TARTU ÜLIKOOL

LOODUS- JA TEHNOLOOGIATEADUSKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

Heleri Kirsip

Viirustelt kärbselistele ülekandunud geeni, *TMV-CP*, integratsiooniaja ning funktsiooni uurimine

Magistritöö

Juhendaja vanemteadur Aare Abroi, PhD

TARTU 2015

SISUKORD

KASUTATUD LÜHENDID

AIC – *Akaike information criterion*

BDV - *Borna disease virus*

CP – kattevalk (*coat protein*)

EBLN – endogeensed bornaviiruste sarnased nukleoproteiinid (*endogenous bornavirus-like nucleocapsid*)

ERV – endogeensed retroviirused (*endogenous retroviruses*)

EVE – endogeensed viiruslikud elemendid (*endogenous viral elements*)

FISH – fluorestsents *in situ* hübridisatsioon (*fluorescence in situ hybridization*)

FM – FlyMine andmebaas

GISH – genoomne *in situ* hübridisatsioon (*genomic in situ hybridization*)

GRD – geminiviiruse sarnased järjestused (*geminivirus related DNA*)

GV – Genevestigator andmebaas

HGT – horisontaalne geeniülekanne (*horizontal gene tranfer*)

HMM – peidetud Markovi mudel (*hidden Markov models*)

MA – miljonit aastat (*milion years*)

MAT – miljonit aastat tagasi (*milion years ago*)

ML – suurima tõepära meetod (*Maximum Likelihood*)

MSA – mitmese järjestuse joondus (*multiple sequence alignment*)

NCBI – *National Center for Biotechnology Information*

NJ – distantsipõhine fülogeneetiliste puude koostamise meetod (*Neighbor joining*)

NRVS – mitte-retroviiruslike RNA viiruste sarnased järjestused (*nonretroviral RNA virus-like sequences*)

PDB – bioloogiliste makromolekulide struktuuride andmebaas (*RCSB Protein Data Bank*)

proto-ORF – proto-avatud lugemisraam (*proto-open readingframe*)

RMS – ruutkeskmine (*root mean square*)

RNP – ribonukleiinhappe-valk kompleks (*ribonucleoprotein*)

RT-PCR – pöördtranskriptsiooni PCR (*reverse transcription-PCR*)

SCOP – valkude struktuuride klassifikatsiooni andmebaas (*Structural Classification of Proteins*)

TMV-CP – tubaka mosaiigiviiruse kattevalk (*tobacco mosaic virus coat protein*)

TSS – transkriptsiooni alguspunkt (*transcription start site*)

SISSEJUHATUS

Uusi geene võib eukarüootsetes organismides tekkida mitmel viisil, kuid käesolevas töös keskendutakse horisontaalsele geeniülekandele (HGT; ing k *horizontal gene tranfer*). HGT on põhjalikult uuritud bakteritel, palju näiteid on teada taimedel, kuid ainult läbi iduree paljunevatel eukarüootidel esineb antud protsess üliharva. HGT üheks näiteks võivad olla endogeensed viiruslikud elemendid (EVE; ing k *endogenous viral elements*) – viiruste genoomi osad, mis on integreerunud peremeesorganismi genoomi ning seal fikseerunud. Antud protsessi on põhjalikult uuritud retroviiruste baasil, kuid ülejäänud viiruste jaoks on seda kuni viimase ajani peetud võimatuks protsessiks. Järjest rohkem on aga leitud mitte-retroviiruslike EVE´de näiteid erinevate organismide genoomidest.

Viiruslike järjestuste detekteerimises esineb probleem just väikeses sarnasuses, kuna viiruslikud järjestused muteeruvad tunduvamalt kiiremini kui eukarüootsed, seega kahe järjestuse võrdluste alusel määratavad EVE´d on limiteeritud. Teades, et valgujärjestus, ja eriti struktuur, on ajas tunduvamalt konserveerunum kui nukleiinhape, püütakse EVE´de detekteerimisvõimet parandada, kasutades selleks näiteks peidetud Markovi mudeleid (HMM; ing k *hidden Markov models*), mis on loodud struktuursete valgudomeenide superperekondadesse klassifitseerimise alusel, näiteks SUPERFAMILY andmebaas (Gough, 2002).

