TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOINFORMAATIKA ÕPPETOOL

Spetsiifiliste *k*-meeride leidmine inimesel toiduallergiat põhjustavate loomaliikide määramiseks

Bakalaureusetöö

12 EAP

Andrea Jõesaar

Juhendaja teadur Reidar Andreson

TARTU 2020

INFOLEHT

“Spetsiifiliste *k*-meeride leidmine inimesel toiduallergiat põhjustavate loomaliikide määramiseks”

Tänapäeval on ligi 2% täiskasvanutest toiduallergiad ja ka ristsaastumisel toitu sattunud allergeen osutuda mõnele allergikule ohtlikuks. Allergeenide tuvastamiseks proovist saab kasutada teise põlvkonna sekveneerimisel põhinevaid *k*-meere ehk *k* pikkuseid DNA lõike kasutavad meetodeid, mis pakuvad suurt tundlikkust ning täpsust. Tänu oma haploidsele pärandumisele, rohkusele keskkonnaproovis ja vähesele rekombinatsioonile on liikide eristamiseks tüüpiliselt kasutusel mitokondriaalne DNA.

Käesoleva töö põhieesmärk oli töötada välja metoodika spetsiifiliste *k*-meeride leidmiseks sihtmärgiks olevate loomaliikide mitokondritest. Loodud meetod, mis põhineb filtreerimisel kasutades programmi *nblast*, leiab ühele sihtmärgile *k*-meeride komplekti keskmiselt 34 minutiga. Leiti, et saadud *k*-meeride komplektid annavad üldiselt tugeva signaali, kui neid testida sihtmärgi täisgenoomil toorlugemitel.

Märksõnad: *k*-meerid, toiduallergia, NGS, GenomeTester4

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

“Finding specific k-mers to identify animal species which cause food allergy”

Around 2% of adults have food allergies and some may experience allergic reactions caused by food, which has had cross-contact with allergens. To detect allergens from a sample NGS-based *k*-mer (*k* length DNA section) methods, which offer high specificity and accuracy, can be used. Thanks to its haploid inheritance, abundance in environment samples and low recombination rates, mitochondrial DNA is often used for species’ detection.

The purpose of this thesis was to develop a method, which can be used to find a set of species-specific k-mers from the mitochondria which could be used to detect the species from an environment sample. The method which was developed is based on filtering with the program *nblast*, can find a set of *k*-mers for a species with an average of 34 minutes, which mostly produces a strong signal on the target species’ raw sequencing reads.

Keywords: *k*-mers, food allergy, NGS, GenomeTester4

CERS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

SISUKORD

KASUTATUD LÜHENDID

bp – aluspaar

CAS – *compare-and-swap*

COI – tsütokroom *c* oksüdaas I

ELISA – immunoensüümmeetod (*Enzyme-linked immunosorbent assay*)

IgE – Immunoglobuliin E

LC – vedelikkromatograafia (*Liquid chromatography*)

MS – massispektromeetria (*Mass spectrometry*)

NGS – teise põlvkonna sekveneerimine (*Next-generation sequencing*)

RAM – muutmälu (*Random Access Memory*)

SPR – pinna plasmaresonantsanalüüs (*Surface plasmon resonance*)

VTA – Veterinaar- ja Toiduamet

SISSEJUHATUS

Toiduallergiad on tänapäeval väga levinud ning tihti on ka väga väikestes kogustes allergeeni toidus piisav allergilise reaktsiooni tekitamiseks. Kuigi suuremates kogustes allergeeni olemasolu tuleb Eestis märkida toidupakendile, on ka väiksemate hulkade tuvastamiseks toidust tarvilik leida tundlikke meetodeid.

Allergeenide tuvastamiseks saab kasutada otseseid ja kaudseid meetodeid. Otsesed meetodid tuvastavad allergiat tekitavat valku, kuid võivad valgu denatureerumisel või teiste valkude interferentsil sihtmärkvalku mitte ära tunda. Kaudsed meetodid kasutavad allergeenide tuvastamiseks toiduallergiat tekitavate liikide DNA-d. Populaarsed DNA triipkoodimise meetodid kasutavad põhilise mehhanismina PCR-i; kasutatakse ka teise põlvkonna sekveneerimismeetodeid (NGS), mis on hea tundlikkusega ja sobilikud degradeerunud DNA järjestuse määramiseks.

Teise põlvkonna sekveneerimismeetoditega saadud toorlugemite (*raw reads*) andmeid on võimalik analüüsida kasutades *k*-meere, mille efektiivse loendamise väljakutset on proovitud lahendada mitmete algoritmidega. Käesolevas töös kasutatakse *k*-meeride loendamiseks programmide paketti GenomeTester4, mis küll ei ole *k*-meeride loendamises teiste rakendustega võrreldes kiireim, kuid võimaldab loendatud *k*-meeridega sooritada mitmeid ülesandeid, mida teised *k*-meeride loendamise programmid ei võimalda.

Käesoleva töö üldeesmärk on töötada välja meetod, millega on võimalik leida liigispetsiifiline *k*-meeride komplekt, mida saab kasutada uuritava liigi tuvastamiseks näiteks keskkonna- või toiduproovist. Töös leitakse taolised 32-meeride komplektid viiekümnele loomset toiduallergiat põhjustavale liigile. Töö teoreetilise osa eesmärk on koostada ülevaade meetoditest, mida kasutatakse inimesel toiduallergiat põhjustavate loomaliikide määramiseks.

1. KIRJANDUSE ÜLEVAADE

1.1 Toidupõhised allergiad ja loomsed allergeenid

Tänapäeval on toiduallergiad inimeste hulgas väga levinud – ligi 2% täiskasvanutest on mõni toiduallergia; ühed kõige levinumatest on koorikloomade (1.9%), kala (0.4%) ja limuste allergiad (Siragakis ja Kizis, 2014). Toiduallergiate all kannatavad tihti nooremad ning vanuse kasvades toiduallergiat omavate inimeste hulk väheneb (Hadley, 2006).

Toiduallergiat põhjustavad enamasti valgud. Näiteks tursa allergeen *Gad c 1* on valk, mis on tähtis musklite lõdvestamisel ning sarnane parvalbumiin on olemas ka enamustes teistes kalaliikides. Kuna see allergeen on stabiilne, ei saa seda kuumutamise, pH muutuste ja keemiliste modifikatsioonidega toidust eemaldada. Taolised omadused on iseloomulikud ka paljudele teistele toiduallergeenidele. (Siragakis ja Kizis, 2014)

Reaktsioon allergeenile võib toimuda väga vähese koguse allergeeni manustamisel, kuigi kogus varieerub inimeseti. Allergilisi reaktsioone on tuvastatud ka vähem kui 3 mg allergiat põhjustavat ainet sisaldava toiduga, kuid keskmine reaktiivsuse piir on ~30 mg. (Reier-Nilsen *et al.*, 2018)

Veterinaar- ja Toiduamet (VTA) nõuab toitudel teadliku ja tahtliku allergeenide sisalduvuse märkimist, kuid võimalike või tahtmatute allergeenide sisalduvus ei ole veel reguleeritud (VTA koduleheküIg, 2019). Valmistoitude puhul võib allergilise reaktsiooni põhjustada ristsaastumisega toidule üle kandunud allergeen (Alvarez ja Boye, 2012). Seetõttu on vaja kasutada tundlikke meetodeid, näiteks mitokondri regioone määrav test, allergeenide tuvastamiseks madalatel kontsentratsioonidel (Eischeid, 2019).

Andmebaas (AllergenOnline, 2019) sisaldab kõiki teaduslikult tõestatud ja ekspertide poolt kontrollitud allergiat põhjustavaid liike ning nende allergeene. Selleks, et potentsiaalne allergeen jõuaks AllergenOnline andmebaasi, peab olema tõestatud, et seostumine allergeeniga on spetsiifiline, tavaliselt läbi homoloogsete valkude testimise. Samuti, et välistada tõenäolised valepositiivsed allergeenid, peavad potentsiaalsed allergeenid vastama järgnevatele tingimustele: peab olema tõendatud Immunoglobuliin E (IgE) seostumine; allergeeni järjestus ei tohi olla patenteeritud; välistatud on parasiitsed valgud ja valkude järjestused, mis on lühemad kui 8 aminohapet; peab leiduma avalik viide allergeensusele,

mis on pikem kui lühike kommentaar NCBI valkude andmebaasis. Andmebaasi kriteeriumeid uuendatakse iga aasta. (Goodman *et al.*, 2016)

Tänu antud kriteeriumitele sisaldab AllergenOnline andmebaas vaid IgE seonduvaid allergeene, kuigi ka alternatiivsed allergilise reaktsiooni tekkimise viisid on võimalikud (Siragakis ja Kizis, 2014). Kuna IgE seondumiseta toiduallergiaid on raske kontrollida ning määratleda (Meyer *et al.*, 2020), on nende käsitlus ka väljaspool käesolevat tööd.

1.2 Meetodid allergeenide detekteerimiseks

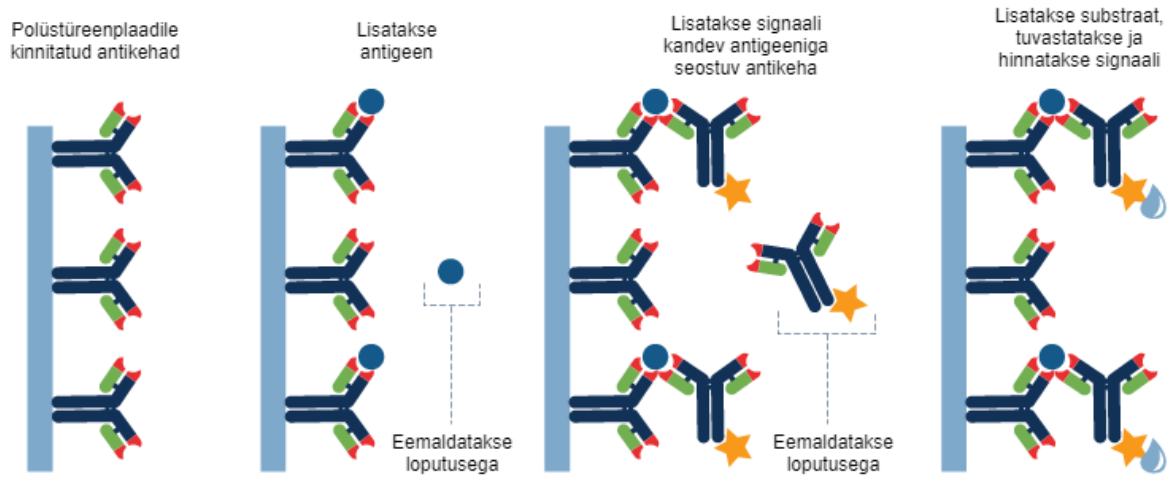
1.2.1 valgupõhised meetodid allergeenide detekteerimiseks

Allergia põhjustajaks on enamasti valgud, seega valgupõhised meetodid on otsesed allergeenide detekteerimise meetodid. Nendest põhilisteks on immunoensüümmeetod (ELISA), massispektromeetria (MS), vedelikkromatograafia (LC) ja pinna plasmaresonantsanalüüs (SPR).

ELISA on allergeenide tuvastamiseks üks enam kasutatud meetodeid tänu oma lihtsusele ning kättesaadavusele (Fernandes *et al.,* 2015). Allergeenide tuvastamine viiakse läbi polüstüreen plaadil, mis seob passiivselt antikehasid ning valke. Allergeeni tuvastamisel seondub antikeha allergeeniga; ülejäänud valgud eemaldatakse. Substraadi lisamisel vabaneb signaal, mis on proportsionaalne seondunud allergeeni kontsentratsiooniga (Engvall ja Perlmann, 1972).

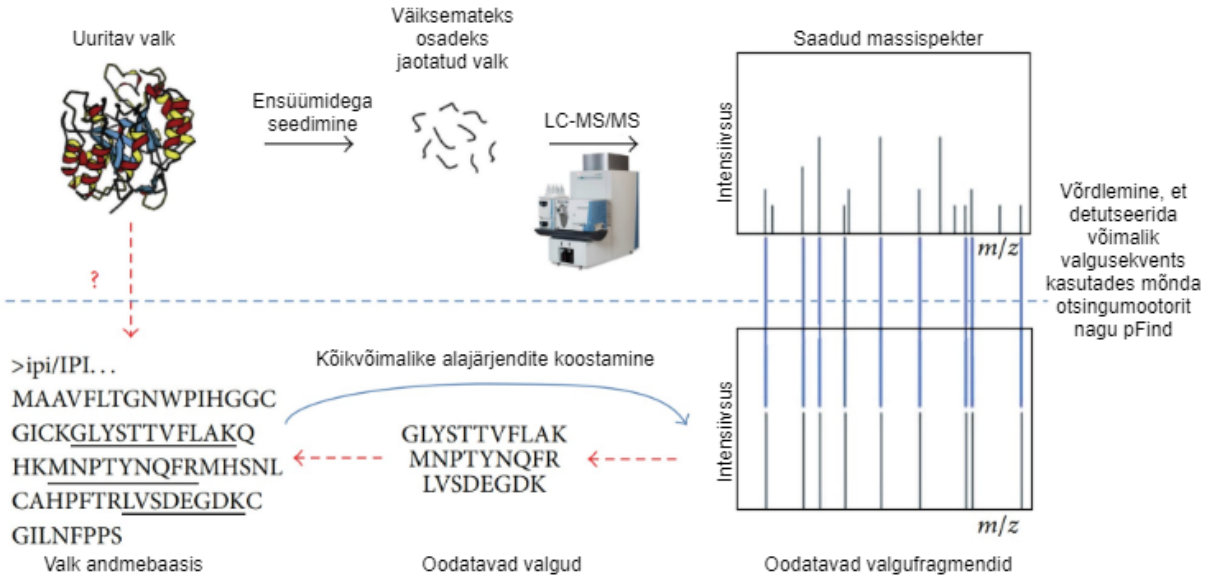
ELISA tööpõhimõte on välja toodud joonisel 1. Joonisel on kirjeldatud otsene kihiline ELISA, kuid allergeeni ja antikeha seondumise ning signaali saamise mehhanisme on erinevaid.

ELISA puudus on vähene spetsiifilisus – lisaks sihtmärgile võivad vahel valepositiivse signaali anda teised valgud. Kui valgu töötlemise tulemusena tekivad selle struktuuris muutused, võib sihtmärk ka tuvastamata jääda, kuna kasutatud antikehad ei tunne seda ära (Fernandes *et al.,* 2015).



