TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOINFORMAATIKA ÕPPETOOL

Marten Rikberg

Putukate taksonoomiline klassifitseerimine erinevate andmebaaside abil

Bakalaureusetöö

12 EAP

Juhendaja Maido Remm

TARTU 2025

Putukate taksonoomiline klassifitseerimine erinevate andmebaaside abil

Bakalaureusetöö

Marten Rikberg

Putukate kasutamine toiduainetööstuses on kasvamas, mistõttu on oluline arendada täpseid viise nende DNA tuvastamiseks toidust. Töö keskendus toidu metagenoomsele analüüsile – meetodile, mis võimaldab kindlaks teha toidus esinevad liigid DNA põhjal. Eesmärk oli simuleerida erineva *Acheta domesticuse* lugemite sisaldusega toidu andmeid ning võrrelda KRAKEN2 abiga kolme erinevat andmebaasi putukate tuvastamiseks: nt_2024 nukleotiidide andmebaasi, mt_2025 mitokondrite andmebaasi ja full_genome_2025 täisgenoomide andmebaasi.

Märksõnad: KRAKEN2, *Acheta domesticus,* metagenoomne sekveneerimine, söödavad putukad

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

Insects' taxonomic classification using different databases

Bachelor's thesis

Marten Rikberg

The use of insects in food industry is increasing and therefore it is important to develop accurate methods for detecting their DNA in food. The study focused on food metagenomic analysis – a method that enables the identification of species present in food based on their DNA. The goal was to simulate food samples with varying Acheta domesticus content and compare the performance of three different databases in insect identification using KRAKEN2: nt_2024nucleotide database, mt_2025 mitochondrial database, and full_genome_2025 database.

Keywords: KRAKEN2, *Acheta domesticus,* metagenomic sequencing, edible insects

CERCS code: B110 (Bioinformatics, health informatics, biomathematics, biometrics)

SISUKORD

INFOLEHT 2

SISUKORD 3

LÜHENDID 4

SISSEJUHATUS 5

1 KIRJANDUSE ÜLEVAADE 6

1.1 Putukate kasutamine toiduainetööstuses 6

1.1.1 Putukate töötlemine 6

1.1.2 Eestis toiduna kasutatavad putukaliigid 7

1.2 Toidu metagenoomne sekveneerimine 8

1.2.1 Metagenoomse sekveneerimise eesmärgid 8

1.2.2 Metagenoomse klassifitseerimise tarkvara 10

1.3 Putukate genoomid 11

2 PRAKTILINE OSA 13

2.1 Metoodika 14

2.1.1 Toidu metagenoomsete andmete simuleerimine 14

2.1.2 Andmebaaside ehitamine 15

2.1.3 Tarkvarade kasutamine 15

2.2 Tulemused 16

2.2.1 Simuleeritud andmed 16

2.2.2 Andmebaaside valik 17

2.2.3 nt_2024 tulemused 17

2.2.4 mt_2025 tulemused 19

2.2.5 full_genome_2025 tulemused 21

2.2.6 Andmebaaside võrdlus 23

2.2.7 Arutelu 25

KOKKUVÕTE 27

SUMMARY 28

KIRJANDUSE LOETELU 29

LIHTLITSENTS 33

LÜHENDID

aw - vee aktiivsus; *water activity*

CDC – *Centers for Disease Control and Prevention*

CFU – kolooniaid moodustav ühik; *colony forming unit*

GWP – globaalne soojenemispotentsiaal; *global warming potential*

LCA – lähim ühine esivanem; *lowest common ancestor*

PSI-BLAST - *Position-Specific Iterative Basic Local Alignment Search Tool*

RTL – juurest sõlmeni; *root to leaf*

SGB – liigitasemeline klaster; *species-level genome bins*

SSO – riknemist põhjustav organism; *specific spoilage organism*

TVC – eluvõimeliste mikroorganismide arv; *total viable count*

SISSEJUHATUS

Tänapäeva toidutööstus seisab silmitsi erinevate väljakutsetega nagu inimpopulatsiooni kasvust tulenev toidu tootmise jätkusuutlikus, loodusvarade piiratus ja tootmisega seonduvad keskkonnamõjud. Seetõttu tuleb mõelda alternatiivsetele toiduallikatele, mis oleksid jätkusuutlikud ning toitaineterikkad. Üheks selliseks alternatiiviks on uuendtoitudena kirjeldatud söödavad putukad nagu *Acheta domesticus, Tenebrio molitor* ja *Locusta migratoria.* Putukad on valgurikkad ning sisaldavad palju erinevaid mineraalaineid ja rasvasid.

Putukate kasutamine toiduainetööstuses on kiiresti kasvamas, mistõttu on oluline arendada välja efektiivsed meetodid, kuidas nende olemasolu toidus DNA põhjal kindlaks teha. Üks selline võimalus on metagenoomne sekveneerimine, mis võimaldab öelda toidu sisalduse liigi täpsusega ilma eelnevalt teadmata, mis liigid toidus esinevad. Metagenoomse sekveneerimise täpsus sõltub suuresti referentsandmebaasi olemasolust ja täpsusest. Kuna suurem osa putukate referentsgenoome on alles viimastel aastatel sekveneeritud ja erinevatesse andmebaasidesse lisatud, on oluline testida nende andmebaaside efektiivsust putukate tuvastamiseks toidust.

Käesoleva töö eesmärk on simuleerida erinevate *Acheta domesticuse* (toakilk) kontsentratsioonidega toidu andmeid ning võrrelda kilgi tuvastamise efektiivsust erinevate referentsandmebaaside abil.

1 KIRJANDUSE ÜLEVAADE

1.1 Putukate kasutamine toiduainetööstuses

Maailma erinevad kultuurid on putuktoite alternatiivse toiduallikana kasutanud juba pikemat aega. Inimkonna arvukuse kasvades ning orgaaniliste toiduressurside nappuse tõttu tuleb mõelda alternatiivsetele energiaallikatele, mis aitab vähendada ka toidu töötlemisega kaasnevat keskkonnasaastet. Üks sellisest alternatiividest on putukate kasutamine toiduallikana. Putukate valguline koostis on umbes 35-61% ning rasvad moodustavad 13-33% kehamassist (Ojha *et al.*, 2021). Nad on rikkad mineraalide poolest nagu magneesium, kaalium, kaltsium jt. Suure rasvasisalduse tõttu on ka putukate toiteväärtus suur. Yee Ling Tani uurimuse järgi, kaheksa valitud söödava putuka keskmine toiteväärtus oli umbes 470 kcal/100 g kohta (Tan et al., 2025)

Väiksem keskkonnamõju ja ökonoomsem ressursside kasutus põhjendab samuti putukate kasutamist alternatiivse toiduallikana. Näiteks putukate kasvatamise globaalse soojenemise potentsiaal (GWP) on 5,5-12,5 korda väiksem, kui see on veiste kasvatamisel; maakasutus 7,9-14,1 korda väiksem ja energiakasutus 1,02-1,58 korda väiksem (Oonincx & De Boer, 2012).

Selliste alternatiivsete toiduallikate kasutusele võtmisel tuleb tähelepanu pöörata toiduohutusele, sest ka putuktoitudes võib esineda keemilisi ühendeid nagu raskemetallid või mikrobioloogilised aspektid nagu bakterid (Skotnicka et al., 2021).

1.1.1 Putukate töötlemine

Putukate töötlemine hõlmab erinevaid tehnikaid. Peamiselt kasutatakse külmutamist, kuumutamist, kuum- ja külmkuivatamist, purustamist ja fraktsioneerimist. Töötlemise keskne eesmärk on mikroobide hävitamine (IPIFF, 2022).

Kuumutamine hõlmab putukate kastmist kuuma vette ja kuuma õhuga töötlemist, et hävitada patogeensed bakterid. Sõltuvalt putukaliigist tuleb kasutada erineva temperatuuriga vett ja hoida neid vees lühemat või pikemat aega (IPIFF, 2022). Mõõdeti näiteks Eestis lubatud söödava putuka *Tenebrio molitori* kogu eluvõimeliste mikroorganismide arvu (TVC), kus peale 10. minutilist keetmist või röstimist, oli TVC (CFU/g) vähenenud ligi 4,5 korda (Klunder et al., 2012).

Täpselt nagu kuumutamisel, erineb ka külmutusprotsessi temperatuur putukaliigist (IPIFF, 2022). Putukaid transporditakse tavaliselt -20°C juures ning külmutamist kasutatakse peamiselt selleks, et hoida putukate toiteväärtust enne, kui nendega tegutsetakse edasi.

Külmkuivatamist kasutatakse peale putukate tapmist, et eemaldada putukatest vesi ja valmistada nad ette purustamiseks. Madal vee aktiivsus takistab mikroorganismide paljunemist, sest enamus mikroorganisme vajavad vähemalt 0,6 aw, et nende metabolism toimiks ning vähemalt 0,4 aw ensümaatiliste reaktsioonide jaoks. (IPIFF, 2022; Peleg, 2022; Krzyżaniak et al., 2022). Poola teadlane Michał Krzyżaniaki uuris, kuidas erineva ajalise pikkusega kuivatamine mõjutab jahumardika vee aktiivsust ning täheldas ligi 14x vee aktiivsuse vähenemist, kui 16h asemel külmkuivatati jahumardikaid 24h (Krzyżaniak et al., 2022)

Kuumkuivatamist kasutatakse kuumusel põhinevat vee eemaldamist putukatest. Näiteks kuumas ahjus, mis on eelnevalt korralikult puhastatud. Siin on oluline stabiilne temperatuur, et piirata mikroobide paljunemist (IPIFF, 2022). Enamasti kuivatatakse putukaid 60-80°C juures 12-16h, kuid suurt vahet eelmainitud Poola teadlase uurimuses ei täheldatud, et pikaajalisem kuivatamine oleks efektiivsem (Krzyżaniak et al., 2022).