Käesolev töö on jätkuks Kirsip (2013) bakaulaureusetööle, kus avastati ning uuriti lähemalt viiruste ning eukarüootsete organismide jagatud valgudomeeni, mille puhul määrati ülekande suunaks viirustelt peremeesorganismidele. Täpsemalt arvati, et tubaka mosaiigiviiruse kattevalk (TMV-CP; ing k *tobacco mosaic virus coat protein*) on integreerunud kärbseliste eellasesse 60-250 milionit aastat tagasi (MAT; ing k *million years ago*).

Käesoleva töö peamiseks eesmärgiks on täpsustada ülekande aega ning uurida kärbseliste TMV-CP võimalikku funktsiooni, kasutades selleks andmekaevet. Lisaks vaadatakse TMV-CP_{fly} detekteerimisvõimet erinevate meetoditega, püüdes leida lisaks üldlevinud BLAST analüüsile paremaid alternatiive.

Kirjanduse osas antakse ülevaade uute geenide tekke võimalustest, keskendudes horisontaalsele geeniülekandele. Lisaks näidatakse, et paljudel uutel tekkinud geenidel esineb tunduvamalt kõrgem ekspressioon testistes, kui teistes kudedes. Samuti antakse lühiülevaade varem avastatud EVE´dest, nende levinumatest deteketeerimismeetoditest ning funktsiooni uurimisest. Teiseks suureks teemaks kirjanduse osas on struktuuri kaasamine bioinformaatilistesse analüüsidesse, täpsemalt kirjeldatud meetodite kaasamine varem esinenud

probleemide lahendamiseks. Lisaks antakse ülevaade ka kärbseliste fülogeneesist, tuues välja suuremad lahkenemised kärbseliste gruppide baasil ning nende sündmuste võimalikud toimumiste ajad. See peaks aitama täpsemalt hinnata uuritava TMV-CP_{fly} integratsiooniaega.

Eksperimentaalses osas kasutatakse erinevaid EVE´de deketeerimismeetodeid, uurimaks nende rakendatavust ning tööefektiivsust. Teiseks proovitakse määrata uuritava TMV-CP funktsiooni kärbselistes. Esmalt uuritakse, kas antud valk võiks olla säilitanud oma algse viirusliku kattevalgu funktsiooni või on kärbselistes suutnud omastada uue ülesande. Teise võimaluse puhul püütakse funktsiooni kirjeldada olemasolevate andmete abil: ekspressioonianalüüside, transkriptide detekteerimistel organismides ning interaktsiooniandmete abil.

Käesolev töö on valminud Tartu Ülikooli Molekulaar- ja Rakubioloogia Instituudis, eksperimentaalne uurimistöö on tehtud Tartu Ülikooli Tehnoloogiainstituudis. Autor soovib tänada oma juhendajat, Aare Abroid abi ning nõuannete eest.

Märksõnad: EVE, tubaka mosaiigiviirus, valgudomeen, struktuur, fülogenees, horisontaalne geeniülekanne

1. KIRJANDUSE ÜLEVAADE

1.1. UUED GEENID

Uute geenide ning funktsioonide teket on peetud tähtsaks adaptiivse evolutsiooni osaks. Pikemat aega peeti nii kromosoomide, geenipiirkondade kui ka üksikute geenide duplikatsiooni ning neo- või subfunktsionaliseerimist peamisteks uute geenide tekkeallikaks. Lisaks kuuluvad selle alla ka retrogeenid – läbi RNA pöördtranskriptsioon tekkivad duplikatsioonid, millel üldjuhul puuduvad intronid.