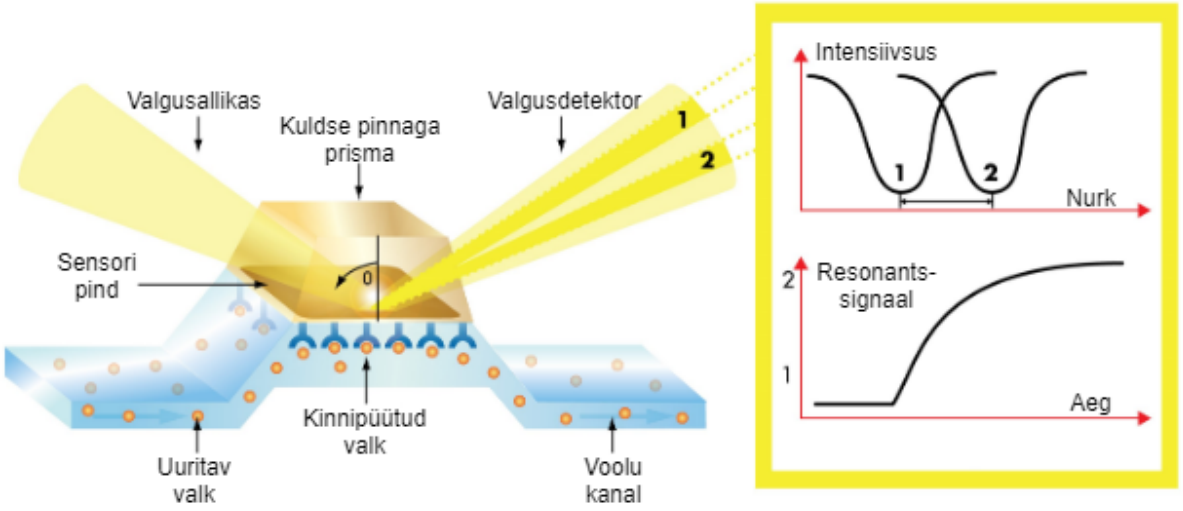
Joonis 1. Otsene kihiline ELISA. Polüstüreenplaadile kinnitatud antikehad on võimelised seonduma nõrkade keemiliste sidemetega spetsiifilise valguga (allergeeniga). Lisatud uuritavas vedelikus olevad spetsiifilised valgud (allergeenid) seonduvad antikehadega ning ülejäänud vedelik eemaldatakse loputusega. Lisatakse signaali kandev antikeha, mis seondub valgu (allergeeni) mõne teise lookusega. Substraadi lisamisel vabaneb mõõdetav signaal. (Kohandatud: Rockland Elisa Assays, i.a)

MS meetodil on mitu variatsiooni – uuritavad valgud läbivad enne analüüsi ensüümidega seedimise (proteolüütiliste ensüümide abiga lagundatakse valgud väiksemateks osadeks) või analüüsitakse terveid valke. Tervete valkude analüüsimise korral on võimalik eristada ka valkude erinevaid isovorme. Mõlema meetodi korral tehakse valgu osadeks jaotamisega kindlaks valgu sõrmejälg, mille järgi on võimalik eristada erinevaid allergeene. Tihti kasutatakse massispektromeetriaga koos ka LC-d, et suurendada meetodi dünaamilist ulatust. Sellised valgu tuvastamise viisid on aga väga aeganõudvad. (Monaci ja Visconti, 2009) Meetodi kirjeldus on toodud joonisel 2.



Joonis 2. Massispektromeetria ülevaade. Tuvastades valku MS meetodil võib esialgu toimuda valgu väiksemateks osadeks jaotamine ning seejärel mõõdetakse valgu (või valgu väiksemate osade) massispekter (massispektrid). Massispektri alusel saab ennustada valgu aminohappejärjestust, mida andmebaasist otsitakse. Kuna valgu väiksemateks osadeks jaotamine võib toimuda suvalise peptiidsideme kohalt, on tarvilik pärast valgu andmebaasis leidmist ka sellest valgust kõikvõimalikkude alajärjendite koostamine, et veenduda leitud valgu sobilikkuses algandmetega. (Kohandatud Zhang ja Zhao, 2013)

SPR on valkude füüsikalistel omadustel põhinev meetod, kus mõõdetakse valguse murdumist sihtmärgi seostumisel sensoril asuva sondiga. SPR tööpõhimõte on toodud joonisel 3. Kuigi meetod on küllaltki täpne ning võimaldab aminohapete üksikahela äratundmist, võivad lahuses olevad mitte-sihtmärkmolekulid signaali häirida. (Šípová ja Homola, 2013)



Joonis 3. SPR tööpõhimõte. Uuritav proov lisatakse voolukanalisse. Kui valk seondub sensoriga, muutub selle pinnale langeva valguse nurk ja intensiivsus. Saadud andmed salvestatakse jooksvalt ning salvestatud signaali profiili järgi on võimalik erinevaid valke omavahel eristada. (Kohandatud: Bruker Surface Plasmon Resonance, i.a)

1.2.2 DNA-põhised meetodid allergeenide detekteerimiseks

Kui toidus on vabasid (loomseid või taimseid) allergeene, leidub selles ka nende allergiat tekitavate liikide DNA-d. Seega allergeenide olemasolu ning hulga hindamiseks saab kasutada ka DNA-põhiseid meetodeid. Enamkasutatud DNA-põhised meetodid on DNA triipkoodimine (*barcoding*), DNA meta-triipkoodimine (*metabarcoding*) ja DNA mini-triipkoodimine (*minibarcoding*). Triipkoodimise ja meta-triipkoodimise ülevaatlik joonis on toodud lisas 1.

DNA triipkood on genoomi erinevates positsioonides asuvate järjestuste kogum, mille abil on võimalik kindlat liiki ära tunda. Selleks otsitakse triipkooditava liigi DNA järjestusest teistest liikidest hästi eristatavaid alasid. (Hebert *et al.*, 2003)

DNA triipkoodimise meetodiga kasutatakse pigem pikemaid (~600 aluspaari ehk bp) järjestusi, sest kodeerivate alade iga kolmas nukleotiid on vähese varieeruvusega ja lühikeste järjestuste korral on raskem lähedasi liike eristada (Hebert *et al.*, 2003). Liigispetsiifilisi järjestusi otsitakse eelkõige mitokondritest nende haploidse pärandumise ja vähese rekombinatsiooni tõttu (Saccone *et al.*, 1999). Enamjaolt kasutatakse mitokondris asuvat tsütokroom *c* oksüdaas I (COI) geeni, mille kiirete evolutsiooniliste omaduste (Knowlton ja Weigt, 1998) tõttu saab väga hästi eristada lähedasi liike. PCR-iga amplifitseeritakse valitud järjestused, kasutades nende otsadega komplementaarseid praimereid. PCR produkti saab Sangeri sekveneerimise abiga lugeda kui triipkoodi.

DNA mini-triipkoodimisega vähendatakse degradeerunud DNA mõju triipkoodimisele analüüsides lühemaid DNA fragmente. Seega eelneva meetodi ~600 bp asemel on ühe analüüsitava järjestuse pikkuseks 100-200 bp. Liigispetsiifilisi mini-triipkoodimise piirkondi on aga tihti raske leida, sest lühemate järjestuste juures on suurem tõenäosus, et sama triipkood on ka mõnel teisel liigil. (Shokralla *et al.*, 2015)

DNA meta-triipkoodimisel amplifitseeritakse universaalsete praimeritega PCR-is üles DNA lõigud, mille järjestust seejärel sekveneerimise abiga analüüsitakse ning võrreldakse referentsandmebaasidega. Amplifikatsioonipiirkondadena on populaarsed COI ning 18S järjestused. (Horton *et al.*, 2017)

Kuna kõik eelnimetatud meetoditest põhinevad PCR-il, siis on neil kõigil sarnane puudus – need töötavad vaid olukordades, kus PCR-il kasutatud polümeraas hästi seondub. Kui aga DNA on liigselt degradeerunud, nagu see töödeldud toitudel tihti on, võib PCR-i efektiivsus olla liigselt madal, et saada usaldusväärseid tulemusi. (Raime ja Remm, 2018)

1.2.3 Teise põlvkonna sekveneerimisandmetest *k*-meeride loendamisel põhinevad meetodid

Teise põlvkonna sekveneerimismeetodeid (NGS) iseloomustab vähese ajaga suure hulga lühikeste järjestuste (70-300 bp) ehk lugemite genereerimine. NGS meetodid kasutavad sekveneerimise põhimehhanismina kas sünteesi (Illumina ja Roche) või otste ligeerimist (Life Technologies). (Chiu, 2015)

Illumina tehnoloogiaga loetakse proovist DNA järjestus kasutades fluorestsentsmarkeriga nukleotiide. Need lisatakse plaadile praimeritega seondunud üksikahelalisele DNA-le igal tsüklil polümeraasi abil. Valgusega ergastumisel annavad fluorestsentsmarkerid nukleotiide diferentseeriva signaali. Seejärel markerid eemaldatakse ja DNA 3' ots vabastatakse edasiseks pikendamiseks. (Chiu, 2015)

K-meeride põhised meetodid kasutavad DNA tuvastamiseks proovist sekveneerimise toorandmeid, millest otsitakse eelnevalt kindlaks tehtud liikide-spetsiifilisi *k* pikkuseid DNA järjestusi ehk *k*-meere. Piisava *k*-meeride hulga leidmise korral proovist saab öelda valdava kindlusega, et proovis on vastav liik.

K-meeridel põhinevad järjestuste analüüsimise meetodid on kasutusel olnud viimasel kahel kümnendil. Esimesena pakuti *k*-meeridele rakendus genoomide assembleerimiseks kasutades *De Bruijni* graafi põhimõtteid (Pevzner *et al.*, 2001). Tänapäeval rakendatakse *k*-meere ka joondamisvaba tuvastusmeetodina. *K*-meeridel põhinevate meetodite aluseks on NGS toorlugemitest *k*-meere loendav algoritm. Kui efektiivsus rolli ei mängiks, ei oleks *k*-meeride loendamiseks loodud suurel hulgal algoritme, sest tegemist on arvutuslikus mõttes lihtsa protsessiga. Järgnevates peatükkides tutvume mõnedega nendest algoritmidest.

1.2.3.1 Algoritmid *k*-meeride loendamiseks lugemitest

Jellyfish kasutab *k*-meeride loendamiseks räsilauda (*hash table*), mille teeb kiireks võimekus mitmel tuumal paralleelselt loendamisoperatsioone läbi viia. Seda toetab *compare-and-swap* (CAS) operatsioon, mis võimaldab mitmete paralleelsete operatsioonide sünkroniseerimist. Kui soovitakse *k*-meeri räsilauale lisada, leitakse selleks kas tühi koht (mille leidmises osaleb CAS ja asukoha määrab räsifunktsioon) või suurendatakse sama *k*-meeri loenduste arvu ühe võrra. Seetõttu ei kustutata eelnevaid räsisid tabelist kunagi ära. Kuna räsilaual säilitatakse vaid pakitud *k*-meerid (võtmed) ning nendega seotud korduste arv (väärtused), on räsilaud

ka mälukasutuselt väike. Kusjuures üks mitmeid kordi samas lugemis sisalduv *k*-meer võib olla räsilaual mitu korda, kus iga järgnev sissekanne sisaldab esimesele viidet, millega hoitakse võtme kirjutamata jätmisega samuti mahtu kokku, sest väärtuste väli ei pea tingimata võimaldama suurima *k*-meeri kordsuse mahutamist. (Marçais ja Kingsford, 2011)

DSK algoritmis jaotatakse *k*-meerid räsifunktsiooni alusel liigendusteks, mis kirjutatakse püsimällu, kus iga *k*-meer kasutab $2^{\lceil log_2(2k) \rceil}$ bitti. Seejärel vaadeldakse liigendusi ükshaaval ning räsitabelit kasutades *k*-meerid loendatakse. Tänu püsimälu kasutamisele kulutab DSK väga vähesel määral vahemälu, kuid protsessi kiirus sõltub otseselt kasutava kõvaketta kiirusest. (Rizk *et al.*, 2013)

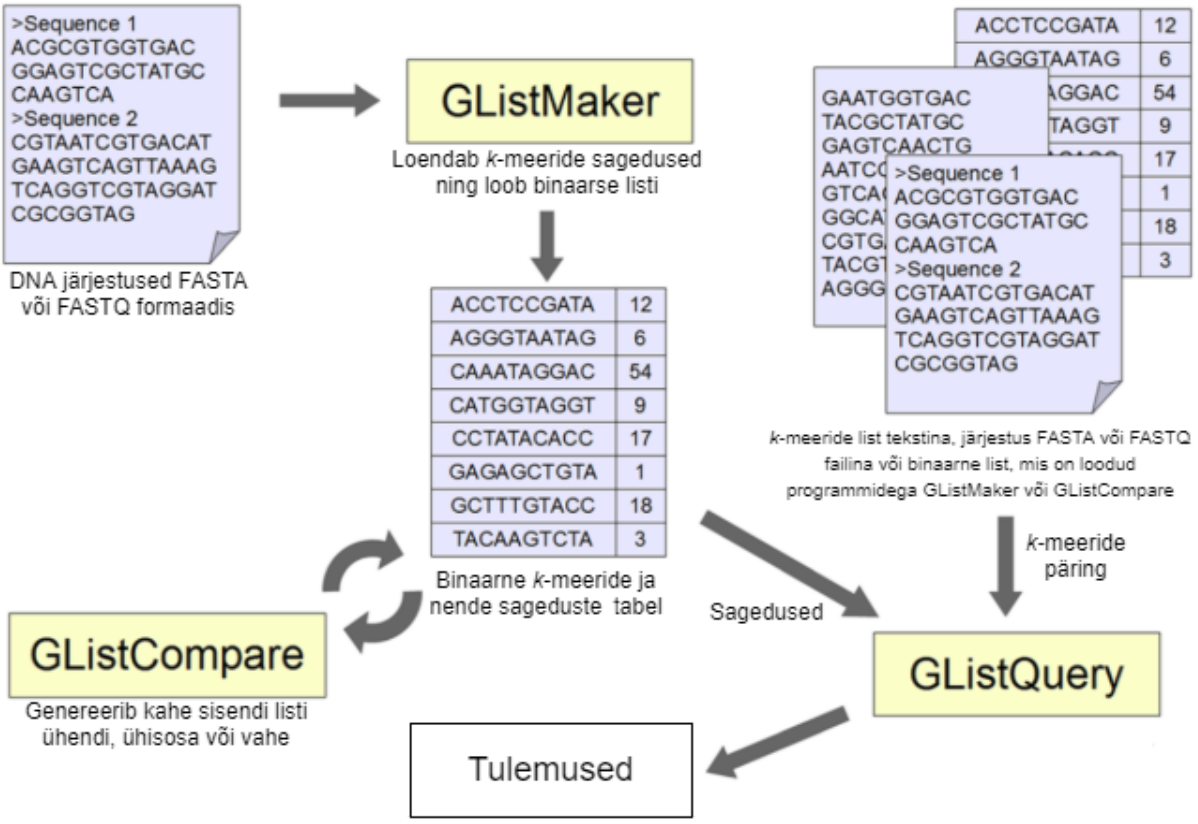
Turtle rakendab loendamisel *bloom* filtrit, millega toorandmetest läbi käimisel tehakse kindlaks need *k*-meerid, mida esineb rohkem kui üks kord. Need lisatakse seejärel jooksvalt listi, kus juhul, kui nende hulk ületab mingi kindlakstehtud lävi, pakitakse *k*-meer, lisatakse listi ning sorditakse. (Roy *et al.,* 2014)

KMC 2 kasutab ära minimiseerijate (leksikograafiliselt vähimat *m*-meeri, kus *m*<*k*) ideed ja (*k,x*)-meere, ehk kanoonilisi (*k+x'*)-meere, kus *x'* on selline, et $0 \leq x' \leq x,$ ehk iga $l \leq k + x$ korral l-meer on kanooniline (s.t päri- ja vastassuunalisi *k*-meere arvestatakse identsena), et *k*-meeride hoiustamisel ja sortimisel mäluruumi vähendada. Programm loendab *k*-meere kahes sammus: esimeses loeb programm sisse osaliselt kattuvad DNA alad, millel on sama signatuur ehk sisaldavad sama minimiseerijat, millel on mälumahtu võtvate polü-A järjestuste vähendamiseks keelatud sisaldada algust "AAA", "ACA" või sisaldada järjestust "AA" kõikjal peale alguse. Need sama signatuuriga "super *k*-meerid" sorditakse samasse "kasti". Mõned signatuurid samastatakse algandmetest loodud histogrammis leitud vähima sageduse alusel. Teises ehk sortimise sammus saadakse super *k*-meeridest (*k*,*x*)-meerid, mis radix sortimisalgoritmi alusel sorditakse. Pärast kanooniliste *k*-meeride alusel edasist sortimist loendatakse saadud *k*-meerid kokku. See algoritm on eelnevalt kirjeldatutest kiirem. (Deorowicz *et al.*, 2015)

1.2.3.2 GenomeTester4

GenomeTester4 kasutab *k*-meeride säilitamiseks listis sarnaselt algoritmile Turtle kahte bitti iga nukleotiidi jaoks. Säilitatakse iga *k*-meeri kanoonilist vormi. Loendid on salvestatud 32 bitiga. (Kaplinski *et al.*, 2015)

Kõikide GenomeTester4 paketti kuuluvate programmide töövoogu kirjeldab joonis 4.