Purustamise ja fraktsioneerimise käigus muudetakse putukad kõigepealt homogeenseks pulbriks, mis seejärel fraktsioneeritakse. Vastavalt erinevatele fraktsioneerimise tehnikatele saadakse putukatest kätte erinevaid bioühendeid nagu valgud, rasvad, kitiin jne (IPIFF, 2022).

1.1.2 Eestis toiduna kasutatavad putukaliigid

Vastavalt Põllumajandus- ja Toiduameti määrusele võib Eestis kasvatada ja toiduna kasutada järgmisi putukaliike: hariliku jahumardika vastsed (*Tenebrio molitor*), toakilgid (*Acheta domesticus*), rändtirtsud (*Locusta migratoria*), meemesilase nukud (*Apis mellifera*) ning *Hermetia illucensise* vastsed (Põllumajandus- ja Toiduamet, 2024, lk 5). Mina enda uurimuses keskendun kolmele putukaliigile: hariliku jahumardika vastsed *Tenebrio molitor*, rändtirtsud *Locusta migratoria*, kuid peamiselt toakilgile *Acheta domesticus*,

Kasutan enda uurimises kolme putukaliiki ja toon välja nendest liikidest valmistatavad toiduained vastavalt Euroopa Parlamendi ja nõukogu määrusele (EL) 2017/2470 uuendtoitude loetelule: saia- ja leivatooted, teraviljabatoonid, pitsa ja pitsalaadsed road, eelsegud pagaritoodete jaoks, kastmed, kartuli- ja köögiviljapõhised road, vadakupulber, lihaanaloogid,

supid ja salatid, laastud ja krõpsud, õllelaadsed joogid, šokolaaditooted, pähklid, õliseemned ja kikerherned, piima ja piimatoodete analoogid, kuivatatud pastatoodete põhised tooted, hommikusöögihelbed, puder, nuudlid, kreekerid ja leivapulgad, vorstid (Komisjoni rakendusmäärus ((EL) 2017/2470, 2017).

1.2 Toidu metagenoomne sekveneerimine

Toidu metagenoomne sekveneerimine on toidu koostise uurimine seal esineva DNA põhjal. Toidu metagenoomse sekveneerimise metoodika jaguneb laias laastus kaheks: amplikoni põhine sekveneerimine, kus enne sekveneerimist amplifitseeritakse kindlat piirkonda PCR-ga üles ning *shotgun* sekveneerimine, kus genoom tehakse lühemateks fragmentideks ning seejärel sekveneeritakse ükshaaval.

Tuntuimad ja enim kasutatavad sekveneerimisplatvormid on PacBio, Nanopore ja Illumina, kus Illumina keskendub lühemate fragmentide skeveneerimisele (1 × 36bp - 2 × 300bp) ning PacBio ja Nanopore pikemate fragmentide sekveneerimisele, vastavalt kuni 60kb ja kuni 100kb (Jagadeesan et al., 2018). Genoomikas on pikkade lugemite sekveneerimise võime oluline, sest suurte genoomide puhul võib esineda palju keerulisi alasid nagu deletsioonid, translokatsioonid, isertsioonid ning lühikeste fragmentidega kogu genoom kokku panna on keeruline (Marx, 2023. Amarasinghe et al., 2020). Kuid toidu metagenoomikas suured fragmendid nii olulised ei ole, kuna toidus juba esinevad lühikesed fragmendid ning minu eesmärk on liik tuvastada, mitte genoomi kokku panna.

1.2.1 Metagenoomse sekveneerimise eesmärgid

Metagenoomse sekveneerimise üks eesmärkidest on toiduohutuse tagamine, mida saab käsitleda kahest vaatepunktist: toiduohutus tarbijale, et mikroobid tarbijat tervist ei ohustaks ja teiseks tootmisprotsessi kontrollimine bakterite suhtes, et seal poleks riknemisega seotud baktereid (SSO).

USA Haiguste Kontrolli ja Tõrje Keskuse statistika (CDC) järgi nakatub toidupatogeenidega üle 48. miljoni inimese, kellest 128 000 jõuab haiglasse ning kellest 3000 iga aasta sureb (CDC, 2025). Peamised mikroobid, kes on tarbija tervisele kahjulikud on *Salmonella typhimurium, Listeria monocytogenes, and Escherichia coli*, kus metagenoomne sekveneerimine võib aidata

tuvastada varakult virulentsusele ja resistentsusele vihjavate geenide järjestusi (Muñoz-Martinez et al., 2025).

Metagenoomne sekveneerimine aitab varakult tuvastada ka patogeenide olemasolu tootmisliinidel. Näiteks Ameerikas, Colorado osariigi ülikoolis uuriti, kuidas *shotgun* sekveneerimist saab kasutada SSO bakterite ja teiste toidupatogeenide tuvastamiseks veiseliha tootmisliinidelt alates hetkest, kui veised jõuavad lauta kuni poeletini jõudva tooteni välja (Yang et al., 2016). Uuriti kuue erineva patogeeni esinemist erinevatel ajahetkedel tootmisliinidel ning kõigi puhul oli näha suurt arvukuse langust, kui toit oli valmis minema poelettidele. Näiteks aitas sekveneerimine tuvastada, et ohutusmeetmete kasutamine vähendas *Listeria monocytogenese* kogust umbes 316 korda ning *E.coli* arvukus langes umbes 100 korda (Yang et al., 2016). Metagenoomne sekveneerimine ise bakterite arvukust ei vähendanud, kuid tänu sellele saame usaldusväärsemalt tootmispindu kontrollida.

Üks võimalikke metagenoomse sekveneerimise eesmärke on toidu päritolu kindlaks tegemine. Häid näiteid selle kohta ei ole, aga on uuritud, kas mees sisalduva DNA põhjal on võimalik öelda, mis taimi mee sisse sattunud on ja taimede kasvupiirkonna järgi määrata mee päritolu. Sellise meetodi põhjal võib tekkida argumentatsioon, kas keegi on ise taimed mee sisse pannud.

Metagenoomset sekveneerimist saab kasutada toidu koostise määramiseks. Iga päevaga lisandub poelettidele uute tootjate toidukaupu, kus tarbijale muutub aina olulisemaks toiduainete puhtus ja nende koostis. Toidu metagenoomse sekveneerimisega saab tuvastada toidus esineva DNA põhjal organism. Ühelt poolt on see kasulik, et tuvastada eelmainitud toidupatogeene. Teiseks on see efektiivne meetod tuvastamaks „võltstoite“, sest aina enam mõtlevad tootjad välja alternatiivseid viise, kuidas tootmisprotsessi odavamaks ja kiiremaks muuta. Kahjuks kasutatakse selleks palju lisandeid, mida tarbijale ei mainita ja mis võivad olla tarbija tervise kahjulikud (Choudhary et al., 2020). Näiteks suhkru tootmisel, kus suhkru massi suurendamiseks kasutatakse söödavat talgi pulbrit või mee massi suurendamiseks kasutatakse suhkrupeedi siirupit ehk melassi (Choudhary et al., 2020). Toidu koostise määramisel aitab metagenoomne sekveneerimine öelda ka erinevate DNA-de protsendilist koostist, millega saab kontrollida, kas tootjate info paki peal on korrektne.

Tartu Ülikoolis uuriti 2020. aastal, kuidas metagenoomse sekveneerimisega saab taimede DNA-d tuvastada toidus (Raime et al., 2020). Nimelt tehti test, kus küpsiste sisse pandi erineva kontsentratsiooniga lupiini jahu (potentsiaalne allergeen) ja tulemusena leiti, et 0,005%-se

koostise juures oli võimalik detekteerida 500 k-meeri, mis klassifitseeriti kui *Lupinus angustifolius* (Raime et al., 2020).

1.2.2 Metagenoomse klassifitseerimise tarkvara

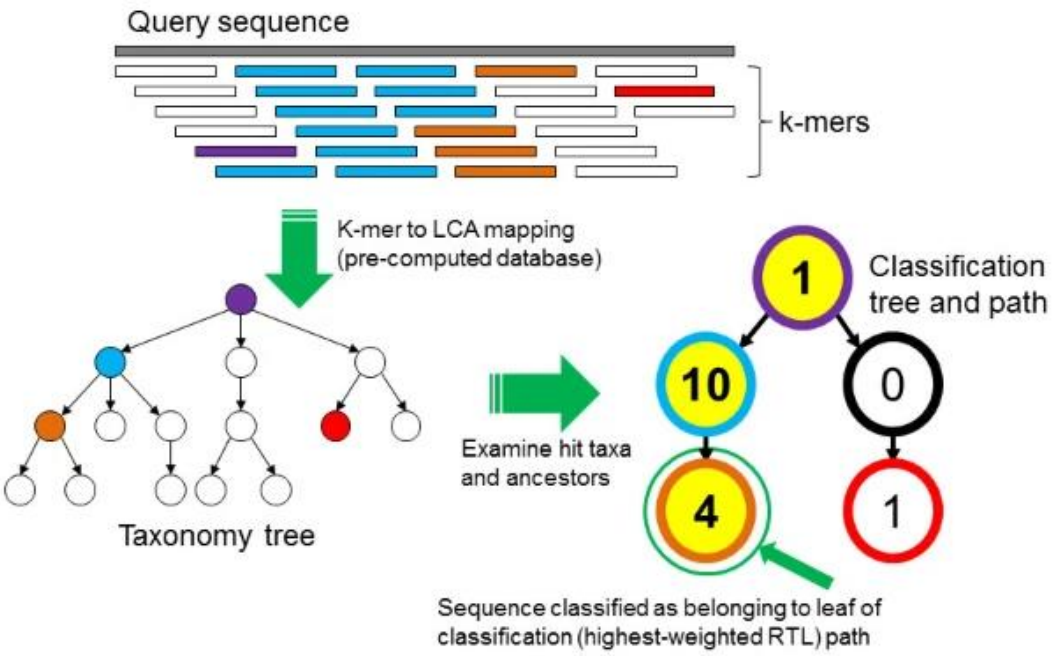
Putuka liikide määratlemiseks kasutavad teadlased erinevaid tööriistu, kuidas organismi taksonoomiline profiil koostada (Ye et al., 2019). Käesolevas töös räägin lühidalt kolmest programmist: BLAST, MetaPhlan, KRAKEN.