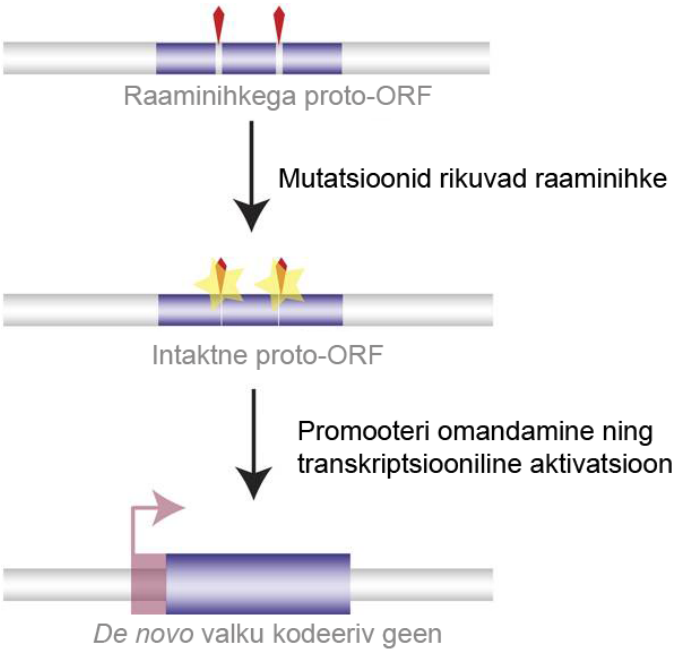
De novo päritolu **valku kodeeriva geeni** teket on pikka aega peetud suhteliselt võimatuks sündmuseks. Jacob (1977) avaldas arvamust, et juhuslike aminohapete assotsieerumisel mittekodeerivas regioonis, on uue funktsionaalse valgu tekke võimalus nulli lähedase tõenäosusega. Samas on viimasel ajal järjest rohkem vastupidist tõestatud. Näiteks *morpheus´*e geen Vana-Maailma primaatide esivanemal (Johnson *et al*., 2001). Johnson *et al*. (2001) avastasid inimeste genoomides 15 duplitseerunud segmenti, mis sisaldasid *morpheuse* geeniperekonna geene. Erinevates primaatide piires teostatud valkude võrdlusel leiti geeniduplikatsiooni sündmus, millele järgnes kodeerivas regioonis suuremamahulised aminohapete muutused. Samas pole uuritud täpsemalt algse järjestuse teket, kuid selle puudumine teistes organsmides viitab *de novo* tekkele.

1.1.1. De novo geenide tekkemehhanismid

De novo geenitekkeks peetakse peamiselt mittekodeerivast regioonist mutatsioonide teel tekkinud geeni koos ekspresseerumiseks vajalike regulaatorelementidega. Tihti loetakse selle alla kuuluvat ka horisontaalsel geeniülekandel (HGT) omandatud uued geenid, seal hulgas parasiitorganismide DNA integratsioon peremeesorganismi.

1.1.1.1. Geenide teke mittekodeerivast regioonist

De novo geenide teke mittekodeerivast regioonist on mitmeetapiline protsess (joonis 1). Esmalt peab vastav regioon omama proto-avatud lugemisraami (proto-ORF) mittekodeerivas regioonis ning järgnevalt on vajalik mutatsioonide teke (insertsioonid, deletsioonid, üksikute nukleotiidide asendused), mis muudaksid avatud lugemisraami segavaid stoppkoodoneid (Kaessmann, 2010). Regulaatorelementide omandamised, mis aitaksid järjestusel saavutada ekspresseeruvat funktsiooni, on viimane vajalik etapp, võimaldades tekitada uut funktsionaalset transkribeeritavat valku (Kaessmann, 2010).



Joonis 1. Valku-kodeerivate geenide tekkimine mittekodeerivast regioonist. Esiteks peab proto-avatud lugemisraami (proto-ORF; sinised kastikesed) tekkima mutatsioone (kollased tähekesed), mis eemaldavad avatud lugemisraami rikkuvaid nukleotiide (punased). Transkriptsiooniliselt aktiivse ORF´i tekkimiseks peab toimuma promootri omandamine (transkriptsiooni alguspunkt (TSS; ing k *transcription start site*) – roosa nool), mille tagajärjel võib toimuda uue *de novo* tekkinud geeni transkriptsioon (funktsionaalne ekson – sinine kastikene). Joonis on koostatud Kaessmann (2010) poolt ning tõlgitud eesti keelde töö autori poolt.