Joonis 4. GenomeTester4 programmide töövoog. Arvutuslikke protsesse viivad läbi *GListMaker* ning *GListCompare* ja programm *GListQuery* väljastab nende programmide poolt koostatud liste või erinevas formaadis failide peal lihtsaid päringuid, näiteks väljastab keskmise GC-nukleotiidide sisalduse failis. Arvutuslikke protsesse läbi viivate programmide väljastatud listifailid ei ole loetavad *GListQuery*t kasutamata, kuid seevastu programmide *GListMaker* ja *GListCompare* binaarne tabeli formaat kasutab vähem püsimälu. (Kohandatud Kaplinski *et al.*, 2015)

GenomeTester4 paketis on *k*-meeride loendamiseks *GListMaker*. Sellega protsessitakse sisendifaili, mille jooksul kogutakse kõik *k*-meerid ühte ajutisse reastusesse, mis seejärel kasutades mitut paralleelset lõimu sorditakse ning järjestikused samad *k*-meerid loendatakse. (Kaplinski *et al.*, 2015)

Algoritmi kiirus sõltub suuresti sisendifaili suurusest. Kuna kõik *k*-meerid kogutakse ühte reastusesse, võivad suured failid kiiresti kogu mälumahu täita. Väikeste lugemite korral, nagu selleks on näiteks bakteri või mitokondri genoomid, toimib algoritm aga rahuldava kiirusega, võimaldades samas väiksemat muutmälu (RAM) kasutust. (Kaplinski *et al.*, 2015)

GenomeTester4 paketis leidub ka erinevate *k*-meeride liste töötlev *GListCompare*. Sellega saab leida kahe listi ühisosa, ühendit ja vahet. Algoritm ei loe *k*-meere otse RAM-is vaid kasutab Linuxi *mmap* funktsiooni, seetõttu on antud operatsiooni RAM kasutus väga väike. (Kaplinski *et al.*, 2015)

Kolmas GenomeTester4 paketis sisalduv programm on *GListQuery*, mis eelneva kahe programmi poolt genereeritud binaarsed listid väljastab kasutajale. Kuigi antud programmil on veel funktsionaalsusi, siis neid töös ei rakendata. (Kaplinski *et al.*, 2015)
Kõiki kolme ülalnimetatud GenomeTester4 paketi programmi kasutatakse ka käesolevas töös.

1.2.3.3 PlantTaxSeeker

Programmide pakett (PlantTaxSeeker, 2018) võimaldab kloroplastist taksoni-spetsiifiliste *k*-meeride leidmist. Paketi programmid põhinevad GenomeTester4 programmidel. Paketti kuuluvad programmid on *identification_of_taxon_specific_kmers.py*, millega leitakse kloroplastidest kloroplastide vastu filtreerides *k*-meerid; *filtering_with_nontargets.py*, millega filtreeritakse välja mittespetsiifilised *k*-meerid kasutades sihtmärkliikidele sarnaste liikide sekveneerimise toorandmeid; ja *plant_taxa_kmers_counter.py*, millega tuvastatakse sekveneerimise toorandmetest taksoni-spetsiifiliste *k*-meeride abil taksoni olemasolu proovist (mille lugemeid analüüsiti). Käesolevas töös on kasutusel vaid esimesed kaks programmi, millega tutvume siinses peatükis.
Programm *identification_of_taxon_specific_kmers.py* võtab sisendiks vastu (lisaks *k* väärtusele) kaks FASTA faili: üks nendest sisaldab vaid huvipakkuva taksoni järjestusi (edaspidi sihtmärgid) ja teine nendega võrreldavate liikide järjestusi (edaspidi mitte-sihtmärgid). Kõikidest sihtmärkidest tehakse *k*-meeride listid programmiga *GListMaker*, võetakse ühisosa programmiga *GListCompare*, eemaldatakse *k*-meerid, mis on liiga vähestes sihtmärkliikides (programmiga *GListCompare*) ning seejärel eemaldatakse (programmiga *GListCompare*) *k*-meerid, mis on mitte-sihtmärkliikide järjestuste *k*-meeride listis. Seega tulemuseks saadakse *k*-meerid, mis sisalduvad *k*-meeri taga oleva numbriga märgitud koguses sihtmärk-taksoni liikides, millest ükski *k*-meer ei sisaldu mitte-sihtmärk liikides.
Programm *filtering_with_nontargets.py* järgib sarnast protsessi, nagu eelnev programm, kuid võtab sisendiks eelnevalt genereeritud sihtmärkide *k*-meeride listi ning mitte-sihtmärkide sekveneerimise toorandmed või FASTA formaadis täisgenoomid. Mittesihtmärkide sisenditest tehakse *k*-meeride listid programmiga *GListMaker*, eemaldatakse väga väikse sagedusega *k*-meerid programmiga *GListCompare* ning eemaldatakse saadud listiga võrreldes mitte-spetsiifilised *k*-meerid sihtmärk-*k*-meeridest programmiga *GListCompare*.

Eelnevaid programme kasutatakse ka modifitseeritud kujul käesolevas töös. Modifitseerimata kujul programme kasutada ei saa järgnevatel põhjustel:

1. Kõik PlantTaxSeekeri paketi programmid on mõeldud ühel liigil korraga kasutamiseks. Käesolev töö soovib leida spetsiifilised *k*-meerid korraga mitmele liigile
2. PlantTaxSeeker kasutab esialgsete *k*-meeride leidmiseks mitme liigi isendi järjestusi, mitte RefSeq järjestust, seega käesolevas töös ei ole sisendi ühisosa leidmine tarvilik
3. PlantTaxSeeker kaotab *k*-meeri ühes genoomis esinemissagedused, mida on proovist tuvastamisel proportsionaalseks võrdluseks käesolevas töös kasutatud.

2. EKSPERIMENTAALOSA

2.1 Töö eesmärgid

Töö põhieesmärk on töötada välja metoodika spetsiifiliste *k*-meeride leidmiseks sihtmärgiks olevate loomaliikide mitokondritest. Täiendavad eesmärgid on:

1. anda ülevaade meetoditest inimesel toiduallergiat põhjustavate loomaliikide määramiseks;
2. leida *k*-meerid, mis oleksid spetsiifilised ka täisgenoomide puhul.

2.2 Materjal ja metoodika

2.2.1 Riistvara, tarkvara ja andmestik

Tarkvarana kasutati GenomeTester4 (versioon 4.0), blast-2.9.0+ (Altschul *et al.*, 1990), ja Python 3.3 (failinime laiend PY viitab selle versiooni Pythoni programmile).

Andmestikuna olid kasutusel (RefSeq mitokondriaalse DNA andmebaas, i.a); (NCBI nt, refseq_genomic ja other_genomic andmebaasid, i.a) ja (SRA andmebaas, i.a).

Kogu protsessi jooksul kasutati CentOS Linux 7 (Close) serverit, millel on 32-tuumaline Intel(R) 2.27GHz protsessor ja 500GB RAM mälu.

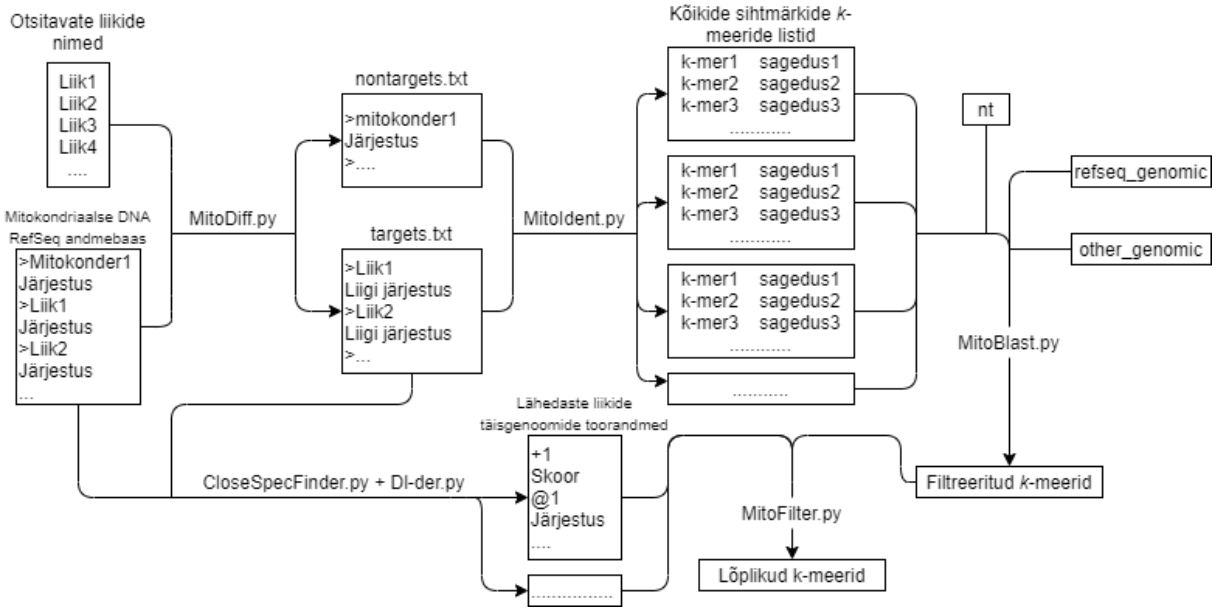
Kõik töö jaoks koostatud programmid ning tulemused ja statistikafail on saadaval (tulemused keskkonnas Github, 2020).

2.2.2 Töövoog

Töövoog põhineb tööl (Raime ja Remm, 2018), kus *k*-meerid leitakse kloroplasti andmebaasist ning filtreeritakse lähedaste liikide toorandmeid kasutades programmide paketiga (PlantTaxSeeker, 2018). Eeltoodud töövoole on käesolevas töös lisatud programmi *blastn* kasutav filtreerimine, mis tänu enda lihtsusele ning kiirusele võib osutuda toorandmetega filtreerimisele paremaks alternatiiviks.

Spetsiifiliste *k*-meeride saamiseks loodi järgmine töövoog: esiteks eraldatakse RefSeq mitokondriaalse DNA andmebaasi järjestustes sihtmärkliigid mitte-sihtmärkidest, seejärel leitakse GenomeTester4 abiga sihtmärkidele spetsiifilised *k*-meerid, nendele tehakse esimene filtreerimine programmiga *blastn* ning teine filtreerimine lähedaste liikide

sekveneerimise toorandmetega. Neid protsesse viivad läbi programmid *MitoDiff.py*, *MitoIdent.py*, *MitoBlaster.py*, *CloseSpecFinder.py*, *Dl-der.py* ja *MitoFilter.py* (edaspidi failinime laiendita), mida saab käivitada käsureal. Kogu töövoog on toodud joonisel 5.



Joonis 5. Liigispetsiifiliste *k*-meeride leidmise töövoog. Failinime laiend PY viitab Pythoni programmile. Põhiline töövoog koosneb programmidest *MitoDiff*, *MitoIdent*, *MitoBlast* ja *MitoFilter*, kuid programmi *MitoFilter* tööks on tarvilik ka lähedaste liikide täisgenoomide toorandmed, mille leiavad ning laadivad alla programmid *CloseSpecFinter* ja *Dl-der*.

Järgnevates alapeatükkides on ülevaatlikult välja toodud antud programmide tööpõhimõtted ning nende juures tehtud määravad otsused. Lisaks kõigele kirjeldatule salvestavad programmid jooksvalt informatsiooni kogutud *k*-meeride kohta faili Statistics.csv.

Programmide tööks on tarvilik, et nendega samas kaustas oleks GenomeTester4 programmid *glistmaker*, *glistcompare* ja *glistquery*, programmi *Dl-der* jaoks ka *sratoolkit*.

2.2.2.1 Programm *MitoDiff*

Eelnevate tööde põhjal (Saccone *et al.*, 1999; Shokralla *et al.,* 2015; Horton *et al.*, 2017; Eischeid, 2019) võib öelda, et mitokondriaalne DNA sobib liikide omavaheliseks eristamiseks. Kuna andmebaasis (RefSeq mitokondriaalse DNA andmebaas, i.a) on 10111 erineva liigi (taksoni) mitokondrite täisjärjestused, on tegemist andmestikuga, millest leiab suure tõenäosusega enamike otsitavate allergiat tekitavate liikide järjestused.

Kõik allergiat tekitavate kalade, limuste ja koorikloomade järjestused saadud andmebaasist AllergenOnline. Programmi *MitDiff* sisend on liikide (ladinakeelsed) nimed reavahetusega eraldatult tekstifailis. Lisaks eeldatakse, et kaustas, milles programm käivitatakse, on olemas

ka lahti pakitud kujul RefSeq mitokondriaalse DNA andmebaas (s.t failid mitochondrion.1.1.genomic.fna ja mitochondrion.2.1.genomic.fna). Programmi tööpõhimõte on järgmine:

1. Avatakse nimekiri otsitavate liikide nimedena ning salvestatakse see vahemällu listina *matching*.
2. Avatakse kolm tekstifaili, millesse hakatakse järgnevate sammudega RefSeq mitokondriaalse DNA andmebaasi ridamisi ümber kirjutama: targets.txt, nontargets.txt ja repeated.txt.
3. RefSeq mitokondriaalse DNA andmebaasi loetakse rida-rea haaval:
    3.1. Kui jõutakse reani, mis algab sümboliga “>”, omandab muutuja *switch* väärtuse 0. Seejärel otsitakse listi *matching*, kas mõni selle element sisaldub loetavas reas. Kuna RefSeq andmebaasis sisalduvad ka liikide ristandid, siis kontrollitakse ka “ x ” sisaldust listi *matching* kandes, et ristandit ei loetaks tahtmatult sobivaks liigiks.
        3.1.1. Kui mõni listi *matching* kanne sisaldub kontrollitavas reas, lisatakse leitud liik listi *found*, kirjutatakse kontrollitav rida faili targets.txt ümber ja muutuja *switch* omandab väärtuse 1.
        3.1.2. Kui listi *matching* kanne sisaldub kontrollitavas reas, kuid see on ka juba listis *found,* siis kirjutatakse rida ümber faili repeated.txt ning muutuja *switch* omandab väärtuse 3.
        3.1.3. Kui list *matching* läbitakse vastet leidmata (s.t muutuja *switch* väärtus on pärast listi läbimist 0), kirjutatakse rida ümber faili nontargets.txt ja muutuja *switch* omandab väärtuse 2.
    3.2. Iga rea korral, mis eelnevale tingimusele ei vasta, kirjutatakse rida muutuja *switch* alusel ümber endale vastavasse faili.
4. Tekstifailid suletakse ning väljastatakse mitte leitud liikide nimed ning nende kogus.