BLASTi *(Basic Local Alignment Search Tool)* mainiti esimesena 1990. aastal ja on üks tuntumaid ja enim kasutatavaid tööriistu, mis otsib DNA või aminohappe järjestuse sarnasusi erinevatest andmebaasidest (Altschul et al., 1990). Peab mainima ka 1997. aastal lisatud uuendused: näiteks sai kasutada sarnasuste otsimisel tühimikega joonduseid ja PSI-BLAST kasutuselevõtt, mis kasutab eelmise otsingu tulemusi, et uus skoorimaatriks luua (Altschul, 1997). BLAST-i olulisemad numbrid, mida jälgida on E-väärtus, mis näitab, et joondus ei olnud juhuslik; *bit-score*, mis näitab, kui hea joondus on, võttes arvesse ka tühimikud; identsuse protsent, mis näitab otsitava järjestuse protsendilist kattuvust andmebaasi järjestusega; *query coverage*, mis näitab, kui palju minu otsitav järjestus on kaetud andmebaasi järjestusega (Samal et al., 2021).

MetaPhlAn on BLAST-i laadne algoritm, mis otsib liigispetsiifilisi markergeene. (Blanco-Míguez et al., 2023). Kõigepealt kogutakse kokku genoomid, mis grupeeritakse liikide tasemel klastriteks (SGB). SGB-d jagunevad teadaolevates ja uuteks SGB-ks ja need on kokku pandud genoomide sarnasuse alusel ehk ühes klastris on genoomid, mille sarnasus on vähemalt 95%. Iga SGB jaoks määratakse unikaalsed markergeenid, mis iseloomustavad ainult seda rühma. Metagenoomi lugemid jooksutatakse markergeenide andmebaasi vastu ning arvutatakse keskmine markergeenide kogus SGB kohta. Lühidalt öeldes võtab MetaPhlAn toiduproovist uued ja teadaolevad liigid ning otsib neile spetsiifilisi geene, mis ainult neid rühmasid iseloomustab.

Eelpool mainitud tarkvaradest põhimõtteliselt erineva algoritmiga on KRAKEN (Wood & Salzberg, 2014; Wood et al., 2019), mida ka mina enda töös kasutan liikide tuvastamiseks. KRAKEN võtab kindla nukleotiidse pikkusega oligomeerid ehk k-meerid ning otsib iga fülogeneesipuu sõlmele spetiifilised k-meerid, mis iseloomustavad seda pere- või sugukonda puu kõige madalama sõlme tasemelt (*Lowest common ancestor* - LCA). Taksonoomilistest

eellastest ehk sõlmedest moodustub klassifitseerimispuu, ning igale sõlmele antakse vastav kaal või skoor, mitu k-meeri sealt üles leiti. Madalaima sõlmeni jõudes, kus k-meer leidis vastavuse, arvutatakse kokku kogu raja (juurest madalaima sõlmeni - RTL) kaalude summa. Rada, millel on suurim kogusumma, määrab LCA otsitavaks organismiks (vt. Joonis 1). KRAKEN suudab tuvastada poolteist miljonit lugemit minutis, mis teeb temast sadu kordi kiirema klassifitseerimise tööriista, kui eelmainitud algoritmid.



Joonis 1. KRAKENI klassifitseerimispuu algoritm. *Query sequence* ehk otsitavast järjestusest teeb KRAKEN lühemad oligonukleotiidid ehk k-meerid ning proovib sobitada neid NCBI taksonoomia andmebaasi alusel loodud puu lähima ühise eellasega (LCA). Igale sõlmele antakse skoor, kust leiti otsitavale k-meerile vastav järjestus ning lõpuks arvutatakse RTL-i koguskoor. Näiteks all paremal oleval klassifitseerimispuul on vasaku RTL-i skoor 15 ja parema RTL-i skoor 2. Kuna vasaku RTL-i skoor on suurem, siis määratud organismiks on oranži ringiga organism, sest tema on madalaim ühine eellane (Wood & Salzberg, 2014).

1.3 Putukate genoomid

Putukate tuvastamiseks toidust on vajalik teada nende DNA järjestusi. *Drosophila melanogaster* ehk harilik äädikakärbes oli esimene putukas, kelle 180 Mb suurune täisgenoom sekveneeriti 2000. aasta paiku (Adams et al., 2000). 2025 aasta aprilli seisuga on *National Centre for Biotechnology Information'i* (NCBI) kodulehel välja toodud 3730 putukate referentsgenoomi, kus kõige väiksemad genoomid on < 1 Mb ja suurimad 9 Gb suurused.

Viimase kümne aasta jooksul on lisatud sinna 3696 referentsgenoomi ehk 99% genoomidest, mis näitab hästi, et sekveneerimine on arenev valdkond. Täielikult sekveneeritud on ainult kuue putuka genoom, kellest viis kuuluvad liblikaliste seltsi. 3730-st referentsgenoomist ainult 400 on annoteeritud ehk geenidele on antud funktsionaalne info juurde ning 400-st 302 on NCBI RefSeq kvaliteeti annotatsiooni saanud. 98 on GenBankist tulnud ehk mingi teadlase või projekti raames lisatud.

Tänu sekveneeritud genoomidele saame me kasutada DNA-d putukate liikide määramiseks. Seda on varem saavutatud lihtsamate meetoditega nagu *metabarcoding,* mis põhineb punktis 1.2 mainitud amplikoni järgsel sekveneerimisel. *Metabarcoding* põhineb kolmel etapil: DNA kogumine mullast, veest või toidust; laboratoorsed etapid nagu DNA markerjärjestuse määramine, markerjärjestuse amplifikatsioon ja sekveneerimine; saadud järjestuste jooksutamine andmebaasi vastu, et määrata liik (Liu et al., 2019). Suure mutatsioonisageduse ja hea praimerite seondumise tõttu on kõige klassikalisem markergeen, mida putukatel kasutatakse tsütokroom c oksüdaasi geen (Elbrecht & Leese, 2017; Liu et al., 2019). Putukate puhul on oluline eraldada DNA valmikueas, sest hiljem on DNA-d putuka kitiinse kesta tõttu keerulisem eraldada kesta paksuse tõttu (Oppert et al., 2019).

2 PRAKTILINE OSA

Käesoleva bakalaureusetöö eesmärk oli katsetada erinevaid andmebaase putukate tuvastamiseks toidu metagenoomist. Selle eesmärgi täitmiseks oli vaja:

1) simuleerida toidu metagenoomseid lugemeid, kus uuritava putukaliigi osakaal varieerub;
2) luua KRAKENI formaadis putukate mitokondri andmebaas ja putukate täisgenoomide andmebaas;
3) jooksutada simuleeritud andmeid erinevate andmebaaside vastu ja hinnata tulemusi.

2.1 Metoodika

2.1.1 Toidu metagenoomsete andmete simuleerimine

Putuka ja taimede genoomid laadisin alla FASTA formaadis NCBI andmebaasist. GenBank ID-d iga genoomi kohta tõin välja Tabelis 1. Järgnevalt kirjutasin koodi, et võtta kogu genoomist täpselt 30 000 000 saja nukleotiidi pikkust lugemit ning konverteerida see FASTQ formaati (Rikberg & Remm, 2025). KRAKEN on programmeeritud lugema FASTQ formaadis andmefaile, mistõttu lisasin programmi käsud kontrollimaks, et FASTA failist ei tuleks kaasa järjestuste sisse jäänud realõpumärke ning lugemid ei sisaldaks väikseid tähti või tühimikke. Programm oli kirjutatud nii, et ta võtaks FASTA failist suvalise positsiooni ning sellele järgnevad sada nukleotiidi. Juhul, kui sadat nukleotiidi ei olnud võimalik võtta, siis valis programm uue positsiooni nii kaua, kuni oli võimalik mainitud pikkusega oligomeer genereerida. FASTQ andmefaili genereerimisel kasutati fiktiivseid kvaliteediskoore.

Tabel 1. Simuleeritud toidu andmete jaoks kasutatud putukate ja taimede genoomide GenBank ID-d

Liik	**GenBank ID**
Acheta domesticus	GCA_031308135.1
Cucurbita pepo	GCA_002806865.2
Cocos nucifera	GCA_003604295.2
Malus domestica	GCA_042453785.1
Macadamia integrifolia	GCA_013358625.1
Citrus sinensis	GCA_022201045.1

2.1.2 Andmebaaside ehitamine

KRAKEN2 formaadis putukate mitokondri ja täisgenoomide andmebaasi ehitasin vastavalt John Hopkinsi Ülikooli kodulehel (JHU, 2025) olevatele juhistele. Kasutasin KRAKENI vaikesätteid, mis määrasid k-meeride pikkuseks k = 35 nukleotiidi. KRAKEN2 andmebaasi kokkupanek koosneb järgmistest etappidest:

1) NCBI taksonoomia allalaadimine käsuga:

```
kraken-build --download-taxonomy --db $andmebaasi_nimi
```

2) referentsgenoomide lisamine andmebaasi käsuga:

```
kraken2-build --add-to-library $fail --db $andmebaasi_nimi
```

3) andmebaasi ehitamine käsuga:

```
kraken2-build --build --db $andmebaasi_nimi
```

, kus $andmebaasi nimi tähistab koostatud andmebaasi nime ja $fail FASTA formaadis ühte referentsgenoomi.