Transkriptsiooni võimaldamiseks on oluline tekkinud geeni asukoht ning selle omadused: genoomi regiooni avatus, transkriptsiooniline aktiivsus ning võime omandada vajalikke regulaatorelemente (Kaessmann, 2010). Neid omadusi võivad soodustada CpG-rikkad järjestused, (Kaessmann *et al*., 2009; Fablet *et al*., 2009), ülavoolu asuvad regulatoorsed elemendid (Fablet *et al*., 2009; Zaiss ja Kloetzel, 1999) ja *de novo* mutatsiooniliste asenduste tagajärjel tekkinud muudatused (Bai *et al*., 2007; Betran ja Long, 2003).

1.1.1.2. Horisontaalne geeniülekanne.

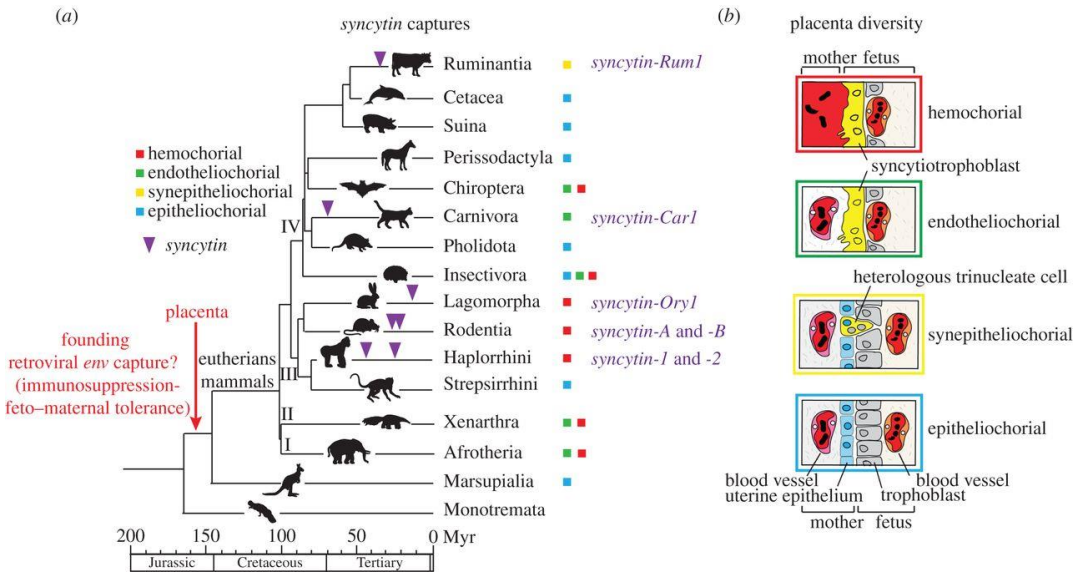
Horisontaalne geeniülekanne (HGT; ing k *horizontal gene tranfer*) on protsess, kus ühe liigi geneetiline materjal kandub üle teise organismi. Bakterite vahelist HGT on põhjalikult uuritud (Boucher *et al*., 2003), lisaks on leitud tõendeid ka üherakuliste eukarüootsete organismide (Keeling ja Palmer, 2008) ning taimede puhul (Rice *et al*., 2013). Näiteks õistaime Amborella mitokondri genoomi on integreerunud rohevetikate, sammalde ning teiste õistaimede mitokondrite geene ja täisgenoome (Rice *et al*., 2013). Kõrgemate eukarüootsete organismide puhul on HGT´st leitud tõendeid vaid endosümbiontide (mitokondri ning plastiidide geenide

ülekanne tuumagenoomi; de Souza ja Motta, 1999; Hoffmeister ja Martin, 2003) ning parasiitorganismide (Wolbachia bakteri geenide/genoomi ülekanne puuviljakärbeste, sääskede, herilaste ning nematoodide genoomidesse; Hotopp *et al*., 2007) puhul.

