-Maria Arroval

Summary

Metagenomics has become a valuable method for examining microbial communities directly from clinical or environmental samples — without the need for cultivation. It plays a key role in understanding the composition and function of microbes in contexts such as the human body, food, and other natural environments. However, the large volume and complexity of sequencing data pose challenges for interpretation. This is where taxonomic profile visualizations become especially important, as they help to make complex microbial data more accessible, comparable, and meaningful.

This bachelor's thesis focuses on evaluating and comparing three taxonomic visualization tools — Krona diagrams via KronaTools, Sankey diagrams via Pavian, and Bubble charts via Tableau Public — in the context of clinical (sepsis-related) and simulated food metagenomic datasets. The study explores how the structure of input data and the customization options of each tool influence the clarity, usability, and interpretability of the resulting diagrams.

Each tool was tested on 16 sepsis samples (cfDNA and pellet) and one synthetic food dataset. Krona provided interactive, radial diagrams that allowed users to drill into taxonomic hierarchies and view percentage compositions. Sankey diagrams via Pavian illustrated the flow of read abundances through taxonomic levels in a linear format and allowed filtering, but lacked color customization. Bubble charts created in Tableau Public offered visual simplicity and manual customization, particularly through color-coding based on pathogenicity, though they were limited in hierarchical detail.

The results revealed that Krona excels in interactive exploration**,** especially when dealing with large, hierarchical datasets. Sankey diagrams are useful for overview comparisons and quick assessments of major taxa. Bubble charts are most suitable when visual emphasis is needed, such as highlighting pathogenic organisms or specific taxa. Importantly, the removal of human DNA from sepsis datasets significantly improved the clarity of bacterial and fungal signals across all visualization types.

In conclusion, this study highlights that the choice of visualization tool should depend on the research context and the user's priorities — whether those are interactivity, simplicity, or customization. For complex clinical samples, Krona and Pavian offer the most insight, whereas

Tableau Public is better suited for simplified overviews and categorical emphasis. The findings underline the value of flexible, context-aware visualization strategies in metagenomic data interpretation.

KASUTATUD KIRJANDUSE LOETELU

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 25(17), 3389-3402. doi:10.1093/nar/25.17.3389

Aplakidou, E., Vergoulidis, N., Chasapi, M., … Baltoumas, F. A. (2024). Visualizing metagenomic and metatranscriptomic data: A comprehensive review. *Computational and Structural Biotechnology Journal*. 3(23), 2011-2033. doi:10.1016/j.csbj.2024.04.060

Belstrøm, D., Constancias, F., Liu, Y., Yang, L., Drautz-Moses, D. I., Schuster, S. C., Kohli, G. S., Jakobsen, T. H., Holmstrup, P. and Givskov, M. (2017). Metagenomic and metatranscriptomic analysis of saliva reveals disease-associated microbiota in patients with periodontits and dental caries. *Npj Biofilms and Microbiomes*. 3(1). doi:10.1038/s41522-017-0031-4

Billington, C., Kingsbury, J. M. and Rivas, L. (2022). Metagenomics Approaches for Improving Food Safety: A Review. *Journal of Food Protection*. 85(3), 448-464. doi:10.4315/jfp-21-301

Bostock, M., Ogievetsky, V. and Heer, J. (2011). D^3 Data-Driven Documents. *IEEE Trransactions on Visualization and Computer Graphics*. 17(12), 2301-2309. doi:10.1109/TVCG.2011.185

Bouchot J., Trimble W. L., Ditzler G., Lan Y., Essinger S. and Rosen G. (2014). Advances in Machine Learning for Processing and Comparison of Metagenomic Data. A. Kriete and R. Elis, *Computational Systems Biology: From Molecular Mechanisms to Disease* (lk 295-329). London: Academic Press.