Korduvalt leitud liikidest jääb selle protsessiga üks faili targets.txt ning ülejäänud faili repeated.txt. Järgnev programm võtab sisendina vaid kaks faili, seega on kasutaja otsustada, mida teha korduvate järjestustega. Käesolevas töös on kõik korduvad (kaasaarvatud üks, mis oli eelnevalt targets.txt failis) liigid lihtsuse huvides välja jäetud (s.t lisatud faili nontargets.txt), sest töö eesmärk ei ole kõikidele allergiat põhjustavatele liikidele *k*-meeride leidmine.

2.2.2.2 Programm *MitoIdent*

Järgmise sammuna, pärast sihtmärkide eristamist mittesihtmärkliikidest, leiab programm *MitoIdent* sihtmärkliikidele *k*-meerid.

Programm *MitoIdent* kasutab muudetud kujul paketi PlantTaxSeeker programmi *identification_of_taxon_specific_kmers.py*, et leida kõikidele sihtmärkliikidele *k*-meerid. Eelnimetatud programm ei sobi muutmata kujul kasutamiseks seetõttu, et see leiab kõikide targets.txt failis toodud liikidele ühised *k*-meerid (võttes seda kui ühte taksonit), kuid käesolevas töös on tarvilik kõikide targets.txt liikide eraldi vaatlemine.

Ühte liiki sihtmärgina vaadeldes ei ole vajalik leida sihtmärktaksoni *k*-meeride listide ühisosa, seetõttu jäetakse need sammud vahele. *K*-meeride leidmiseks kasutatakse GenomeTester4 programmi *GListMaker*, listide ühendi ja vahe leidmiseks programmi *GListCompare*.

Programmi *MitoIdent* tööpõhimõte on järgmine:

1. Sisendina võetakse eelneva programmi poolt eristatud sihtmärkliikide fail ning mittesihtmärkliikide tekstifail.
2. Leitakse mittesihtmärkide *k*-meerid.
3. Kogu järgnev programm on tsüklis. Tsükkel lõppeb, kui kõikidele sihtmärkidele on leitud *k*-meerid.
4. Avatakse targets.fna ja nontargets.fna failid (eelneva tsükli sisu kustutatakse).
5. Igal tsükli iteratsioonil lisatakse üks sihtmärkliik faili targets.fna ning ülejäänud liigid ja mittesihtmärkliigid lisatakse faili nontargets.fna.
6. Failidest targets.fna ja nontargets.fna luuakse *k*-meeride listid.
7. Leitakse nontargets.fna *k*-meeride listi ühend mittesihtmärkide *k*-meeridega, et saada list kõikidest sellel tsüklil mitte-sihtmärkliikide *k*-meeridest.
8. Leitakse sihtmärgi *k*-meerid, eemaldatakse nendest mitte-sihtmärkide *k*-meerid.

Programmi tulemus on igale sihtmärkliigile list *k*-meeridest, mille listis igale *k*-meerile järgneb selle sagedus sihtmärgis. *K*-meerid on liigispetsiifilised RefSeq mitokondrite andmebaasis.

2.2.2.3 Programm *MitoBlast*

BLAST paketi programmid on lühikeste (32-pikkuste oligomeeride) järjestuste korral andmebaasist homoloogsete vastete leidmiseks antud andmestiku korral piisavalt kiired ning

joondamine on lihtne protsess. Seega on mõistlik liigispetsiifiliste *k*-meeride leidmisel teha järgmine filtreerimine kasutades BLAST-i.

BLAST paketi progammid võtavad sisendina FASTA formaadis faile, seega eelnevast programmist saadud formaat ei sobi. Seetõttu muudab programm *MitoBlast* esimese sammuna iga liigi tabuleeritud väljundifaili iga rea

k-meer n

järgmiseks (FASTA formaadis):

>liigi_nimi_n_kmer_i
k-meer

kus *i* on rea järjekorranumber. Seejärel käivitatakse iga sihtmärkliigi *k*-meeridega andmebaase nt, refseq_genomic ja refseq_other kasutades programm *blastn* järgmiste parameetritega:

-out *liigi_nimi*_kmers_*andmebaas*.txt -outfmt '6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore stitle' -perc_identity 90 -qcov_hsp_perc 100 -num_threads 20")

Parameeter -qcov_hsp_perc 100 tagab, et joondatakse kogu *k*-meer, mitte vaid osa sellest.

Seejärel eemaldatakse iga liigi BLAST-i tulemusfailide põhjal mitte-liigispetsiifilised *k*-meerid järgnevalt:

1. Iga andmebaasi tulemusfaili igal real kontrollitakse, kas vaste andnud liigi nimi erineb *k*-meeri liigi nimest. Positiivse vastuse korral lisatakse *k*-meeri number *i* listi *found*.
2. Korduvad numbrid eemaldatakse listist
3. Loetakse BLAST-i sisendina antud *k*-meeride FASTA faili. Kui nimereal olev *k*-meeri number *i* ei ole listis, kirjutatakse *k*-meer liigi väljundifaili ümber.

Programmi väljundiks on igale liigile FASTA formaadis fail *k*-meeridega, mis ei esine kontrollitud andmebaasides.

2.2.2.4 Programmid *CloseSpecFinder* ja *Dl-der*

Et viia läbi *k*-meeride filtreerimine sihtmärkliikidele lähedaste liikide sekveneerimise toorandmetega, tuvastatakse lähedased liigid programmiga *CloseSpecFinder* ja laetakse nende sekveneerimistoorandmed andmebaasist alla programmiga *Dl-der*.

Enne programmi CloseSpecFinder.py tuleb eelnevalt käsureal käivitada järgmised käsud:

“cat mitochondrion.1.1.genomic.fna > database.txt”
“cat mitochondrion.2.1.genomic.fna >> database.txt”
“module load blast-2.9.0+”
“makeblastdb -in database.txt -dbtype nucl -parse_seqids”

“blastn -query targets.txt -db database.txt -out blasted_targets.txt -outfmt '6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore stitle' -perc_identity 20 -num_threads 20”

Sellega on BLAST leidnud RefSeq mitokondrite andmebaasis sihtmärkliikide järjestustega kattuvad liigid. Programm *CloseSpecFinder* kasutab selle BLAST-i otsingu väljundit, et selgitada välja lähedased liigid. Programm töötab järgnevalt:

1. Kuna BLAST-i tulemusfail jaotab iga järjestusega joondatud järjestused bittskoori (*BLAST bit-score*) järgi, saab tulemusfaili lugedes iga liigi esimesi vasteid lugeda parimateks. Listi *tncs* salvestatakse iga liigi kohta list, kus esimene element on selle liigi nimi, teine element on list lähedastest perekonnast ning kolmas element list nende perekonna sihtmärkliigile lähedastest liikidest. Igale liigile otsitakse 30 vastet.
2. Kasutajal palutakse otsida SRA andmebaasist kõikide vastete liike (OR seosega eraldatud) lisaparameetritega

   AND ("biomol dna"[Properties] AND "platform illumina"[Properties] AND "filetype fastq"[Properties])

   Kaustajal tuleb alla laadida fail SraRunInfo.csv, seejärel saab programm jätkata
3. Failist SraRunInfo.csv leidakse sekveneerimised, mis on mahult suurimad, kuid mitte üle 20GB. Need salvestatakse failina todownload.txt, kus on igal real komaga eraldatuna accession kood (võimaldab hiljem toorandmeid alla laadida) ja koodile vastav liigi nimi.

Sammul 2 on lisaparameetrid antud selleks, et SRA andmebaasi kirjel oleks juures FASTQ formaadis toorandmed – neid on vaja järgnevaks filtreerimiseks. Kuna antud tarkvara on optimeeritud Illumina lugemite andmetega töötama ja teisi platvorme ei ole katsetatud, välistatakse need otsingust.

Kolmanda sammu 20GB ülempiir on kehtestatud selleks, et suurendada programmi kiirust. Käesolevas uurimistöös on 50 sihtmärkliiki ning nende sihtmärkliikide lähedaste liikide ühend on *ca* 100 liiki. Ülempiirita toorandmete valimine ei oleks taolise koguse andmete juures mõeldav.

Järgneva programmi, *Dl-der*, pika tööaja tõttu on tegemist eraldiseisva programmiga. Tööpõhimõte on järgnev:

1. Loetakse faili todownload.txt rida-rea haaval.
2. Igal real käivitatakse accession koodi kasutades SRA toolkiti protseduur prefetch.

Kui programmi käivitada käsuga nohup, saab seejärel otsida nohup.out failist ebaõnnestunud protseduurid ning programmi uuesti nende SRA koodidega käivitada.

Seejärel tuleb kõikide alla laetud kaustade lahti pakkimiseks käsureal käivitada SRA toolkiti fastq-dump järgmiselt:

“fastq-dump --skip-technical -F --split-3 SRR*”

Ning uuesti lõpuga ERR*.

2.2.2.5 Programm *MitoFilter*

Sarnaselt programmile *MitoIdent*, kasutab ka programm *MitoFilter* alusena PlantTaxSeekerit ja ühendi, vahe ja *k*-meeride leidmiseks programmide komplekti GenomeTester4. Programm *filtering_with_nontargets* võimaldab filtreerimist vaid ühe *k*-meeride listi ja selle liigi lähedaste mitte-sitmärkide korral, seega on vaja programmi moditifseerida. Modifitseeritud programmi tööpõhimõte on järgnev:

1. Luuakse *k*-meeride listid kõikidest alla laetud sekveneerimise toorandmetest:
    1.1. Esiteks muudetakse FASTQ formaadis failid FASTA formaadis failideks, sest GenomeTester4.0 ei toeta SRA andmebaasi FASTQ formaadis failide analüüsi.
    1.2. Kui toorlugem on antud nii päri- kui ka vastassuunalisena, leitakse nende ühend.
2. Filtreeritakse ükshaaval sihtmärkliike:
    2.1. Tehakse kindlaks faili blasted_targets.txt info alusel, millised kuni 20% mitokondriaalse DNA kattuvusega liigid on alla laetud ning nimetatakse need lähedasteks liikideks.
    2.2. Leitakse lähedaste liikide *k*-meeride listide ühend, filtreeritakse välja vabalt valitud madala sagedusega *k*-meerid.
    2.3. Leitakse sihtmärkliikide ning lähedaste liikide listide ühendi vahe.

Programmi väljund on lõplikud listid sihtmärkliikide *k*-meeridest.

2.3 Tulemused

2.3.1 Allergiat põhjustavate liikide valik

Sihtmärkliigid valiti AllergenOnline andmebaasist, otsides täisliiginime märgisega “type: Food Animal”. Valimist eemaldati maismaaloomad, sest käesoleva uurimistöö eesmärgiks on leida liigispetsiifilised *k*-meerid limustele, kaladele ja koorikloomadele. Tabelis 1 on toodud kõik andmebaasist välja valitud tingimustele vastavad liigid.

Tabel 1. Töös käsitletud kalad, koorikloomad ja limused, mis andmebaasi AllergenOnline alusel põhjustavad allergiat. Tabelis on toodud 87 liiki oma ladinakeelsete nimedega, mis on tähestiku järjekorras. Liigid on käsitsi otsitud andmebaasist AllergenOnline.

Amphioctopus fangsiao	*Haliotis discus discus*	*Mimachlamys nobilis*	*Saccostrea glomerata*
Balanus rostratus	*Haliotis diversicolor*	*Neptunea polycostata*	*Salmo salar*
Batillus cornutus	*Haliotis laevigata x Haliotis rubra*	*Octopus vulgaris*	*Salvelinus fontinalis*
Bos grunniens mutus	*Helix aspersa*	*Ommastrephes bartramii*	*Sardinops sagax*
Charybdis feriatus	*Homarus americanus*	*Oncorhynchus keta*	*Scapharca broughtonii*
Chionoecetes opilio	*Lates calcarifer*	*Oncorhynchus mykiss*	*Scomber japonicus*
Clupea harengus	*Lepidorhombus whiffiagonis*	*Oratosquilla oratoria*	*Scomber scombrus*
Crangon crangon	*Litopenaeus vannamei*	*Oreochromis mossambicus*	*Scylla paramamosain*
Crassostrea gigas	*Macrobrachium rosenbergii*	*Pandalus borealis*	*Scylla serrata*
Crassostrea virginica	*Macruronus magellanicus*	*Panulirus stimpsoni*	*Sebastes marinus*
Cyprinus carpio	*Macruronus novaezelandiae*	*Paralithodes camtschaticus*	*Sepia esculenta*
Erimacrus isenbeckii	*Marsupenaeus japonicus*	*Penaeus monodon*	*Sepioteuthis lessoniana*
Eriocheir sinensis	*Melicertus latisulcatus*	*Perna viridis*	*Sinonovacula constricta*
Euphausia pacifica	*Merluccius australis australis*	*Pontastacus leptodactylus*	*Solen strictus*
Euphausia superba	*Merluccius bilinearis*	*Portunus pelagicus*	*Theragra chalcogramma*
Evynnis japonica	*Merluccius capensis*	*Portunus sanguinolentus*	*Thunnus albacares*
Farfantepenaeus aztecus	*Merluccius gayi*	*Portunus trituberculatus*	*Todarodes pacificus*
Fenneropenaeus chinensis	*Merluccius merluccius*	*Procambarus clarkii*	*Trachurus japonicus*
Fenneropenaeus merguiensis	*Merluccius paradoxus*	*Protortonia cacti*	*Tresus keenae*
Fulvia mutica	*Merluccius polli*	*Pseudocardium sachalinensis*	*Venerupis philippinarum*
Gadus callarias	*Merluccius productus*	*Quercus mongolica*	*Xiphias gladius*
Gadus morhua	*Metapenaeus ensis*	*Rastrelliger kanagurta*	

Tabelis toodud liike otsiti seejärel RefSeq mitokondriaalse DNA andmebaasist kasutades programmi *MitoDiff*. Programm jaotas andmebaasi sihtmärkliikideks, mis on toodud tabelis 2, ja mittesihtmärkliikideks.