Üldiselt arvatakse, et loomadel esineva HGT puudumine on põhjustatud segregeerunud ning kaitstud sugurakkude poolt (Keeling ja Palmer, 2008). Gladyshev *et al*. (2008) näitas aga väikestel närilistel ja värskevee keriloomadel tuumagenoomi integreerunud bakterite, seente ja taimede võõr-DNA´d, mis on seotud keskkonna stressivastusega. Täpsemat funktsionaalset seost tuleks aga põhjalikumalt uurida.

1.1.1.3. Parasiitorganismidest pärit geenide ülekanne peremeesorganismi

Genoomi parasiitelementideks peetakse tavaliselt transposoone ning endogeenseid retroviiruseid, mis on integreerunud organismide genoomi, aja jooksul fikseerunud ning on päritavad Mendeli seaduste järgi (Lavialle *et al*., 2013). Neid peetakse peamisteks uute geenide tekkeallikaks, võimaldades geeniduplikatsioonide abil tekitada uusi funktsioone (Kaessmann, 2010).

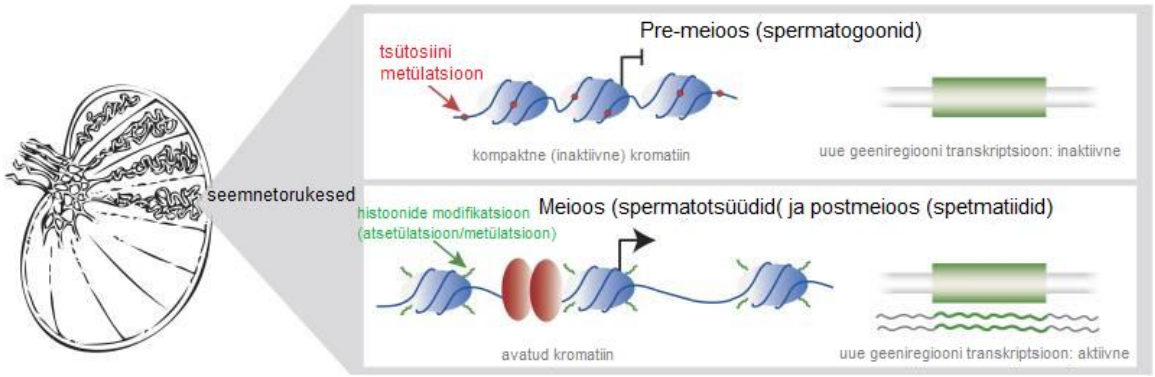


Joonis 2. *Syncytin* geen imetajates ning nende platsentade omapärad. (A) Imetajate fülogeneetilise puu, *Eutheria* nelja suurima klaadi esindajatega: (I) Afrotheria, (II) Xenarthra, (III) Euarchontoglires ja (IV) Laurasiathria (baseerub Meredith *et al*. (2011) tööl). Värvilised ruudukesed märgivad nelja platsentatüübi esinemist antud gruppides. Lilla kolmnurgaga on märgitud *syncytin* geenide insertsiooniajad. Harude pikkused on võrdelised ajaga (miljonites aastates, ing k *Myr*). (B) Värvilistes kastikestes on vastavalt (A) poolele välja toodud neli peamist platsenta struktuuri tüüpi. Joonis on koostatud Lavialle *et al*. (2013) poolt.

Tuntud näiteks on *syncytin* geen, mis on omandanud uue funktsiooni peremeesorganismis. Antud geen pärineb endogeense retroviiruse kattevalgu geenist ning on integreerunud **sõltumatult** primaatide, näriliste, kiskjaliste, mäletsejate ning jäneseliste genoomidesse (joonis 2; Mi *et al*., 2000; Dupressoir *et al*., 2009; Heidmann *et al*., 2009, Vernochet *et al*., 2011, Cornelis *et al*., 2012; Cornelis *et al*., 2013). Joonisel 2 on lillade kolmnurkadega välja toodud toimunud integratsioon ning lisaks organismidele omased erinevad platsenta tüübid. Huvitavaks on see, et kõigis nendes liikides võeti valk **iseseisvalt** kasutusele platsenta formeerumise protsessis (ing k *plancentation*), mis on vajalik süntsüütsiumi arenguks (Mi *et al*., 2000; Dupressoir *et al*., 2009; Heidmann *et al*., 2009).