Breitwieser, F. P. and Salzberg, S. L. (2019). Pavian: interactive analysis of metagenomics data of microbiome studies and pathogen identification. *Bioinformatics*. 36(4), 1303-1304. doi:10.1093/bioinformatics/btz715

Brumfield, K. D., Hasan, N. A., Leddy, M. B., Cotruvo, J. A., Rashed, S. M., Colwell, R. R. and Huq, A. (2020). A comparative analysis of drinking water employing metagenmics. *PloS ONE*. 15(4). doi:10.1371/journal.pone.0231210

Carlino, N., Blanco-Míguez, A., Punčochář, M., … Pasolli, E. (2024). Unexplored microbial diversity from 2,500 food metagenomes and links with the human microbiome. *Cell*. 187(29), 5775-5795. doi:10.1016/j.cell.2024.07.039

Moeckel, C., Mareeboina, M., Konnaris, M. A., Chan, C. S. Y., Mouratidis, I., Montgomery, A., Chantzi, N., Pavlopoulos, G. A. and Georgakopoulos-Soares, I. (2024). A survey of k-mer methods and applications in bioinformatics. *Computational and Structural Biotechnology Journal*. 3, 2289-2303. doi:10.1016/j.csbj.2024.05.025

Ondov, B. D., Bergman, N. H. and Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*. 12(385). doi:10.1186/1471-2105-12-385

Pérez-Cobas, A. E., Gomez-Valero, L. and Buchrieser, C. (2020). Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. *Microbial Genomics*. 6(8). doi:10.1099/mgen.0.000409

Qi, Y., Lin, W.-Q., Liao, B., Chen, J.-W. and Chen, Z. (2023). Blood plasma metagenomic next-generation sequencing for identifying pathogens of febrile neutropenia in acute leukemia patients. *Scientific Reports*. 13(1). doi:10.1038/s41598-023-47685-6

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*. 9(8), 811-814. doi:10.1038/nmeth.2066

Shah, N., Nute, M. G., Warnow, T. and Pop, M. (2018). Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. *Bioinformatics*. 35(9), 1613-1614. doi:10.1093/bioinformatics/bty833

Thomas, T., Gilbert, J. and Meyer, F. (2012). Metagenomics – a guide from sampling to data analysis. *Microbial Informatics and Experimentation*. 2(1). doi:10.1186/2042-5783-2-3

Urrea, V., Páez-Triana, L., Velásquez-Ortiz, N., … Muñoz, M. (2025). Metagenomic Analysis of Surface Waters and Wasterwater in the Colombian Andean Highlands: Implications for Health and Disease. *Current Microbiology*. 82(4). doi:10.1007/s00284-024-04019-7

Wang, W.-L., Xu, S.-Y., Ren, Z.-G., Tao, L., Jiang, J.-W. and Zheng, S.-S. (2015). Application of metagenomics in the human microbiome. *World Journal of Gastroentereology*. 21(3), 803-814. doi: 10.3748/wjg.v21.i3.803

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*. 15(3). doi:10.1186/gb-2014-15-3-r46

Xu, M., Gao, J., Li, S., Zeng, M., Jianming, W. and Luo, M. (2020). Metagenomic analysis and identification of emerging pathogens in blood form healthy donors. *Scientific Reports*. 10(1). doi:10.1038/s41598-020-72808-8

Xu, P. and Gunsolley, J. (2014). Application of metagenomics in understanding oral health and disease. *Virulence*. 5(3), 424-432. doi:10.4161/viru.28532

Yen, S. and Johnson, J. S. (2021). Metagenomics: a path to understanding gut microbiome. *Mammalian Genome*. 32(4), 282-296. doi:10.1007/s00335-021-09889-x

KASUTATUD VEEBIAADRESSID

Docker. (n.d.). *Docker*. Vaadatud 17.03.2025 https://www.docker.com/

Marbl. (n.d.). *Krona*. Vaadatud 29.11.2024 https://github.com/marbl/Krona/wiki/Installing

Segre J. (2025). *Metagenomics*. National Human Genome Research Institute. Vaadatud 15.03.2025 https://www.genome.gov/genetics-glossary/Metagenomics

Tableau Public. (n.d.). *Tableau Public*. Vaadatud 18.12.2024 https://www.tableau.com/community/public

LIHTLITSENTS

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks.

Mina, Leiki-Maria Arroval, annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Taksonoomiliste profiilide visualiseerimine“, mille juhendaja on Maido Remm.

1. 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.

Leiki-Maria Arroval
26.05.2025