Tabel 2. DNA RefSeq andmebaasis olemasolevad kalade, limuste ja koorikloomade liigid. Tabelis on toodud 50 liiki oma ladinakeelsete nimedega tähestiku järjekorras. Iga tabelis olev liik on leitud AllergenOnline andmebaasist ning on olemas DNA RefSeq andmebaasis.

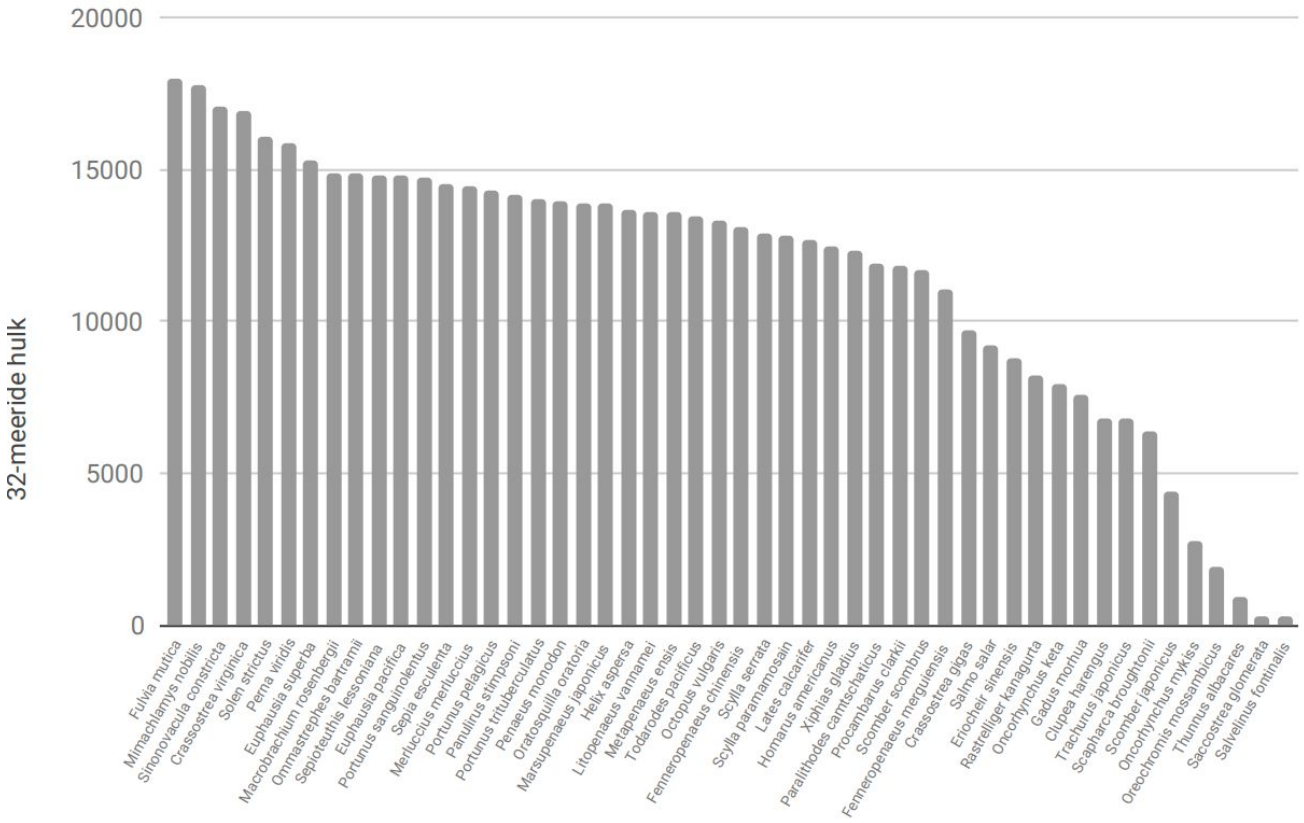
Clupea harengus	*Litopenaeus vannamei*	*Paralithodes camtschaticus*	*Scomber scombrus*
Crassostrea gigas	*Macrobrachium rosenbergii*	*Penaeus monodon*	*Scylla paramamosain*
Crassostrea virginica	*Marsupenaeus japonicus*	*Perna viridis*	*Scylla serrata*
Eriocheir sinensis	*Merluccius merluccius*	*Portunus pelagicus*	*Sepia esculenta*
Euphausia pacifica	*Metapenaeus ensis*	*Portunus sanguinolentus*	*Sepioteuthis lessoniana*
Euphausia superba	*Mimachlamys nobilis*	*Portunus trituberculatus*	*Sinonovacula constricta*
Fenneropenaeus chinensis	*Octopus vulgaris*	*Procambarus clarkii*	*Solen strictus*
Fenneropenaeus merguiensis	*Ommastrephes bartramii*	*Rastrelliger kanagurta*	*Thunnus albacares*
Fulvia mutica	*Oncorhynchus keta*	*Saccostrea glomerata*	*Todarodes pacificus*
Gadus morhua	*Oncorhynchus mykiss*	*Salmo salar*	*Trachurus japonicus*
Helix aspersa	*Oratosquilla oratoria*	*Salvelinus fontinalis*	*Xiphias gladius*
Homarus americanus	*Oreochromis mossambicus*	*Scapharca broughtonii*	
Lates calcarifer	*Panulirus stimpsoni*	*Scomber japonicus*	

2.3.2 Sihtmärkliikidele spetsiifiliste *k*-meeride leidmine

Kasutades programmi *MitoIdent*, tehti RefSeq andmebaasi mitokondri genoomidest *k*-meeride listid ning eemaldati need *k*-meerid, mis esinesid ka mittesihtmärkliikide RefSeq mitokondrites. Tulemused on toodud joonisel 6.

Edaspidi on kõikidel järgnevatel joonistel *x*-teljel toodud liigid samas järjekorras. *K*-meeri pikkuseks valiti maksimaalne GenomeTester4 poolt võimaldatav, s.t 32 aluspaari – protsessi käigus ei olnud põhjust aluspaaride hulga vähendamiseks ning GenomeTester4 töötab kõige mälu-efektiivsemalt just sellel aluspaaride hulgal.

Kõige enam 32-meere oli liigil *Fulvia mutica* (17965), kõige vähem liigil *Salvelinus fontinalis* (289) ja keskmiselt oli liikidel 11532 32-meeri. Sihtmärkliikide *k*-meeride hulga standardhälve oli 4615.5.

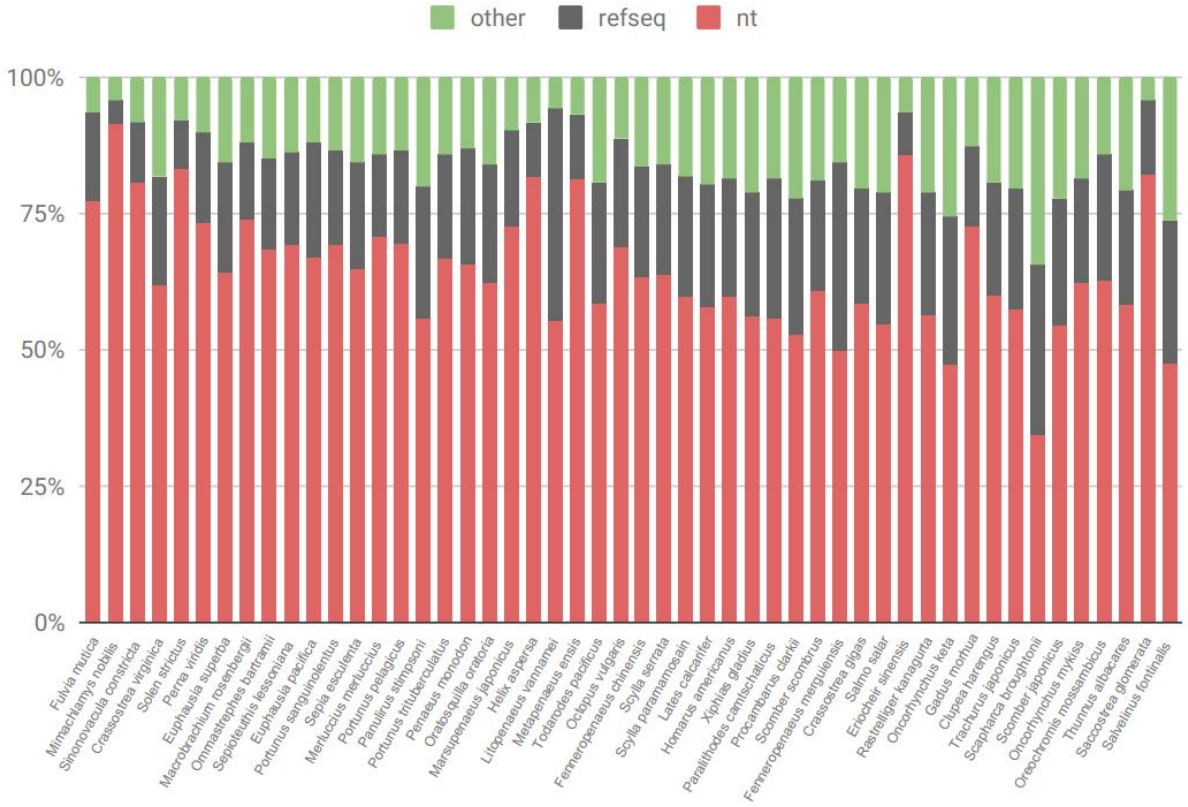


Joonis 6. Programmiga *MitoIdent* leitud liigispetsiifilised 32-meerid. Tulbale vastava liigi ladinakeelne nimi on toodud *x*-teljel; ja *y*-teljel on 32-meeride hulk, ehk 32-meeride komplektis leiduvate üksteisest erinevate 32-meeride arv.

2.3.3 Mittespetsiifiliste *k*-meeride väljafiltreerimine genoomsete järjestuste abil

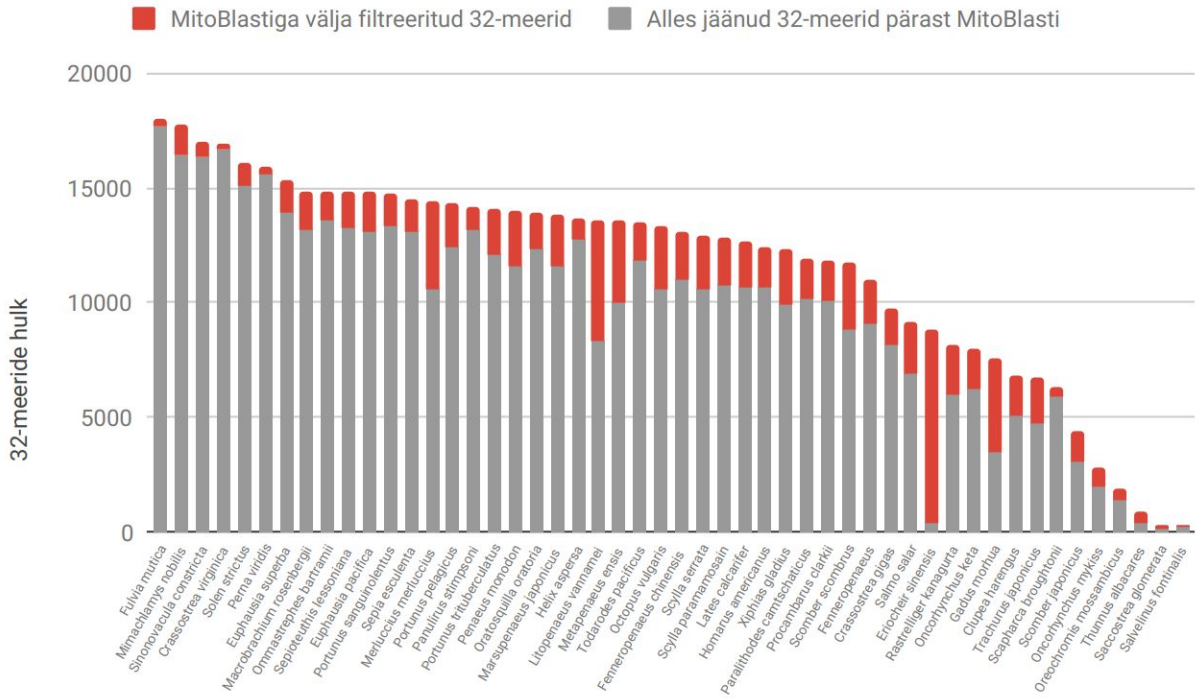
Järgmine filtreerimine viidi läbi kasutades programmi *MitoBlast*, mis filtreeris leitud 32-meeridest välja NCBI andmebaasides nt, refseq_genomic ja other_genomic teistele liikidele vasteid andvad 32-meerid. Joonisel 7 on toodud vasteid andnud 32-meeride päritolu andmebaaside kaupa.

Kõige enam (64%) vastetest leiti andmebaasist nt, seejärel (21%) andmebaasist refseq_genomic ja 15% vasteid andnud 32-meeridest andmebaasist other_genomic. 53% vasteid andnud 32-meeridest esinesid enam kui ühes eelnimetatud andmebaasidest. Joonisel 7 on iga andmebaasi loetelus arvestatud ka neid 32-meere, mis mitmes andmebaasis vasteid andsid.



Joonis 7. Vasteid andnud 32-meeride andmebaasiline päritolu. Andmed on toodud osakaaludena ning *x*-teljel on igale tulbale vastava liigi ladinakeelne nimi. Punasega on toodud andmebaasis nt vasteid andnud 32-meeride osakaal, halliga andmebaasi refseq_genomic vasteid andnud 32-meeride osakaal ja rohelisega andmebaasi other_genomic vasteid andnud 32-meeride osakaal. Kui mõni 32-meer andis vasteid mitmes andmebaasis, suurenevad nende andmebaaside osakaalud võrdselt. Osakaalude arvestamisel ei ole arvestatud vasteid andnud 32-meeride sagedustega sihtmärkliigi mitokondris ehk iga andmebaasi osakaalus on igat 32-meeri arvestatud üks kord.

Joonisel 8 on toodud punasega need 32-meerid, mis välja filtreeriti. Kõige enam (95%) 32-meere filtreeriti välja liigil *Eriocheir sinensis* ning kõige vähem (1.5%) liigil *Fulva Mutica*. Keskmiselt filtreeriti 20% 32-meeridest.