2.1.3 Tarkvarade kasutamine

Simuleeritud toidu andmete genereerimisel kasutasin Pythoni versiooni number 3.11.1 ning toidu andmete erinevate andmebaaside vastu jooksutamisel kasutasin KRAKEN2 versiooni number 2.1.2. Toon välja käsurea etapid, mis käivitab KRAKENi otsingualgoritmid:

1) KRAKENi versiooni aktiveerimine käsuga:

```
ml load kraken2/2.1.2
```

2) Otsingualgoritmi käivitamine käsuga:

```
kraken2 --threads 4 --confidence 0.25 --db $andmebaasi_nimi --
report $report_fail --output $output_fail
```

, kus `$andmebaasi_nimi` tähistab koostatud andmebaasi nime; `$report_fail` märgib klassifitseerimise aruande faili ja `$output_fail` tähistab klassifitseerimise tulemuste faili. *Threads* tähistab protsessorite arvu ja *confidence* usalduskünnist, kus lugem klassifitseeritakse, kui vähemalt 25% k-meeridest viitavad samale taksonile.

2.2 Tulemused

2.2.1 Simuleeritud andmed

Toidu metagenoomsete andmete simuleerimise eesmärk oli võimalikult täpselt jäljendada Eesti ettevõtte GreenBite kilgijahubatooni, kus ligi 10% batooni sisaldusest on kilgijahu (*Acheta domesticus*) ning ülejäänud 90% on n-ö. täiteaine ehk taimed. Minu eesmärk oli simuleerida nelja tüüpi andmeid, kus igas andmestikus on 30 miljonit 100. nukleotiidi pikkust järjestust ning putuka genoomist tulnud lugemite sisaldus varieerub (Tabel 2). Erinevad putuka lugemite osakaalud võimaldasid mul hinnata, millised erinevad andmebaasid annavad mulle sarnase lugemite koguse ja sarnase osakaalu tagasi. Samuti tahtsin näha, kui suure lahjenduse korral suudab KRAKEN putuka lugemid ikka ära tuvastada. Putukas, kelle genoomist võtsin lugemid oli *Acheta domesticus* (toakilk) ning täiteainena kasutasin kõrvitsat (*Cucurbita pepo*), kookost (*Cocos nucifera*), koduõuna (*Malus domestica*), makadaamiapähkel (*Macadamia integrifolia*) ja apelsini (*Citrus sinensis*). Valisin neli erinevat putukate lugemite osakaalu, et näha, millises andmebaasis saan sarnase osakaalu tagasi kõigi andmestike puhul.

Tabel 2. Simuleeritud metagenoomsed andmed. Eesmärk oli simuleerida nelja erinevat tüüpi andmeid, kus on erineva osakaaluga *Acheta domesticuse* lugemeid. Kui kilgi kontsentratsioon oli näiteks 10% ehk 3 000 000 tuleb kilgi genoomist, siis ülejäänud 27 miljonit lugemit jagas minu kood võrdselt tabelis näha olevate taimede vahel ära.

Liik	***A. domesticus* 10%**	***A. domesticus* 1%**	***A. domesticus* 0,1%**	***A. domesticus* 0,01%**
Acheta domesticus	3 000 000	300 000	30 000	3000
Cucurbita pepo	5 400 000	5 940 000	5 994 000	5 999 400
Cocos nucifera	5 400 000	5 940 000	5 994 000	5 999 400
Malus domestica	5 400 000	5 940 000	5 994 000	5 999 400
Macadamia integrifolia	5 400 000	5 940 000	5 994 000	5 999 400
Citrus sinensis	5 400 000	5 940 000	5 994 000	5 999 400

Neid nelja andmestikku kasutasin edasistes katsetes kolme andmebaasi võrdlemisel.

2.2.2 Andmebaaside valik

Kasutasin enda töös kolme erinevat referentsandmebaasi: Osaliselt NCBI Nucleotide andmebaasil põhinev nt_2024, enda koostatud mitokondri andmebaas mt_2025 ja enda koostatud täisgenoomide andmebaas full_genome_2025.

nt_2024 uuendati viimati 2024. aasta mai kuus ning koosneb paljudest erinevatest GenBank ja RefSeq järjestustest. nt_2024 andmebaasis esineb 1 070 968 erinevat liiki, millest 19 049 on putukate liigid. Valisin nt_2024, sest see oli ainuke teadaolev ja avalikult kättesaadav KRAKENi formaadis andmebaas, mis oli liikide poolest kõige täielikum ja sisaldas ka putukate genoomist pärinevaid järjestusi.

Lõin ka enda putukate mitokondri järjestustega andmebaasi, et võrrelda tulemusi nt_2024 andmebaasi tulemustega. Kui mõlemas on putukate mitokondri järjestused, peaksid tulemused tulema sarnased. *Acheta domesticus, Tenebrio molitor* ja *Locusta migratoria* – lisaks nendele kolmele valisin mardikaliste ja sihktiivaliste seltsist veel 50 putuka mitokondri järjestused. Lisaks putukatele lisasin viie taime mitokondri järjestused, kelle leiab üles tabelistest 1 ja 2. Kokku oli mt_2025 andmebaasis 55 erinevat liiki.

full_genome_2025 andmebaasi ehitamiseks lisasin referentsjärjestustena kolme putuka ja viie taime täisgenoomid (vt. Tabel 1). Putukad, kelle lisasin olid *Acheta domesticus, Tenebrio molitor, Locusta migratoria*. Kokku oli full_genome_2025 andmebaasis 8 erinevat liiki, millest 3 olid putukaliigid.

2.2.3 nt_2024 tulemused

nt_2024 andmebaasi testimise eesmärk analüüsida, a) mitu uuritava liigi algset lugemit KRAKENI abil tuvastan ning b) võrrelda tuvastatud liikide lugemite osakaalu algandmetes esineva osakaaluga

Katse läbiviimise käigus sain teada, et nt_2024 andmebaas sisaldab putukate genoomist pärinevaid mitokondri järjestusi, sest kui võtsin putukateks klassifitseeritud järjestused ja tegin nendega BLAST otsingu, siis kõik järjestused asusid mitokondrites.

Tabelis 3 tõin välja *Acheta domesticuse* lugemite arvu, mida KRAKEN seostas kilgi taksonoomilisel puul põhineva ID-ga. Huvitav tulemus oli see, et KRAKEN tuvastas *Acheta*

domesticuse lugemeid väga vähe, kuid *Gryllidae* sugukonna lugemeid rohkem (vt. Tabel 3 ja 4). Seetõttu lõdvendasin liigi määramise kriteeriumeid, sest *Acheta domesticus* kuulub taksonoomilise puu järgi *Gryllidae* sugukonna alla ning algandmestikus *Gryllidae* sugukonna liike ei olnud. Järgnevates analüüsides arvestan tulemustes *Acheta domesticuse* lugemeid *Gryllidae* sugukonna alla liigitunud lugemitena. Ilmselt oli tegemist mingi tarkvara või andmebaasi veaga.

Tabel 3. KRAKENI poolt määratud *Acheta domesticuse* lugemite arv nt_2024 andmebaasis. Erinevate *Acheta domesticuse* kontsentratsioonide korral KRAKENI määratud lugemite arv, mille KRAKEN seostas kilgi taksonoomilise ID-ga.