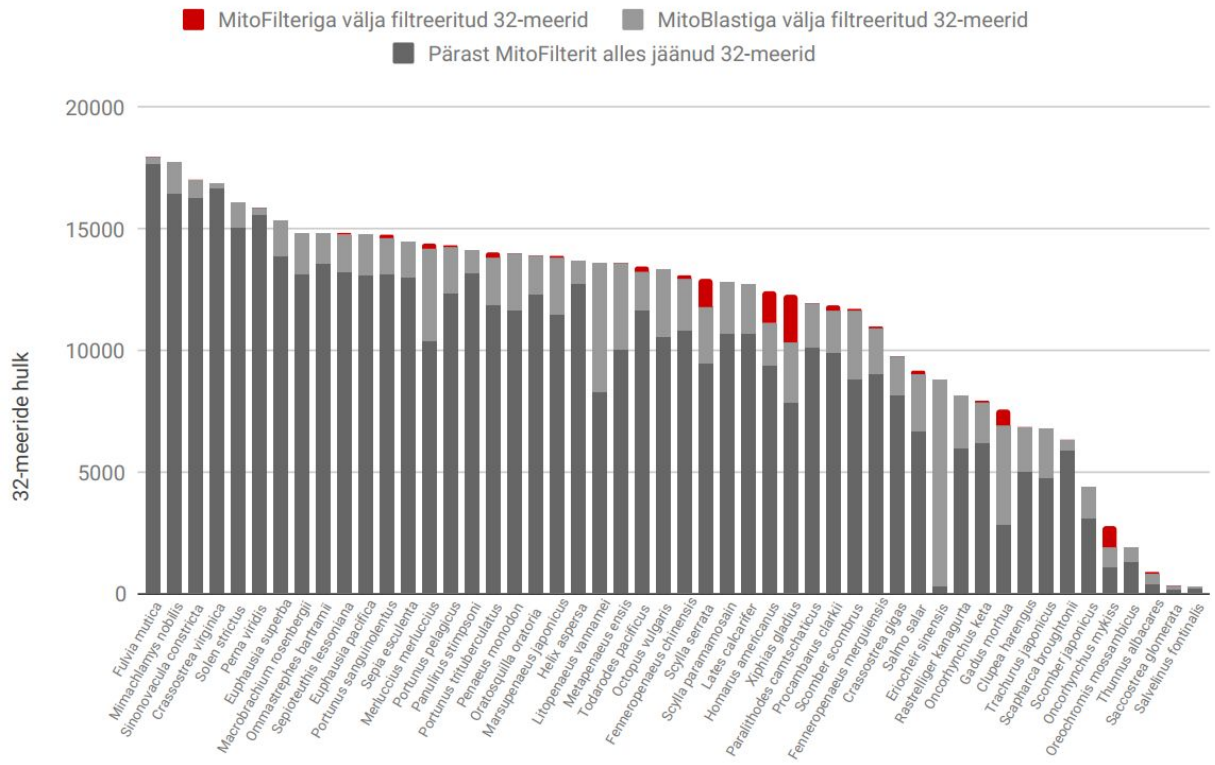


Joonis 8. Programmiga *MitoBlast* filtreerimise tulemus. Üksteisest erinevate 32-meeride hulk on toodud *y*-teljel; ja *x*-teljel on toodud sihtmärkliikide ladinakeelsed nimed. Punasega on toodud MitoBlastiga vasteid andnud 32-meerid ning halliga need 32-meerid, mis MitoBlasti filtreerimisega vasteid ei andnud.

2.3.4 Mittespetsiifiliste *k*-meeride väljafiltreerimine täisgenoomi toorlugemite abil

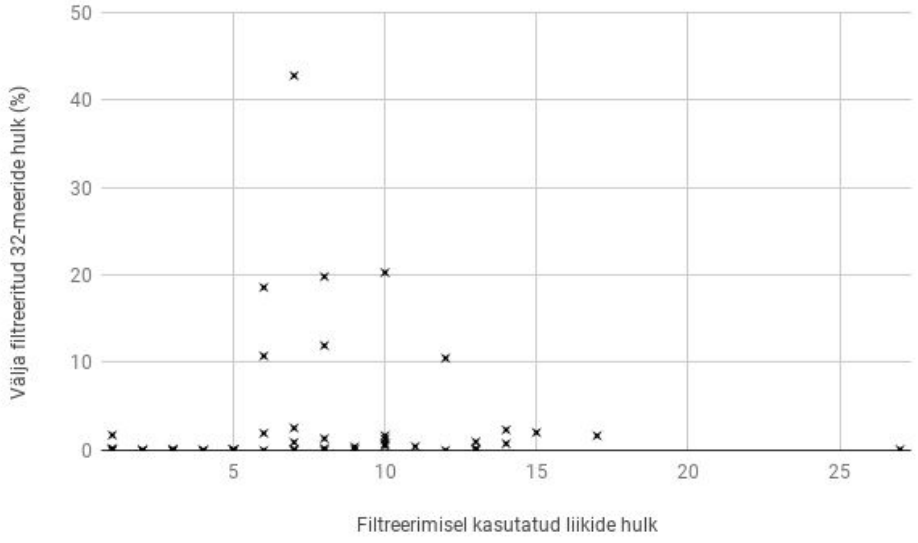
Täisgenoomi toorlugemite abil *k*-meeride filtreerimiseks oli esialgu vaja kindlaks teha sihtmärkliikide lähedased liigid, mis on saadaval SRA andmebaasis. Lisas 2 on toodud kõik filtreerimisel kasutatud sihtmärkliikidele lähedaste liikide nimed, mille määramine on kirjeldatud peatükis 2.2.2.5.

Joonisel 9 on toodud punasega täisgenoomi toorlugemitega filtreerimisel vasteid andnud 32-meerid. Kõige enam (30.5%) filtreeriti liigilt *Oncorhynchus mykiss* ning mitmetel liikidel ei filteeritud ühtegi 32-meeri. Keskmiselt filtreeriti välja 2% 32-meeridest. Pärast programmiga *MitoFilter* filtreerimist oli igal liigil keskmiselt 9526 32-meeri.



Joonis 9. Programmiga *MitoFilter* filtreerimise tulemus. Üksteisest erinevate 32-meeride hulk on toodud *y*-teljel; ja *x*-teljel on toodud sihtmärkliikide ladinakeelsed nimed. Punasega on toodud need 32-meerid, mis filtreeriti välja programmiga *MitoFilter*, helehalliga need 32-meerid, mis filtreeriti välja eelmises sammus, ehk programmiga *MitoBlast* ning tumehalliga need 32-meerid, mis on pärast programmiga *MitoFilter* filtreerimist alles.

Kui vaadelda, kas välja filtreeritud 32-meeride hulk on seoses filtreerimisel kasutatud 32-meeride hulgaga (joonis 10), märkame, et seos puudub.



Joonis 10. Programmiga *MitoFilter* välja filtreeritud 32-meeride hulk võrreldes filtreerimisel kasutatud liikide hulgaga. Iga *x*-tähis märgib ühe sihtmärkliigi andmepunkti, *x*-teljel on MitoFilteriga filtreerimisel kasutatud sihtmärkliigile lähedaste liikide hulk ja *y*-teljel programmiga *MitoFilter* välja filtreeritud 32-meeride osakaal kõikidest 32-meeridest.

2.3.5 Leitud *k*-meeride kontroll täisgenoomi toorlugemitel

Kontrollimaks, kas leitud 32-meerid tuvastavad liiki geneetilise materjali rohkusel, s.t ideaaltingimustes (*in vitro*), tehti kontroll nendel liikidel, mille täisgenoomi toorlugemid olid eelnevast etapist juba olemas. 32-meeride tuvastamise hindamiseks loodi skoor, mis on summa 32-meeride sagedusest toorlugemis (ülempiiriga 100) korrutatud selle sagedusega liigispetsiifiliste 32-meeride listis. Kuigi antud skoor ei ole objektiivne hinnang funktsionaalsusele toiduproovis, saab selle järgi hinnata, kas 32-meeride valik on hea ning hulk piisav signaali saamiseks. Tehti ka negatiivne kontroll ühe juhusliku liigiga, mis ei olnud eelneval filtreerimisel kasutusel. Tabelis 3 on vastavad tulemused.

Tabel 3. Pisteline kontroll 32-meeride funktsionaalsusest. Igal liigil, mille toorandmed olid eelneva sammu tõttu olemas, kontrolliti, kas 32-meerid annavad toorandmetel ning (mitte filtreerimisel kasutatud) negatiivsel kontrollil vasteid. Skoor peegeldab endas 32-meeride sagedusi toorandmetes (ülempiiriga 100), mis on korrutatud sagedustega sihtmärgi 32-meeride komplektis ning seejärel kokku liidetud. Võrdluseks on viimases veerus toodud üksteisest erinevate 32-meeride hulk sihtmärgil.

Liik	Skoor	Negatiivne kontroll	Skoor	sihtmärgi 32-meeride hulk
Penaeus monodon	1031889	Engaeus lyelli	0	11607
Portunus trituberculatus	1078062	Phalangium opilio	0	11861
Salmo salar	506224	Ruditapes philippinarum	0	6704
Oncorhynchus mykiss	27	Cherax holthuisi	0	1135
Eriocheir sinensis	21305	Cambaroides dauricus	0	332
Crassostrea gigas	749238	Engaeus quadrimanus	0	8174
Scylla paramamosain	1028823	Pomacea canaliculata	0	10728
Thunnus albacares	16712	Trachyrincus murrayi	0	348
Sinonovacula constricta	1228541	Upogebia bowerbankii	0	16323
Gadus morhua	210149	Drosophila yakuba	0	2851
Portunus pelagicus	119977	Sardina pilchardus	0	12355

2.3.6 Tööaeg ja arvutuslikud ressursid

Tabelis 4 on toodud töös kasutatud tegevuste kestvused. Märkame, et programmidel BLAST ja GenomeTester4 põhinevad sammud võtsid kokku 65.5 tundi, s.t ühe liigi kohta vaid 1.31 tundi; kuid kõige ajamahukam samm – toorandmete alla laadimine ja lahti pakkimine – mis

põhineb SRA toolkiti käskudel prefetch ja fastq-dump, võttis kõige kauem aega – ühe fastq faili alla laadimiseks ja lahti pakkimiseks kulus ligi 10 tundi.

Programmides käivitatakse üks BLAST või GenomeTester4 protsess korraga ning muu pythoni protsess ei ole väga arvutusmahukas, seega meetodi kasutatavad arvutuslikud ressursid on proportsionaalsed kasutatud programmide ressurssidega – BLAST käivitati kahekümnel lõimul, et see protsess oleks võimalikult kiire ning GenomeTester4 jaoks kasutati baas-sätteid.

Tabel 4. Kasutatud programmide kestvused. Programmide *MitoIdent*, *MitoBlast* ja *MitoFilter* kestvused on mõõtetud kasutades pythoni käsklust time.time(). Toorandmete alla laadimine ja lahti pakkimine ning programmide *MitoDiff* ja *CloseSpecFinder* kestvus on antud ligikaudu, sest see oli nende tegevuste korral liiga pikk või liiga lühike, et täpne mõõtmine oleks tarvilik.

Tegevus	**Kestvus**
MitoDiff	<1 minutit
MitoIdent	0.3 tundi
MitoBlast	28.1 tundi
CloseSpecFinder	<10 minutit
toorandmete alla laadimine ja lahti pakkimine	~1.5 kuud
MitoFilter	37.1 tundi

2.4 Arutelu

Käesoleva töö raames töötati välja meetod, millega saab leida *k*-meeride komplekti, mida oleks võimalik kasutada allergiat põhjustava liigi tuvastamisel sekveneerimise toorandmetest. *K*-meeridel põhinevaid meetodeid on varasemates töödes (Roosaare *et al.*, 2017; Wang *et al.*, 2020; Galata *et al.,* 2018) kasutatud bakteriliikide tuvastamiseks, kuid eukarüootidel tehtud töid on vähe. Artiklis (Raime ja Remm, 2018) leitakse liigile *Solanum lycopersicum* sarnase meetodiga *k*-meerid, kuid leitud 882 *k*-meeri on tunduvalt väiksem käesoleva töö keskmisest – 9526 32-meeri, mis võib viidata käesoleva meetodi suuremale tundlikkusele.

RefSeq mitokondriaalse DNA andmebaasi kasutades leiti keskmiselt 11532 esialgset 32-meeri, kuid liikidevaheline standardhälve on suur (joonis 6). Variatsiooni võivad seletada mitmed faktorid: mitokondri suurus (kuid sihtmärkliikidest vaid liigil *Scapharca broughtonii* on märkimisväärselt, s.t kaks korda, suurem mitokondri genoom, kui teistel liikidel, millest kõik on keskmisest *ca* 2000 bp suurusvahemikus) ja evolutsiooniline distants teistest liikidest (mida käesolevas töös ei uuritud). Varasemalt on näidatud (Spolsky ja Uzzell, 1984; Schiavo *et al.*, 2017), et mitokondriaalne DNA võib ka ühelt liigilt teisele täielikul kujul hüpata või mõne teise liigi tuuma DNAs esineda. *K*-meeride komplekti suuruste märkimisväärse variatsiooni põhjuste välja selgitamisega võiks tegeleda järgnevates töödes.

Vaadeldes BLAST-i filtreerimise tulemust, on märkimisväärne liigil *Eriocheir sinensis* välja filtreeritud 95% 32-meeridest. BLAST-i tulemusfailist (tulemused keskkonnas Github, 2020) on näha, et enamus 32-meeride vastetest on andnud liik *Eriocheir japonica sinensis,* mis on liigi *Eriocheir sinensis* alternatiivne nimetusviis, mida kasutatakse paralleelselt lühema ladinakeelse nimega (Tang *et al.*, 2020). Programm *MitoBlast* ei lugenud alternatiivset nime samale liigile kuuluvana, sest otsiti fraasi "*Eriocheir sinensis*"; kusjuures RefSeq mitokondriaalse DNA andmebaasis kasutatakse lühemat nime. Sarnaselt eelnevaga filtreeriti ka välja *Saccostrea glomerata k*-meere, mis olid andmebaasis nimega "*Saccostrea commercialis*".

Eelnevad näited toonitavad meetodi viimistlemise vajadust, s.t tuleks esialgsete liiginimede sisestamisel välja tuua ka populaarsed ladinakeelsed sünonüümid.

BLAST-i tulemusfailidest on näha (joonis 7), millistest andmebaasidest 32-meeride vasted pärit on. Andmebaasi other_genomic vastete osakaal kõige väiksem, sest tegemist on andmebaasiga, mis sisaldab vaid mikroorganisme ning enamus vastetest kattub ka RefSeq

andmebaasi omadega. RefSeq andmebaasi kasutamine on tarvilik, sest andmebaasis nt andmebaasi RefSeq andmed puuduvad. RefSeq andmebaasi vähene vastete arv on põhjendatav faktiga, et esialgsed 32-meerid leiti kasutades mitokondriaalset RefSeq andmebaasi, seega enamus vastetest, mis RefSeq andmebaas uuritavatele 32-meeridele annab, on genoomsetest järjestustest. Kõigi kolme andmebaasi rakendamine on siiski vajalik, et maksimaalne filtreerimine saavutada ning keskmine ajakulu liigi kohta (keskmiselt 34 minutit liigi kohta) ei ole piisavalt märkimisväärne, et see tingiks andmebaaside hulga vähendamist.