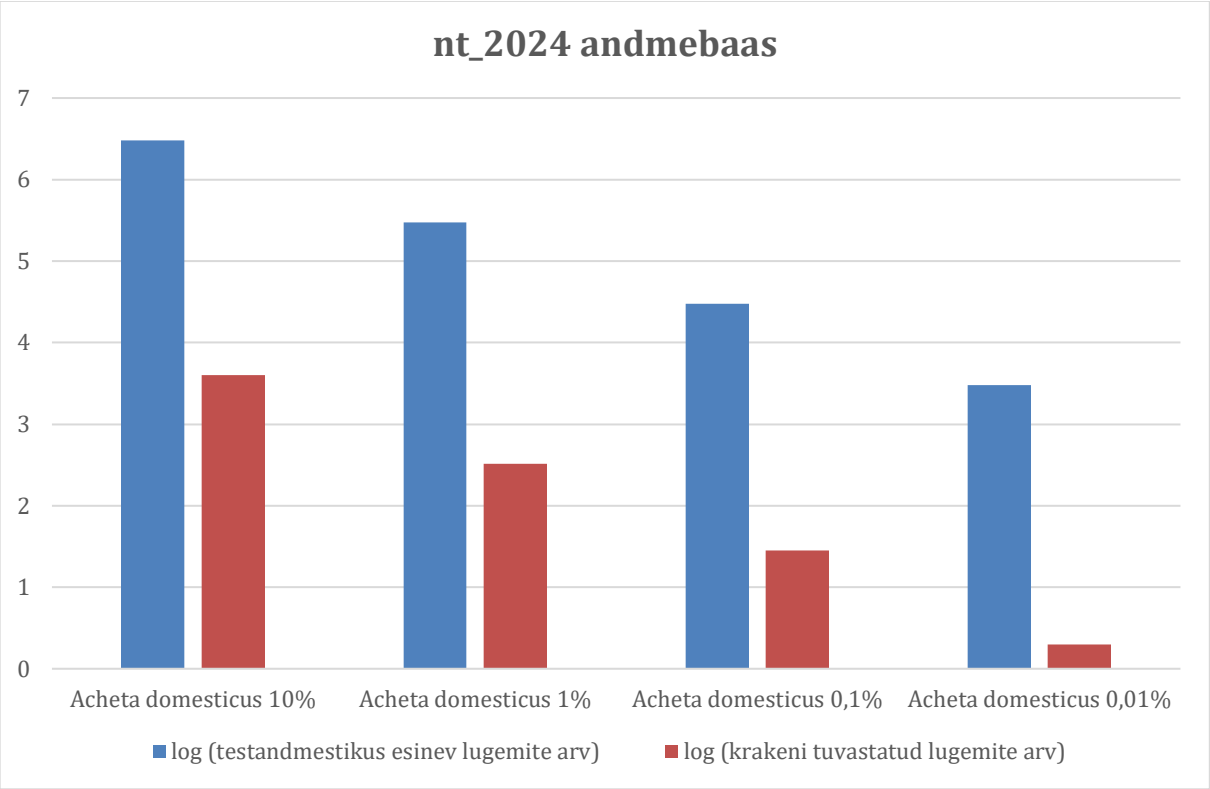
Andmestik	Andmestikus olevate lugemite arv	KRAKENI määratud lugemite arv
Acheta domesticus 10%	3 000 000	471
Acheta domesticus 1%	300 000	56
Acheta domesticus 0,1%	30 000	7
Acheta domesticus 0,01%	3000	2

Tulemusi kirjeldan seega Tabeli 4 põhjal. KRAKEN tuvastas nt_2024 andmebaasi puhul keskmiselt 0,1% lugemitest (vt. Tabel 4) ehk KRAKEN tuvastas nelja erineva andmestiku korral keskmiselt 1000x vähem lugemeid, kui testandmestikus esines. Sarnast lugemite arvu ühegi *Acheta domesticuse* andmestiku kontsentratsiooni korral tagasi ei saanud.

Tabel 4. KRAKENI poolt määratud *Gryllidae* sugukonna lugemite arv nt_2024 andmebaasis. *Acheta domesticuse* erinevate osakaalude korral KRAKENI määratud lugemite arv, mille KRAKEN seostas *Gryllidae* sugukonna taksonoomilise ID-ga.

Andmestik	Andmestikus olevate lugemite arv	KRAKENI määratud lugemite arv
Acheta domesticus 10%	3 000 000	4 027
Acheta domesticus 1%	300 000	325
Acheta domesticus 0,1%	30 000	28
Acheta domesticus 0,01%	3000	2

KRAKEN tuvastas ja klassifitseeris ka suure lahjenduse korral *Gryllidae* sugukonna lugemeid, kuid nende arv jäi mõnes andmestikus alla kümne lugemi, mis ei pruugi olla usaldusväärne määramistase. Tabeli 4 põhjal koostasin ka joonise. (vt. Joonis 2).



Joonis 2. Testandmestikus esinev lugemite arv vs KRAKENi tuvastatud lugemite arv nt_2024 andmebaasis. Joonisel on võetud lugemite arvust kümnendlogaritm, et joonis oleks arusaadavam.

2.2.4 mt_2025 tulemused

mt_2025 andmebaasi testimise eesmärk oli võrrelda, kas minu tehtud andmebaas on parem kui nt_2024 andmebaas. Selleks võrdlesin a) mitu uuritava liigi algset lugemit KRAKENI abil tuvastan ning b) tuvastatud liikide lugemite osakaalu algandmetes esineva osakaaluga.

Tabelis 5 tõin välja KRAKENI poolt määratud *Acheta domesticuseks* klassifitseeritud lugemite tulemused mt_2025 andmebaasis, kuid tulemusi kirjeldan Tabeli 6 põhjal, kus *Acheta domesticuse* lugemid on arvestatud *Gryllidae* sugukonna taksoni alla.

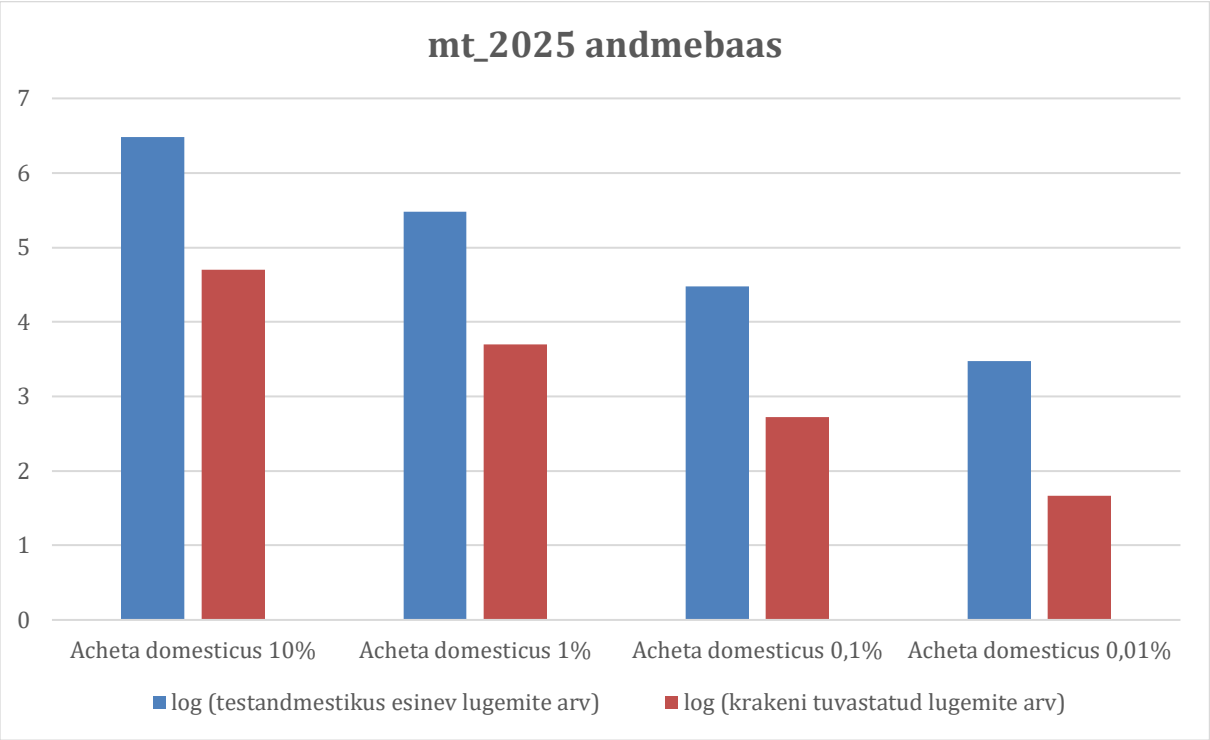
Tabel 5. KRAKENI poolt määratud *Acheta domesticuse* lugemite arv mt_2025 andmebaasis. Neli erinevat testandmestikku, kus *Acheta domesticuse* lugemite osakaal varieerus ning KRAKENI tuvastatud lugemite kogus võrreldes testandmestikes olevate lugemite arvuga.

Andmestik	Andmestikus olevate lugemite arv	KRAKENI määratud lugemite arv
Acheta domesticus 10%	3 000 000	710
Acheta domesticus 1%	300 000	93
Acheta domesticus 0,1%	30 000	7
Acheta domesticus 0,01%	3000	2

KRAKEN tuvastas mt_2025 andmebaasis nelja erineva kilgi kontsentratsiooni korral keskmiselt 1,66% lugemitest, mis on ligi 17x parem tulemus, kui nt_2024 andmebaasi korral. KRAKEN tuvastas mt_2025 andmebaasis keskmiselt 60x vähem lugemeid, kui algandmetes esines. Kuigi mt_2025 andmebaasi tulemustes ei olnud ühtegi testandmestikku, mille puhul KRAKENi tuvastatud lugemid oleks jäänud nendega samasse suurusjärku, suutis KRAKEN ka väikseima kilgi lugemite osakaalu korral tuvastada arvestatava koguse lugemeid (vt. Tabel 6). Jooniselt 4 on näha, et KRAKENi tuvastatud lugemite arv kahaneb ilusti koos testandmestikus esinevate lugemite arvuga.

Tabel 6. KRAKENI poolt määratud *Gryllidae* sugukonna lugemite arv mt_2025 andmebaasis. *Acheta domesticuse* erinevate osakaalude korral KRAKENI määratud lugemite arv, mille KRAKEN seostas *Gryllidae* sugukonna taksonoomilise ID-ga.

Andmestik	Andmestikus olevate lugemite arv	KRAKENI määratud lugemite arv
Acheta domesticus 10%	3 000 000	50 284
Acheta domesticus 1%	300 000	4968
Acheta domesticus 0,1%	30 000	533
Acheta domesticus 0,01%	3000	46



Joonis 4. Testandmestikus esinev lugemite arv vs KRAKENi tuvastatud lugemite arv mt_2025 andmebaasis. Lugemite arvust on võetud logaritm, et joonist paremini iseloomsutada.