Analüüsides programmi *MitoFilter* tulemusi märkame, et kuigi protsess võttis keskmiselt 23 tundi sihtmärkliigi kohta, filtreeriti välja vaid keskmiselt 2% sammu *MitoBlast* läbinud 32-meeridest, mis ajamahukuse tõttu võib mitmele liigile korraga *k*-meeride leidmisel osutuda ebaefektiivseks. Kui filtreerimine lõpetada sammuga *MitoBlast*, on küll tõenäosus, et mõni lähedane liik annab antud *k*-meeride komplektiga signaali, suurem, kuid tänu *MitoFilter*-ile eelnevatel sammudel tehtud filtreerimistele võib eeldada, et sihtmärkliigi signaal on märkimisväärselt kõrgem teiste liikide poolt antud signaalidest. Ka fakt, et filtreerimisel kasutatud liikide arv ei muuda märkimisväärselt filtreerimise tulemust räägib kaasa väitele, et selle sammu võib ära jätta. Kui lõplikuks töövooks lugeda programmid *MitoDiff, MitoIdent* ja *MitoBlast,* kulub ühele liigile keskmiselt 36 minutit, mis on märkimisväärne arvestades, et kui *k*-meeride komplekt on olemas, ei tule protsessi enam korrata.

Leitud 32-meeride kontrolli tulemusi (tabel 3) vaadeldes on näha, et enamus leitud 32-meeride komplektidest annavad märkimisväärse signaali täisgenoomi sekveneerimise toorandmetega, kuid liigil *Oncorhynchus mykiss,* kuigi signaal on olemas, on see väike. Kuna antud sihtmärgi 32-meeride hulk on samas ka väike, võib eeldada, et signaal on proportsionaalne liigi 32-meeride hulgaga, kuid liigi *Eriocheir sinensis k*-meeride komplekti hulk on samuti väike (tänu liigisiseste vasteid saanud *k*-meeride eemaldamisele) ja signaal on siiski tugevam. Liigisiseseid *k*-meere antud liigil liigselt programmiga *MitoBlast* filtreerimise sammus ei eemaldatud. Kontrollimiseks kasutatud täisgenoomi toorlugem (ERR322054) oli samuti piisava pikkusega (12.8 Gbp). Signaali vähesust võib seletada toorlugemi saamiseks kasutatud metoodika või isendi ebasobivusega (mille kohta täiendavad andmed puuduvad) või käesoleva töö puudujääkidega.

Siiski 91% kontrollitud liikidest andis leitud *k*-meeride komplektiga signaali, seega töö eesmärgid võib lugeda saavutatuks.

Kui soovida laiendada töö mahtu kõigile 10111 RefSeq mitokondriaalse DNA andmebaasis olemasolevale liigile kasutades vaid ühte filtreerimist (programmiga *MitoBlast*), võtaks *k*-meeride leidmine aega umbes 8 kuud, kuid saaks luua tööriista, millega on võimalik tuvastada kõikvõimalikke liike, millel on olemas RefSeq mitokondrite andmebaasis viide.

KOKKUVÕTE

Käesoleva töö põhieesmärk oli töötada välja metoodika spetsiifiliste *k*-meeride leidmiseks sihtmärkliikide mitokondritest ja leida *k*-meerid, mis ka täisgenoomide puhul oleksid spetsiifilised.

Töö metoodikas pakuti välja töövoog, mis koosneb *k*-meeride leidmisest DNA RefSeq mitokondriaalse DNA andmebaasist ja kahest filtreerimise sammust. Filtreeritakse kasutades NCBI andmebaase (nt, refseq_genomic ja other_genomic) ja SRA toorlugemite andmebaasi.

Metoodikas välja pakutud töövooga leiti igale RefSeq andmebaasis olemasolevale sihtmärkliigile 32-meeride komplekt, mis koosnevad keskmiselt 9526 üksteisest erinevast 32-meerist. 91% leitud 32-meeridest andsid *in vitro* pistelises kontrollis liigi enda täisgenoomi toorlugemiga ka märkimisväärse signaali.

Leiti, et välja pakutud töövoo teise filtreerimise, ehk lähedaste liikide toorlugemitega filtreerimise sammu võib väga vähese spetsiifilisuse kaoga ka ära jätta, mis võimaldab keskmiselt 23 tundi kiiremalt *k*-meeride komplekti leida, olenevalt filtreerimisel kasutatud lähedaste liikide arvust ning toorlugemite mahust.

Eksperimentaalse osa tulemuseks on töövoog, millega on võimalik leida poolautomaatselt ja valdavalt vähese ajakuluga *k*-meeride komplekt liikidele, mille järjestused on olemas DNA mitokondriaalses RefSeq andmebaasis.

Finding specific *k*-mers to identify animal species which cause food allergy

Andrea Jõesaar

Summary

Today, food allergies are common and often very small amounts of the allergen is enough to cause an allergic reaction. Although in Estonia the contents of food, including allergen contents of larger doses, must be marked on the food container, it is necessary to develop sensitive methods to detect small amounts of an allergen from food.

For the detection of allergens, direct and indirect methods are in use. Direct methods detect the protein which causes the allergy, but they may fail if proteins are denatured or there is interference from other proteins. Indirect methods use the DNA of the allergy causing species to detect allergen content, as there if there are proteins of the species in the sample, there is also its DNA.

Indirect methods are divided into PCR-based methods and NGS-based methods. PCR-based methods, such as DNA barcoding, and minibarcoding (and in part, metabarcoding) use specific DNA primers to amplify a region of DNA, usually COI gene in the mitochondria, and sequence it using Sanger sequencing to differentiate between species. NGS-based methods, such as *k*-mer methods, derive the species from DNA raw sequencing reads.

In this thesis the program GenomeTester4 is used to count and sort *k*-mers to create *k*-mer lists, which are species-specific. GenomeTester4 is a *k*-mer counting program used mostly on small reads, which offers a high amount of functions in addition to the *k*-mer counting.

The purpose of this thesis was to develop a method, which can be used to find a set of species-specific *k*-mers which could be used to detect the species from an environment sample. In this thesis such 32-mers are found for 50 food allergy causing species.

The experimental part of this thesis consists of the identification of *k*-mers and two filtering steps. The filter is achieved by using BLAST on the initial *k*-mer sets, comparing the *k*-mers to the nt, refseq_genomic and other_genomic databases, and removing the *k*-mers that are also in the close species' raw sequencing reads. It was found that similar results could be achieved using only the first filtering step with a significant time remission.

The result of the thesis is a method which may be used to find *k*-mers for species which are in the mitochondrial RefSeq database relatively automatically and at a low time cost.

KIRJANDUSE LOETELU

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., ja Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410.

Alvarez, P. A., ja Boye, J. I. (2012). Food Production and Processing Considerations of Allergenic Food Ingredients: A Review. Journal of Allergy, 1–14. Academic Search Complete.

Spolsky, C, ja Uzzell, T. (1984). Natural Interspecies Transfer of Mitochondrial DNA in Amphibians. Proceedings of the National Academy of Sciences of the United States of America, 81(18), 5802. JSTOR Journals.

Deorowicz, S., Kokot, M., Grabowski, S., ja Debudaj-Grabysz, A. (2015). KMC 2: Fast and resource-frugal k-mer counting. Bioinformatics, 31(10), 1569–1576.

Eischeid, A. C. (2019). A method to detect allergenic fish, specifically cod and pollock, using quantitative real-time PCR and COI DNA barcoding sequences. Journal of the Science of Food and Agriculture, 99(5), 2641–2645.

Engvall, E., ja Perlmann, P. (1972). Enzyme-Linked Immunosorbent Assay, Elisa. The Journal of Immunology, 109(1), 129.

Fernandes, T. J. R., Costa, J., Oliveira, M. B. P. P., ja Mafra, I. (2015). An overview on fish and shellfish allergens and current methods of detection. Food & Agricultural Immunology, 26(6), 848–869. Academic Search Complete.

Galata, V., Backes, C., Laczny, C. C., ... Keller, A. (2018). Comparing genome versus proteome-based identification of clinical bacterial isolates. Briefings in Bioinformatics, 19(3), 495–505.

Goodman, R. E., Ebisawa, M., Ferreira, F., ... Taylor, S. L. (2016). AllergenOnline: A peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. Molecular Nutrition & Food Research, 60(5), 1183–1198.

Haarsma, A.-J., Siepel, H., ja Gravendeel, B. (2016). Added value of metabarcoding combined with microscopy for evolutionary studies of mammals. Zoologica Scripta, 45, 37–49.

Hadley, C. (2006). Food allergies on the rise? Determining the prevalence of food allergies, and how quickly it is increasing, is the first step in tackling the problem. EMBO Reports, 7(11), 1080–1083.

Hebert, P. D. N., Cywinska, A., Ball, S. L., ja deWaard, J. R. (2003). Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences, 270(1512), 313–321.

Horton, D. J., Kershner, M. W., ja Blackwood, C. B. (2017). Suitability of PCR primers for characterizing invertebrate communities from soil and leaf litter targeting metazoan 18S ribosomal or cytochrome oxidase I (COI) genes. European Journal of Soil Biology, 80, 43–48.

Kaplinski, L., Lepamets, M., ja Remm, M. (2015). GenomeTester4: A toolkit for performing basic set operations—Union, intersection and complement on k-mer lists. GigaScience, 4(s13742-015-0097-y).

Knowlton, N., ja Weigt, L. A. (1998). New dates and new rates for divergence across the Isthmus of Panama. Proceedings of the Royal Society of London. Series B: Biological Sciences, 265(1412), 2257–2263.

Chiu, K.-P. 2015. Next-Generation Sequencing and Sequence Data Analysis. 17-37. Bentham Science Publishers.

Marçais, G., ja Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 27(6), 764–770.

Meyer, R., Chebar Lozinsky, A., Fleischer, D. M., ... Venter, C. (2020). Diagnosis and management of Non-IgE gastrointestinal allergies in breastfed infants—An EAACI Position Paper. Allergy, 75(1), 14–32.

Monaci, L., ja Visconti, A. (2009). Mass spectrometry-based proteomics methods for analysis of food allergens. Applying combinations of chemical analysis and biological effects to environmental and food samples - I, 28(5), 581–591.

Pevzner, P. A., Tang, H., ja Waterman, M. S.. (2001). An Eulerian Path Approach to DNA Fragment Assembly. Proceedings of the National Academy of Sciences of the United States of America, 98(17), 9748. JSTOR Journals.

Raime, K., ja Remm, M. (2018). Method for the Identification of Taxon-Specific k-mers from Chloroplast Genome: A Case Study on Tomato Plant (Solanum lycopersicum). Frontiers in Plant Science, 9, 6.

Reier-Nilsen, T., Michelsen, M. M., Lødrup Carlsen, K. C., Carlsen, K. -H., Mowinckel, P., Nygaard, U. C., Namork, E., Borres, M. P., ja Håland, G. (2018). Predicting reactivity threshold in children with anaphylaxis to peanut. Clinical & Experimental Allergy, 48(4), 415–423.

Rizk, G., Lavenier, D., ja Chikhi, R. (2013). DSK: k-mer counting with very low memory usage. Bioinformatics, 29(5), 652–653.

Roosaare, M., Vaher, M., Kaplinski, L., Mols, M., Andreson, R., Lepamets, M., Koressaar, T., Naaber, P., Koljalg, S., ja Remm, M. (2017). StrainSeeker: Fast identification of bacterial strains from raw sequencing reads using user-provided guide trees.

Roy, R. S., Bhattacharya, D., ja Schliep, A. (2014). Turtle: Identifying frequent k-mers with cache-efficient algorithms. Bioinformatics, 30(14), 1950–1957.

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G., ja Reyes, A. (1999). Evolutionary genomics in Metazoa: The mitochondrial DNA as a model system. Gene, 238(1), 195–209.

Schiavo, G., Hoffmann, O. I., Ribani, A., Utzeri, V. J., Ghionda, M. C., Bertolini, F., Geraci, C., Bovo, S., ja Fontanesi, L. (2017). A genomic landscape of mitochondrial DNA insertions in the pig nuclear genome provides evolutionary signatures of interspecies admixture. DNA Research, 24(5), 487–498.

Shokralla, S., Hellberg, R. S., Handy, S. M., King, I., ja Hajibabaei, M. (2015). A DNA Mini-Barcoding System for Authentication of Processed Fish Products. Scientific Reports, 5(1), 15894.

Siragakis, G., ja Kizis, D. 2013. Food allergen testing: Molecular, immunochemical and chromatographic techniques. 4, 84-92.

Šípová, H., ja Homola, J. (2013). Surface plasmon resonance sensing of nucleic acids: A review. Analytica Chimica Acta, 773, 9–23.

Zhang, W., ja Zhao, X. (2013). Method for Rapid Protein Identification in a Large Database. BioMed Research International, 2013, 414069.

Tang, B., Wang, Z., Zhang, H., ... Li, Y. (2020). High-Quality Genome Assembly of Eriocheir japonica sinensis Reveals Its Unique Genome Evolution. Frontiers in Genetics, 10, 1340.

Wang, Y., Fu, L., Ren, J., Yu, Z., Chen, T., ja Sun, F. (2018). Identifying Group-Specific Sequences for Microbial Communities Using Long k-mer Sequence Signatures. Frontiers in Microbiology, 9, 872.

KASUTATUD VEEBIAADRESSID

AllergenOnline. (2019). Kasutatud 07.11.19, www.allergenonline.org

Bruker Surface Plasmon Resonance. (i.a). Kasutatud 17.04.2020, www.bruker.com/products/surface-plasmon-resonance

RefSeq mitokondriaalse DNA andmebaas andmebaas (i.a). Kasutatud 11.07.2019, ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/

NCBI nt, refseq_genomic ja other_genomic andmebaasid (i.a). Kasutatud 26.11.2019, ftp://ftp.ncbi.nlm.nih.gov/blast/db/