2.2.5 full_genome_2025 tulemused

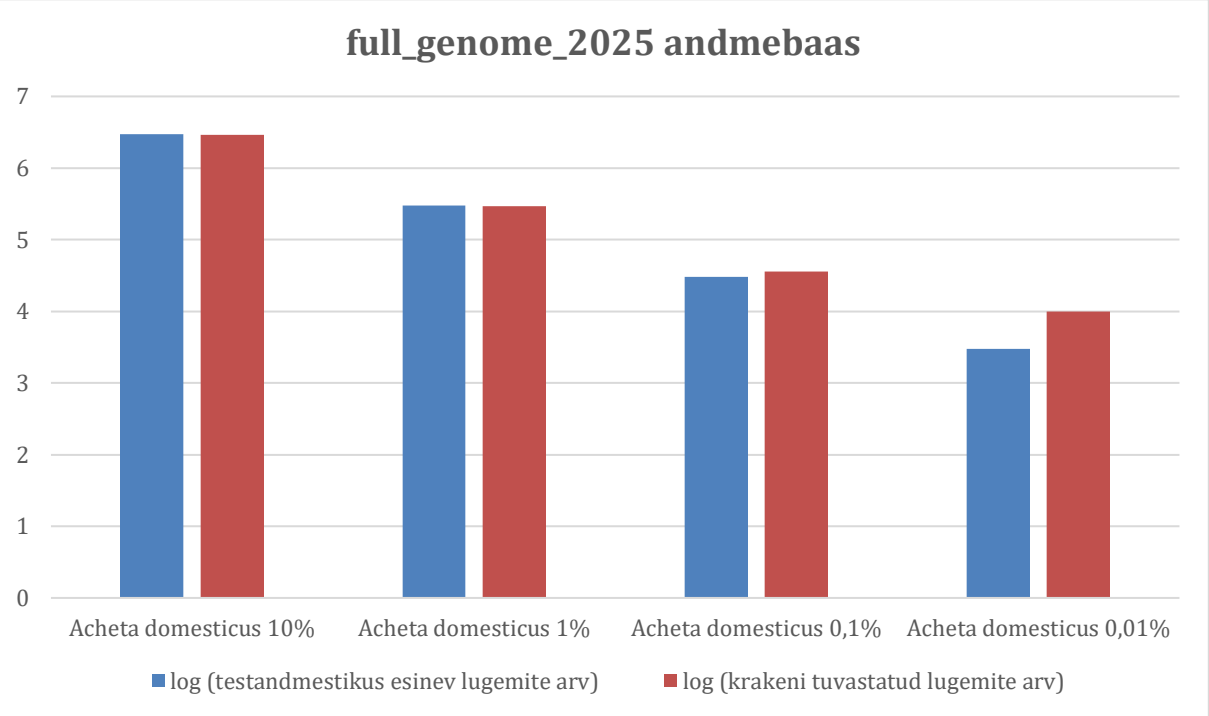
full_genome_2025 andmebaasi testimise eesmärk analüüsida, a) mitu uuritava liigi algset lugemit KRAKENI abil tuvastan ning b) võrrelda tuvastatud liikide lugemite osakaalu algandmetes esineva osakaaluga.

Eelnevates andmebaasides kujutasin *Acheta domesticuse* tulemust koos *Gryllidae* sugukonna tulemustega. Kuigi full_genome_2025 andmebaas *Gryllidae* sugukonda ei sisaldanud, loen arusaadavuse mõttes *Acheta domesticuse* lugemite tulemused nimeliselt *Gryllidae* sugukonna tulemuste alla.

Tabel 7. KRAKENI poolt määratud *Gryllidae* sugukonna lugemite arv full_genome_2025 andmebaasis. *Acheta domesticuse* erinevate osakaalude korral KRAKENI määratud lugemite arv, mille KRAKEN seostas *Gryllidae* sugukonna taksonoomilise ID-ga.

Andmestik	Andmestikus olevate lugemite arv	KRAKENI määratud lugemite arv
Acheta domesticus 10%	3 000 000	2 888 636
Acheta domesticus 1%	300 000	295 476
Acheta domesticus 0,1%	30 000	35 917
Acheta domesticus 0,01%	3000	9923

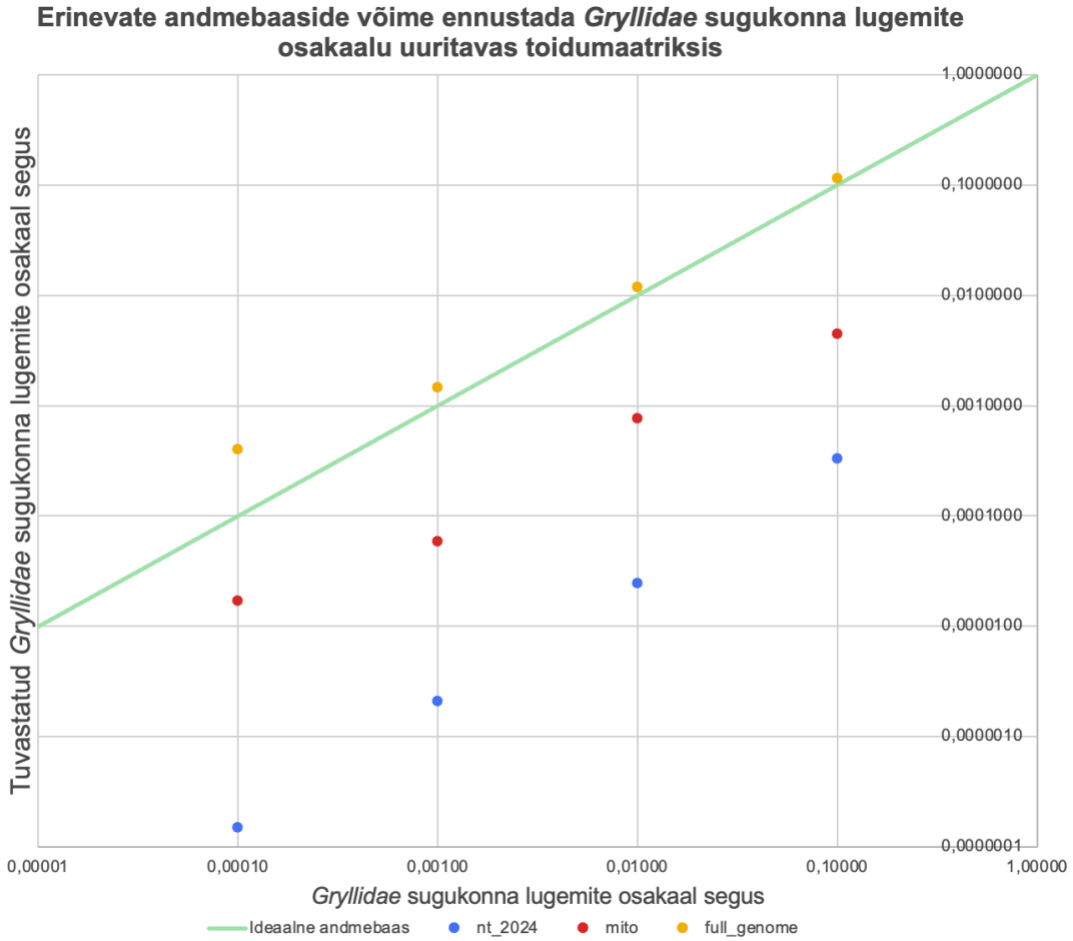
full_genome_2025 andmebaasi puhul oli nelja algandmestiku keskmine tuvastamismäär võrreldes teiste andmebaasidega 1-le kõige lähemal. Kõige väiksema kilgi lugemite osakaaluga andmestiku korral oli tuvastamise määr ümardatult 3, mis sisuliselt tähendab, et KRAKEN tuvastas 3 korda rohkem lugemeid, kui algselt andmebaasi vastu jooksutati (vt. Tabel 7). Kõige täpsemini töötas full_genome_2025 andmebaas suurte kilgi lugemite osakaalu korral ning väiksemate puhul tuvastas KRAKEN rohkem lugemeid, kui algandmetes esines.



Joonis 5. Testandmestikus esinev lugemite arv vs KRAKENi tuvastatud lugemite arv full_genome_2025 andmebaasis. Joonise paremaks iseloomustamiseks võtsin Tabelis 7 toodud lugemite arvust kümnendlogaritmid.

2.2.6 Andmebaaside võrdlus

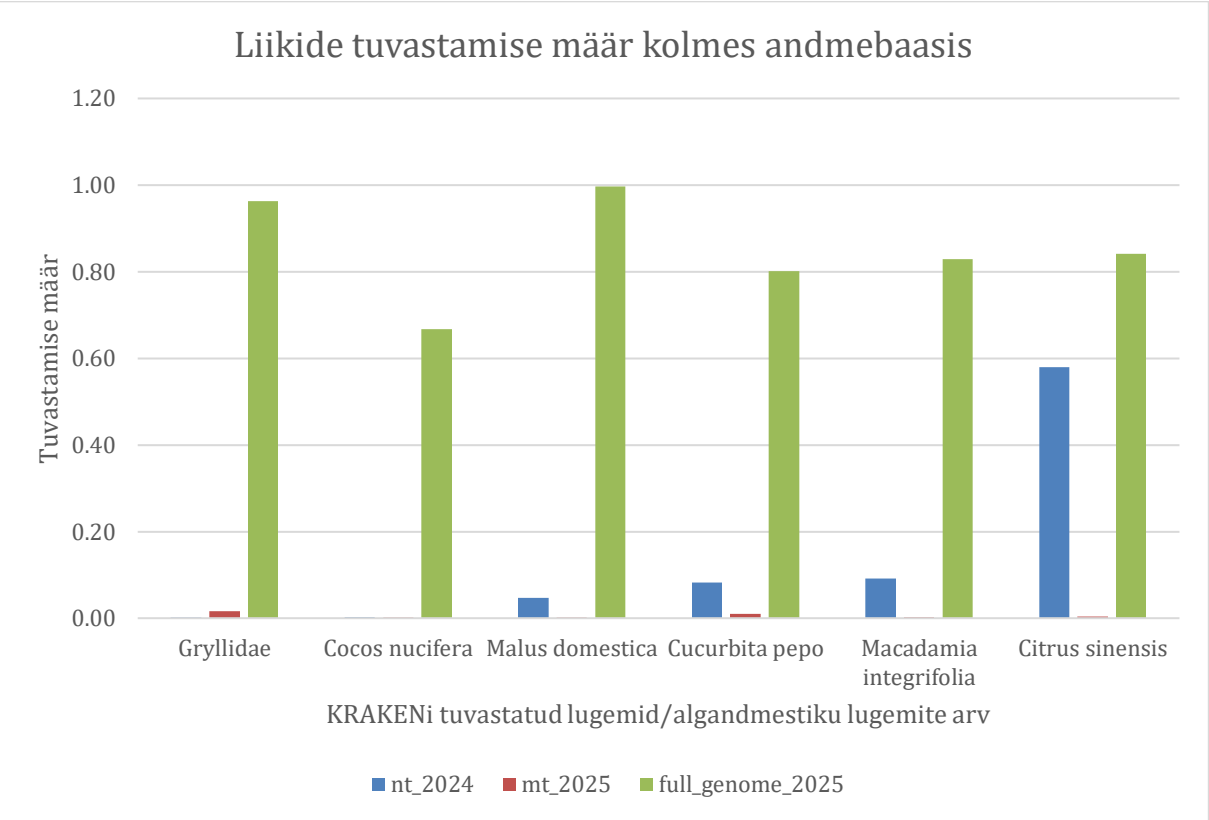
Tõin välja joonise, et illustreerida, kui erinevad võivad olla tuvastamise tulemused sõltuvalt andmebaasist ja selle sisust (vt. Joonis 6). Samuti võimaldas see hinnata, kas *Acheta domesticuse* lugemite osakaalu kasvu korral algandmestikes, toimus lineaarne kasv ka KRAKENi tuvastatud lugemite arvus. Tõin välja ka sirge, kus ideaalse andmebaasi korral tulemused peaksid asetsema. Ükski andmebaas ideaalse andmebaasi sirgel täies ulatuses ei asetsenud. Kõige lähedasem tulemus sellele oli full_genome_2025 andmebaas, kõige kaugemale jäi nt_2024 andmebaasi tulemused. Ka mt_2025 andmebaasi korral ei tuvastanud KRAKEN sarnases osakaalu suurusjärgus lugemeid.



Joonis 6. Kolme andmebaasi võrdlus *Gryllidae* sugukonna tuvastamisel. X- ja Y-telje järgi asetsevad puntkid graafikul on algandemetes esineva *Acheta domestiucse* lugemite osakaalude ja tuvastatud *Gryllidae* sugukonna lugemite osakaalude suhe.

Hindasin ka erinevate andmebaaside tuvastamismäära. Ideaalse andmebaasi korral oleks tuvastamise määr alati 1 ehk tuvastatakse sama palju lugemeid, kui palju algandmetes esines.

Andmebaas, mille tuvastamise määr kõikide liikide puhul oli kõige lähemal 1-le oli full_genome_2025 (vt. Joonis 7). Joonis 7 iseloomustab hästi, kuidas nt_2024 ja mt_2025 andmebaaside tuvastamise määr on võrreldes full_genome_2025 andmebaasiga marginaalne. Taimeliikide puhul saab samuti öelda, et enamuste liikide puhul on full_genome_2025 andmebaas lugemite arvu määramiseks parem, kui teised andmebaasid. Kõige väiksema tuvastamise määraga andmebaas on mt_2025. Huvitav tulemus on apelsini puhul nt_2024 andmebaasiga, sest BLAST otsingut tehes tuli välja, et enamuste liikide puhul oli nt_2024 andmebaasis kajastatud nende mitokondri järjestused, aga vaadates Joonist 7, siis apelsini puhul tundub, et nt_2024 andmebaasis on apelsini täisgenoom. Liikide algandmed pärinesid kõik nende täisgenoomidest (vt. Tabel 1).



Joonis 7. Andmebaaside tuvastamise määr erinevate liikide korral. X-teljel on toodud liigid, mis esinesid kõikides andmebaasides. Joonisel on toodud tuvastamise määr *Acheta domesticus* 10% andmestiku korral (vt. Tabel 2).

2.2.7 Arutelu

Minu eesmärk oli katsetada erinevaid andmebaase, et hinnata, kui suure koguse lugemeid KRAKEN tuvastab võrreldes algandmestikuga ning võrrelda andmebaaside tulemusi.

Andmebaaside nt_2024 ja mt_2025 puhul pidin *Acheta domesticuse* lugemid arvestama *Gryllidae* sugukonna lugemiteks. Tegin sellise otsuse, sest *Acheta domesticuse* lugemeid tuli mõlema andmebaasi puhul üsna vähe tagasi ja taksonoomilisel puul kuulub *Acheta domesticus Gryllidae* sugukonna alla. Edaspidi peaks *Acheta domesticuse* puhul kasutama andmebaasi, kus *Gryllidae* sugukonda referentsina ei esine või kui on soov *Gryllidae* sugukonna liike ja *Acheta domesticust* korraga tuvastada, siis peaks mõtlema välja mingi mooduse, kuidas neid eristada. Tõenäoliselt oli see mingi tarkvara või programmi viga miks *Acheta domesticuse* lugemid läksid *Gryllidae* sugukonna alla.

nt_2024 andmebaas on hea laialdane andmebaas, sest ta on kaetud väga paljude erinevate liikidega, aga pole väga usaldusväärne, sest osade liikide puhul tuli välja, et andmebaasis esineb selle liigi mitokondri järjestused, osade puhul oli tegemist täisgenoomidega ja osade puhul ilmselt mingi osaga genoomist. Selleks, et saaks toiduandmeid kvaliteetselt analüüsida oleks vaja, et kõik liigid oleks esindatud, kas ainult mitokondrite tasemel või ainult täisgenoomide tasemel. Üldiselt sobis nt_2024 andmebaas liikide tuvastamiseks, kuid osakaalu väga täpselt ei suutnud määrata. Ilmselt ei saanud nt_2024 andmebaasi tulemustes kätte väga head tuvastamise määra, sest algandmed olid võetud erinevate liikide täisgenoomidest, kuid andmebaasi referentsjärjestusteks olid mõnedel liikidel mitokondri järjestused, mõnedel täisgenoomid ja mõnedel mingi osa genoomist.

mt_2025 oli usaldusväärsem andmebaas, kui nt_20245, sest ma teadsin kindlalt millised referentsjärjestused andmebaasi lisati. mt_2025 andmebaas sobis samuti toidumaatriksis esinevate liikide kirjeldamiseks. Osakaalu määramise poolest oli ta täpsem, kui nt_2024 andmebaas ilmselt sellepärast, et mt_2025 andmebaasis oli referentsjärjestusena vähem liike ning KRAKENil on vähem võimalusi liigi valesti klassifitseerimiseks.

full_genome_2025 oli teistest andmebaasidest kõige efektiivsem. Üldiselt sobib full_genome_2025 andmebaas üsna hästi sarnase lugemite arvu tuvastamiseks algandmetest, sest tema tuvastamise määr oli 1-le kõige lähemal. Küll aga full_genome_2025 andmebaasi puhul tuvastas KRAKEN väiksemate *Acheta domesticuse* osakaalude juures rohkem lugemeid,

kui algandmetes esines, sest ilmselt läksid mõnede taimeliikide lugemid ka *Gryllidae* sugukonna lugemite alla. full_genome_2025 oli kõikidest andmebaasidest kõige täpsem liikide osakaalu kirjeldamiseks, kuid tema miinuseks on see, et ta nõuab hulgaliselt arvutusressurssi ja seetõttu on raske suuri täisgenoomi andmebaase ehitada/täiendada. Toidu koostise täpsemaks uurimiseks peaks seda andmebaasi edasi arendama ja lisama juurde teiste liikide täisgenoome.

KOKKUVÕTE

Toidu metagenoomne sekveneerimine võimaldab tuvastada toidu liigilist koosseisu DNA põhjal. Putuktoitude tarbimise kasvutrendi tõttu on vaja aina rohkem tähelepanu pöörata usaldusväärsete analüüsimeetodite arendamise poole, mis annaks võimalikult täpse ülevaate toidus esinevast DNA-st ja millisele liigile see DNA kuulub. See eeldab kvaliteetsete referentsandmete olemasolu.

Selle bakalaureusetöö eesmärk oli võrrelda erinevaid andmebaase (nt_2024, mt_2025 ja full_genome_2025) putukate identifitseerimiseks toidu metagenoomsetest andmetest. Eesmärgi täitmiseks simuleerisin toidu metagenoomsed lugemid varieeruva *Acheta domesticuse* osakaaluga (10%; 1%; 0,1%; 0,1%), lõin KRAKENi formaadis piiratud liikide arvuga mitokondri ja täisgenoomide andmebaasi ja jooksutasin toidu andmeid referetnsandmebaaside vastu.