PlantTaxSeeker (2019). https://github.com/bioinfo-ut/PlantTaxSeeker/

Rockland Elisa Assays. (i.a). Kasutatud 14.04.2020, rockland-inc.com/elisa.aspx

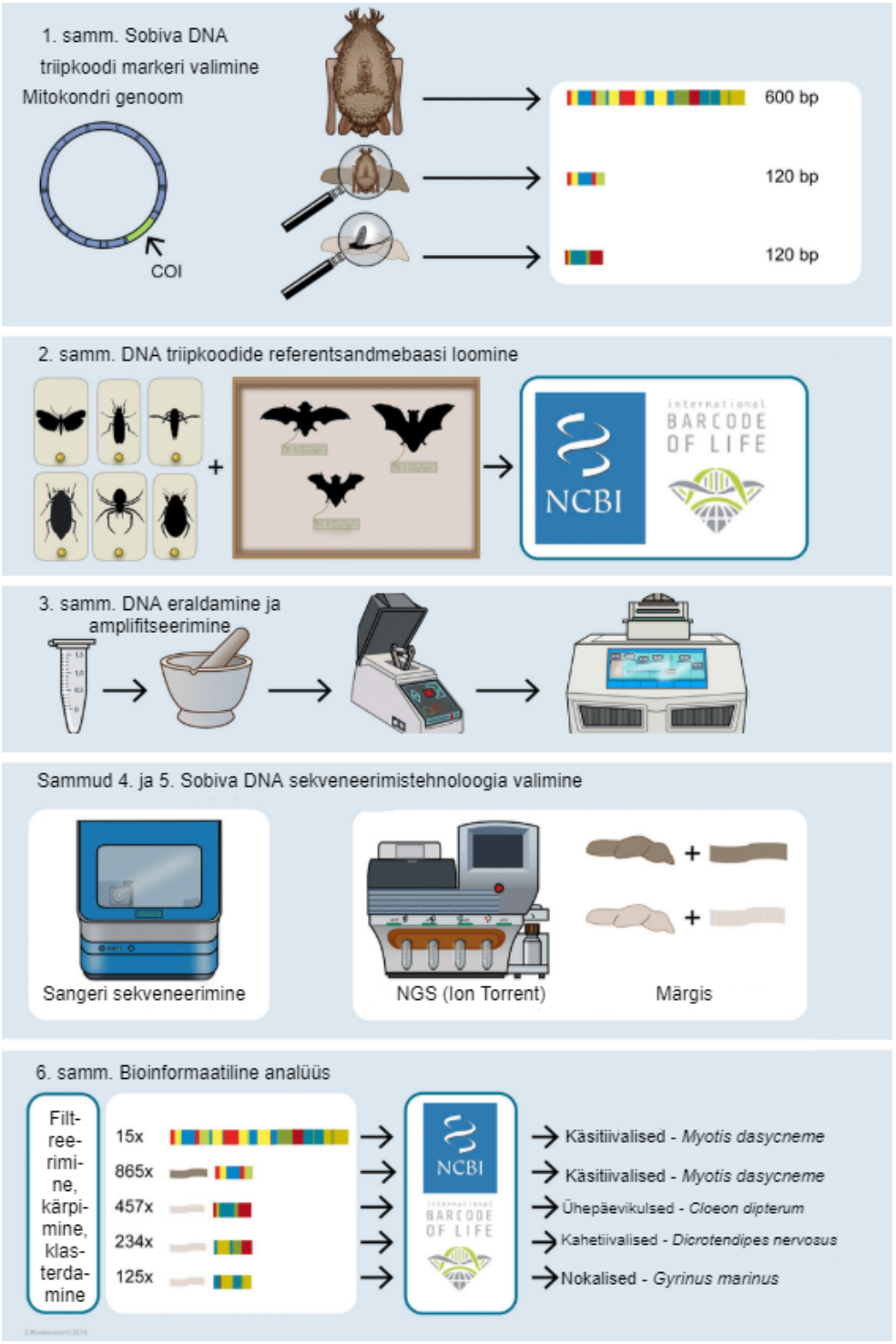
Tulemused keskkonnas GitHub (2020). https://github.com/B47895/MitoKmerSeeker

SRA andmebaas (i.a), Kasutatud 05.03.2020, https://www.ncbi.nlm.nih.gov/sra

VTA kodulehekülg. (2019). Kasutatud 02.12.19, vet.agri.ee/et/toit/allergeenid

LISAD

LISA 1



DNA triipkoodimise ja meta-triipkoodimise kuus põhilist sammu. DNA triipkoodimisel on tarvilik valida sobiv DNA marker, mis on eristatav erinevatel liikidel. Luues nendele markeritele andmebaasi, saab sama triipkoodi alusel erinevaid liike eristada. Pärast markeri PCR abil amplifitseerimist ja (üldjuhul Sangeri) sekveneerimist, saab sekveneeritud järjestuse abil kasutades eelnevalt loodud referentsandmebaasi eristada markeri abil liike üksteisest. (Kohandatud Haarsma *et al.*, 2016)

LISA 2

Sihtmärkliigid ning nendele lähedased liigid. Töös kasutati viimases filtreerimises sihtmärkliikidele lähedaste liikide sekveneerimise toorandmeid. Kasutatud lähedased liigid on toodud tabelis. Lähedased liigid leiti kasutades BLAST otsingut sihtmärkliikide mitokondritel otsides ülimalt 80% kokkusobimatute aladega liike ning valides 30 paremini sobivat liki ning otsides neid liike SRA andmebaasist. Olemasolevatest toorandmetest kasutati iga sihtmärkliigi kõiki ülimalt 80% kokkusobimatusega liike, mis loetakse ka allolevas tabelis sihtmärkideks.

Sihtmärkliik	**Lähedane liik**
Fulvia mutica	Erinaceus europaeus, Remiarctus bertholdii
Mimachlamys nobilis	Mizuhopecten yessoensis
Sinonovacula constricta	Ruditapes decussatus, Ruditapes philippinarum, Labritermes buttelreepeni, Clibanarius infraspinatus
Crassostrea virginica	Crassostrea ariakensis, Crassostrea gigas, Architeuthis dux, Engaeus quadrimanus, Engaeus sericatus, Penaeus monodon, Gebiacantha plantae, Engaeus lengana, Axianassa australis, Epixanthus frontalis, Austruca lactea, Labritermes buttelreepeni, Laurentaeglyphea neocaledonica, Pachygrapsus marmoratus, Gecarcoidea natalis, Portunus trituberculatus, Tubuca capricornis, Paranephrops planifrons, Portunus pelagicus, Gramastacus lacus, Cherax dispar, Upogebia bowerbankii, Tubuca polita, Gelasimus borealis, Remiarctus bertholdii, Tenuibranchiurus glypticus, Cherax destructor
Solen strictus	Ruditapes philippinarum, Drosophila yakuba, Drosophila simulans, Drosophila melanogaster, Cambaroides japonicus
Perna viridis	Pacifastacus leniusculus, Papilio protenor, Danaus chrysippus, Trypaea australiensis
Euphausia superba	Thalamita crenata, Cambaroides japonicus, Gelasimus borealis, Engaeus quadrimanus, Clibanarius infraspinatus, Upogebia bowerbankii, Cherax dispar, Pachygrapsus marmoratus, Euastacus spinifer, Engaeus sericatus, Paranephrops planifrons, Engaewa walpolea, Austruca lactea
Macrobrachium rosenbergii	Macrobrachium bullatum, Tenuibranchiurus glypticus, Thunnus thynnus, Thunnus albacares, Thunnus tonggol
Ommastrephes bartramii	Architeuthis dux, Upogebia bowerbankii, Remiarctus bertholdii, Engaeus lengana, Engaewa subcoerulea, Gebiacantha plantae, Ruditapes philippinarum
Sepioteuthis lessoniana	Architeuthis dux
Euphausia pacifica	Scylla paramamosain, Thalamita crenata, Drosophila melanogaster, Drosophila yakuba, Portunus trituberculatus, Gelasimus borealis, Gebiacantha plantae, Clibanarius infraspinatus, Munida isos
Portunus sanguinolentus	Portunus trituberculatus, Portunus pelagicus, Upogebia bowerbankii, Cardisoma carnifex, Cherax dispar, Cherax destructor, Euastacus spinifer, Tenuibranchiurus glypticus, Engaewa subcoerulea, Gelasimus borealis, Gecarcoidea natalis, Austruca lactea, Remiarctus bertholdii
Sepia esculenta	Architeuthis dux
Merluccius merluccius	Lota lota, Gadus chalcogrammus, Gadus morhua, Arctogadus glacialis, Melanogrammus aeglefinus, Trachyrincus murrayi
Portunus pelagicus	Portunus trituberculatus, Thalamita crenata, Penaeus monodon, Cherax dispar, Cherax holthuisi, Cherax preissii, Tenuibranchiurus glypticus, Pachygrapsus marmoratus, Epixanthus

	frontalis, Macrobrachium bullatum, Paranephrops planifrons, Gramastacus lacus, Remiarctus bertholdii, Upogebia bowerbankii
Panulirus stimpsoni	Thenus orientalis, Tenuibranchiurus glypticus, Engaeus lyelli, Sinonovacula constricta, Axianassa australis
Portunus trituberculatus	Portunus pelagicus, Thalamita crenata, Scylla paramamosain, Upogebia bowerbankii, Cherax dispar, Cherax holthuisi, Engaeus sericatus, Cherax preissii, Cherax destructor, Gebiacantha plantae, Tenuibranchiurus glypticus, Cherax tenuimanus, Pachygrapsus marmoratus, Austruca lactea, Clibanarius infraspinatus, Engaeus lengana, Remiarctus bertholdii
Penaeus monodon	Upogebia bowerbankii, Paranephrops planifrons, Trypaea australiensis, Gelasimus borealis, Portunus pelagicus, Gecarcoidea natalis
Oratosquilla oratoria	Gecarcoidea natalis, Austruca lactea, Cherax tenuimanus
Marsupenaeus japonicus	Penaeus monodon, Axianassa australis, Paranephrops planifrons, Birgus latro, Euastacus spinifer, Gebiacantha plantae, Cherax destructor, Clibanarius infraspinatus, Gelasimus borealis, Gecarcoidea natalis, Bactrocera oleae
Helix aspersa	Pomacea maculata, Pomacea canaliculata
Litopenaeus vannamei	Penaeus monodon, Paranephrops planifrons, Gebiacantha plantae, Upogebia bowerbankii, Clibanarius infraspinatus, Thenus orientalis, Pachygrapsus marmoratus, Austruca lactea
Metapenaeus ensis	Gramastacus lacus, Paranephrops planifrons, Upogebia bowerbankii, Euastacus spinifer, Tenuibranchiurus glypticus, Cherax dispar, Cherax preissii, Cherax holthuisi, Munida isos, Remiarctus bertholdii, Engaeus sericatus, Austruca lactea
Todarodes pacificus	Architeuthis dux
Octopus vulgaris	Fundulopanchax gardneri, Littorina saxatilis, Architeuthis dux
Fenneropenaeus chinensis	Penaeus monodon, Upogebia bowerbankii, Paranephrops planifrons, Thalamita crenata, Clibanarius infraspinatus, Tetranychus urticae, Cardisoma carnifex, Gecarcoidea natalis, Austruca lactea, Gebiacantha plantae
Scylla serrata	Scylla paramamosain, Thalamita crenata, Portunus trituberculatus, Trypaea australiensis, Austruca lactea, Eriocheir sinensis
Scylla paramamosain	Thalamita crenata, Portunus trituberculatus, Gecarcoidea natalis, Upogebia bowerbankii, Homarus gammarus, Remiarctus bertholdii, Eriocheir sinensis
Lates calcarifer	Sarotherodon linnellii, Heniochus diphreutes
Homarus americanus	Homarus gammarus, Cherax holthuisi, Paranephrops planifrons, Cherax preissii, Birgus latro, Axianassa australis, Remiarctus bertholdii, Upogebia bowerbankii
Xiphias gladius	Caesio cuning, Thunnus tonggol, Gymnosarda unicolor, Thunnus albacares, Thunnus thynnus, Katsuwonus pelamis, Euthynnus affinis, Trachinotus ovatus, Euthynnus alletteratus, Auxis rochei
Paralithodes camtschaticus	Laurentaeglyphea neocaledonica, Pacifastacus leniusculus, Upogebia bowerbankii, Munida isos, Drosophila simulans, Austruca lactea, Gecarcoidea natalis, Cardisoma carnifex
Procambarus clarkii	Pacifastacus leniusculus, Cambaroides japonicus, Cambaroides dauricus, Paranephrops planifrons, Cherax preissii, Engaewa walpolea, Cherax dispar, Upogebia bowerbankii, Engaeus quadrimanus, Engaeus sericatus, Tenuibranchiurus glypticus, Remiarctus bertholdii, Birgus latro, Gramastacus lacus, Cherax tenuimanus
Scomber scombrus	Thunnus albacares, Thunnus tonggol, Katsuwonus pelamis, Thunnus thynnus, Euthynnus affinis, Gymnosarda unicolor, Euthynnus alletteratus, Auxis rochei, Caesio cuning, Oreochromis niloticus
Fenneropenaeus merguiensis	Penaeus monodon, Remiarctus bertholdii, Upogebia bowerbankii, Clibanarius infraspinatus, Engaeus lengana, Gelasimus borealis, Austruca lactea

Crassostrea gigas	Crassostrea ariakensis, Eriocheir sinensis, Strahlaxius plectrorhynchus, Ibacus alticrenatus, Munida isos
Salmo salar	Salmo trutta, Oncorhynchus mykiss, Salvelinus namaycush, Thymallus thymallus, Hiodon alosoides, Protosalanx chinensis, Trachinotus ovatus
Eriocheir sinensis	Cardisoma carnifex, Gecarcoidea natalis, Pachygrapsus marmoratus, Gelasimus borealis, Austruca lactea, Tubuca polita, Tubuca capricornis, Paranephrops planifrons, Scylla paramamosain, Engaeus quadrimanus, Epixanthus frontalis, Thalamita crenata
Rastrelliger kanagurta	Katsuwonus pelamis, Euthynnus alletteratus, Auxis rochei, Thunnus thynnus, Thunnus albacares, Euthynnus affinis, Thunnus tonggol, Gymnosarda unicolor, Caesio cuning, Salmo trutta
Oncorhynchus keta	Oncorhynchus mykiss, Salmo trutta, Salvelinus namaycush, Salmo salar, Thymallus thymallus, Protosalanx chinensis, Caesio cuning, Thunnus thynnus
Gadus morhua	Gadus chalcogrammus, Arctogadus glacialis, Melanogrammus aeglefinus, Lota lota, Trachyrincus murrayi, Gymnosarda unicolor
Clupea harengus	Sardina pilchardus, Hiodon alosoides, Gymnosarda unicolor
Trachurus japonicus	Trachinotus ovatus, Caesio cuning, Thunnus albacares, Auxis rochei, Thunnus tonggol, Thunnus thynnus, Katsuwonus pelamis, Gymnosarda unicolor, Euthynnus alletteratus, Euthynnus affinis, Sarotherodon lohbergeri, Sarotherodon linnellii, Stomatepia pindu
Scapharca broughtonii	Phalangium opilio, Apostictopterus fuliginosus, Bactrocera oleae, Mustela erminea
Scomber japonicus	Katsuwonus pelamis, Thunnus tonggol, Thunnus thynnus, Thunnus albacares, Auxis rochei, Euthynnus affinis, Gymnosarda unicolor, Euthynnus alletteratus, Caesio cuning
Oncorhynchus mykiss	Salmo salar, Salmo trutta, Salvelinus namaycush, Thymallus thymallus, Protosalanx chinensis, Thunnus thynnus, Euthynnus alletteratus
Oreochromis mossambicus	Oreochromis niloticus, Sarotherodon lohbergeri, Sarotherodon linnellii, Stomatepia pindu, Caesio cuning, Katsuwonus pelamis, Auxis rochei, Gymnosarda unicolor, Thunnus thynnus, Euthynnus affinis, Thunnus albacares, Thunnus tonggol, Euthynnus alletteratus, Heniochus diphreutes
Thunnus albacares	Thunnus tonggol, Thunnus thynnus, Katsuwonus pelamis, Euthynnus affinis, Euthynnus alletteratus, Auxis rochei, Gymnosarda unicolor, Caesio cuning
Saccostrea glomerata	Crassostrea ariakensis, Crassostrea gigas, Thenus orientalis, Ibacus alticrenatus, Munida isos
Salvelinus fontinalis	Salvelinus namaycush, Salmo trutta, Salmo salar, Oncorhynchus mykiss, Thymallus thymallus, Hiodon alosoides, Protosalanx chinensis, Thunnus thynnus, Euthynnus alletteratus, Katsuwonus pelamis

LIHTLITSENTS

Mina, Andrea Jõesaar,

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

   Spetsiifiliste *k*-meeride leidmine inimesel toiduallergiat põhjustavate loomaliikide määramiseks,

   mille juhendaja on Reidar Andreson,

   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 3.0, 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.

Andrea Jõesaar

08.06.2020