Eesmärk sai täidetud ja tulemused tõestasid, et sõltuvalt andmebaaside sisust saame erinevaid tulemusi. Kõige täpsema osakaalu võrreldes algandmetega taastas KRAKEN full_genome_2025 andmebaasi abiga, mis sisaldas piiratud liikide arvuga taimede ja putukate täisgenoome. Samuti tuvastas täisgenoomide andmebaas kõige täpsemini lugemite arvu, mis esines algandmetes. Üldiselt saab KRAKEN2 algoritmidega edukalt metagenoomseid andmeid analüüsida, kuid tulemuste täpsus sõltub suuresti referentsandmebaaside kvaliteedist. Edaspidi tuleks arendada suurema liikide varieeruvusega täisgenoomide andmebaas, mida liikide tuvastamiseks kasutada.

Insects' taxonomic classification using different databases

Bachelor's thesis

Marten Rikberg

Edible insects are becoming more popular and getting more attention year by year. They are rich in nutrients and environment friendly alternative to animal based foods. However, the wider adoption of insect-based foods requires reliable mechanisms for detecting and identifying such organisms' DNA within complex food matrices.

The aim of this bachelor's study was to evaluate the effectiveness of metagenomic sequencing in detecting insect content in food by using KRAKEN algorithms. The study focused how accurately different databases (nt_2024, mt_2025, full_genome_2025) are able to detect and classify insects', mainly *Acheta domesticus'* DNA in food. The experimental part of the study was based on simulating food metagenomic datasets, where four different *Acheta domesticus* DNA concentrations (10%, 1%, 0,1%; 0,01%) was mixed with the genomes of five different plant species. Three reference databases were tested: nt_2024 (a general NCBI nucleotide database), mt_2025 (manually made mitochondrial database) and a full_genome_2025 (manually made full-genome databases).

The results demonstrated differences in classification performance between the databases in terms of the number of reads detected and their relative proportions compared to the original dataset. The most accurate database for achieving the described results was the full_genome_2025 database, where the read detection rate and the proportions were the closest to the original value. In contrast, the nt_2024 database performed the worst, with results differing the most from the ideal database's results.

The study concludes that metagenomic analysis is a suitable method for detecting novel food components such as insects but the accuracy of the results is dependent on the choice of reference database. The best performing database was the full-genome database where the next step would be to add more reference species.

KIRJANDUSE LOETELU

Adams, J. (2008) Complex genomes: Shotgun sequencing. Nature Education 1(1):186. https://www.nature.com/scitable/topicpage/complex-genomes-shotgun-sequencing-609/

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., . . . Venter, J. C. (2000). The Genome Sequence of Drosophila melanogaster. *Science*, *287*(5461), 2185–2195. https://doi.org/10.1126/science.287.5461.2185

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/s0022-2836(05)80360-2

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-1935-5

Belete, A. a. E. (2015). *Food Adulteration: Its challenges and impacts*. Ermias Belete | Food Science and Quality Management. https://www.iiste.org/Journals/index.php/FSQM/article/view/23873

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., Nickols, W. A., Piccinno, G.,

Piperni, E., Punčochář, M., Valles-Colomer, M., Tett, A., Giordano, F., Davies, R., Wolf, J., . . . Segata, N. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nature Biotechnology, 41(11), 1633–1644. https://doi.org/10.1038/s41587-023-01688-w

Choudhary, A., Gupta, N., Hameed, F., & Choton, S. (2020). An overview of food adulteration: Concept, sources, impact, challenges and detection. International Journal of Chemical Studies, 8(1), 2564–2573. https://doi.org/10.22271/chemi.2020.v8.i1am.8655

Elbrecht, V., & Leese, F. (2017). Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. Frontiers in Environmental Science, 5. https://doi.org/10.3389/fenvs.2017.00011

Havelaar, A. H., Brul, S., De Jong, A., De Jonge, R., Zwietering, M. H., & Ter Kuile, B. H. (2009). Future challenges to microbial food safety. *International Journal of Food Microbiology*, *139*, S79–S94. https://doi.org/10.1016/j.ijfoodmicro.2009.10.015

Jagadeesan, B., Gerner-Smidt, P., Allard, M. W., Leuillet, S., Winkler, A., Xiao, Y., Chaffron, S., Van Der Vossen, J., Tang, S., Katase, M., McClure, P., Kimura, B., Chai, L. C., Chapman, J., & Grant, K. (2018). The use of next generation sequencing for improving food safety: Translation into practice. Food Microbiology, 79, 96–115. https://doi.org/10.1016/j.fm.2018.11.005

Kashgouli, Sana. A Kraken-like Tool with K Given at Query Time and an Index for Finding Approximately Longest Common Substrings. Diss. Dalhousie University, 2023.

Klunder, H., Wolkers-Rooijackers, J., Korpela, J., & Nout, M. (2012). Microbiological aspects of processing and storage of edible insects. Food Control, 26(2), 628–631. https://doi.org/10.1016/j.foodcont.2012.02.013

Krzyżaniak, M., Aljewicz, M., Bordiean, A., & Stolarski, M. J. (2022). Yellow Mealworm Composition after Convective and Freeze Drying—Preliminary Results. Agriculture, 12(2), 149. https://doi.org/10.3390/agriculture12020149

Linard, B., Crampton-Platt, A., Gillett, C. P., Timmermans, M. J., & Vogler, A. P. (2015). Metagenome Skimming of insect specimen pools: Potential for Comparative Genomics. Genome Biology and Evolution, 7(6), 1474–1489. https://doi.org/10.1093/gbe/evv086

Liu, M., Clarke, L. J., Baker, S. C., Jordan, G. J., & Burridge, C. P. (2019). A practical guide to DNA metabarcoding for entomological ecologists. Ecological Entomology, 45(3), 373–385. https://doi.org/10.1111/een.12831

Marx, V. (2023). Method of the year: long-read sequencing. *Nature Methods*, *20*(1), 6–11. https://doi.org/10.1038/s41592-022-01730-w

Muñoz-Martinez, T. I., Rodríguez-Hernández, B., Rodríguez-Montaño, M., Alfau, J., Reyes, C., Fernandez, Y., Ramos, R. T., De Los Santos, E. F. F., & Maroto-Martín, L. O. (2025). Unlocking the hidden microbiome of food: The role of metagenomics in analyzing fresh produce, poultry, and meat. *Applied Microbiology*, *5*(1), 26. https://doi.org/10.3390/applmicrobiol5010026

Ojha, S., Bekhit, A. E., Grune, T., & Schlüter, O. K. (2021). Bioavailability of nutrients from edible insects. Current Opinion in Food Science, 41, 240–248. https://doi.org/10.1016/j.cofs.2021.08.003

Oonincx, D. G. a. B., & De Boer, I. J. M. (2012). Environmental impact of the production of mealworms as a protein source for humans – a life cycle assessment. PLoS ONE, 7(12), e51145. https://doi.org/10.1371/journal.pone.0051145

Oppert, B., Stoss, S., Monk, A., & Smith, T. (2019). Optimized extraction of insect genomic DNA for Long-Read sequencing. *Methods and Protocols*, *2*(4), 89. https://doi.org/10.3390/mps2040089

Raime, K., Krjutškov, K., & Remm, M. (2020). Method for the identification of plant DNA in food using Alignment-Free Analysis of sequencing reads: A case study on Lupin. Frontiers in Plant Science, 11. https://doi.org/10.3389/fpls.2020.00646

Rikberg, M., Remm, M (2025). Create random reads from a large genome FASTA [Source code]. GitHub. https://github.com/bioinfo-ut/random/blob/95c98179bd5cb02768d9da3efd9a3c86a030a33b/create_random_reads_from_a_large_genome_FASTA.py

Samal, K. C., Sahoo, J. P., Behera, L., & Dash, T. (2021). Understanding the BLAST (Basic Local Alignment Search Tool) Program and a Step-by-step Guide for its use in Life Science Research. Bhartiya Krishi Anusandhan Patrika, Of. https://doi.org/10.18805/bkap283

Skotnicka, M., Karwowska, K., Kłobukowski, F., Borkowska, A., & Pieszko, M. (2021). Possibilities of the development of edible Insect-Based foods in Europe. Foods, 10(4), 766. https://doi.org/10.3390/foods10040766

Srinivas, M., O'Sullivan, O., Cotter, P. D., Van Sinderen, D., & Kenny, J. G. (2022). The application of metagenomics to study microbial communities and develop desirable traits in fermented foods. Foods, 11(20), 3297. https://doi.org/10.3390/foods11203297

Tan, Y. L., Tan, F. A., & Chye, F. Y. (2024). Nutritional Properties of Selected Edible Insects. Biology and Life Sciences Forum, 40(1), 43. https://doi.org/10.3390/blsf2024040043

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1891-0

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

Yang, X., Noyes, N. R., Doster, E., Martin, J. N., Linke, L. M., Magnuson, R. J., Yang, H., Geornaras, I., Woerner, D. R., Jones, K. L., Ruiz, J., Boucher, C., Morley, P. S., & Belk, K. E. (2016). Use of Metagenomic Shotgun Sequencing Technology To Detect Foodborne Yang X, Noyes NR, Doster E, Martin JN, Linke LM, Magnuson RJ, Yang H, Geornaras I, Woerner DR, Jones KL, Ruiz J, Boucher C, Morley PS, Belk KE.2016. Use of Metagenomic Shotgun Sequencing Technology To Detect Foodborne Pathogens within the Microbiome of the Beef Production Chain. Appl Environ Microbiol 82:. https://doi.org/10.1128/AEM.00078-16

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell*, *178*(4), 779–794. https://doi.org/10.1016/j.cell.2019.07.010

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Marten Rikberg ,
(autori nimi)

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

   Putukate taksonoomiline klassifitseerimine erinevate andmebaaside abil ,
   (lõputöö pealkiri)

   mille juhendaja(d) on Maido Remm ,
   (juhendaja nimi)

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;

2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi 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.

Marten Rikberg

15.05.2025