1.1.2. De novo geenide esmane ekspressioon organismides

Paljudel uutel *de novo* tekkinud geenidel on leitud esmast ekspressiooni ühes organis – testistes (Betran *et al*., 2002; Marques *et al*., 2005). „Testistest välja“ (ing k „*out of the testis*“) teooria kohaselt on testis oluline uute geenide evolutsiooniks, andes neile võimaluse tekkida ja areneda ning võimaluse korral omandada aja jooksul funktsioone teistes, somaatilistes kudedes (Marques *et al*., 2005; Vincknbosh *et al*., 2006; Kaessmann *et al*., 2009).



Joonis 3. Uute geenide teke „testistest välja“ hüpoteesi baasil. Hüpoteesi järgi uute geenide koopiate (rohelised kastikesed) transkriptsioon on hõlbustatud kindlates testiste sugurakkudes – meiootilistes spermatotsüütides ning post-meiootilistes spermatiidides tänu üldisele kromatiini avatumale olekule ning transkriptsiooniks vajalike baaselementide üleekspressioonile. Antud olukord tekib üldjuhul tänu laiahaardelisele CpG rikaste promootorite demetüleerimisele ning histoonide (sinised ovaalid) modifikatsioonidele (atsetüleerimine ning metüleerimine), mida abistab transkriptsiooni masinavärk (punased ovaalid). Tänu sellele võib toimuda uute geenide transkriptsioon (transkriptid roheliste laineliste joontega). Joonis koostatud Kaessmann *et al*. (2010) poolt, tõlgitud eesti keelde töö autori poolt.

Testis on tugeva selektiivse surve ning seksuaalse, sugulise või reproduktiivse valiku all, tänu millele toimub kiire evolutsioon (Nielsen *et al*., 2005). Eriti imetajate puhul on selle protsessi puhul oluline transkriptsiooni omadused meiootilistes ning postmeiootilistes spermatogeensetel rakkudel – vastavalt spermatotsüütides ja spermatiidides (joonis 3). Molekulaarsed analüüsid viitavad erinevatele spetsiifilistele histooni-variatsioonidele ning modifikatsioonidele, mis võivad soodustada avatud kromatiini antud rakkudes (joonis 3; Kleene, 2001; Sassone-Corsi, 2002; Kimmins ja Sassone-Corsi, 2005; Vincknbosh *et al*., 2006). Samuti on leitud naabergeenide avatud kromatiini ja/või regulaatorelemente kasutatavust uuritava geeni jaoks (Vincknbosh *et al*., 2006).

Vincknbosh *et al*. (2006) leidsid, et inimese transkribeeritud noored retrogeeni koopiad, mis puuduvad hiirtes, on transkribeeritud madalal tasemel inimestes, omades sellest suurt proportsiooni (10.7%) testistes. Samas transkribeeritud vanematel retrokoopiatel, mille ortoloogid esinevad ka hiirtes, esineb tunduvamalt väiksem testiste kallutatus (5.4%). Seega tekkinud ekspressioon sugurakkudes võib võimaldada geenil tekitada uusi reaalseid funktsioone teistes kudedes. Kuna sugurakkudes on vajalik vaid väga lihtne promootor (Kleene, 2005), siis toimuv kiire evolutsioon võib võimaldada mõningate mutatsioonide abil tekitada sobilikke promootoreid, mis aitavad järgnevalt omandada uusi funktsioone teistes rakkudes või jääda vajalikeks geenideks testiste piires (Kaessmann *et al*., 2010). Sellisteks geenideks võivad ka olla segmentaalsel duplikatsioonil tekkinud või kimäärsed geenid:

- *Tre2* (*USP6*) – onkogeen, mis on tekkinud kimäärselt kahe geeni (*UPS32* ja *TBC1D3*) fuseerumisel. Hominiidide spetsiifiline geen ekspresseerub ainult testistes, erinevalt eellasgeenide laiast ekspressioonispektrist, ning arvatakse, et antud geen võib olla seotud rakkude proliferatsiooni ning spetsifikatsiooniga. (Paulding *et al*., 2003)
- She *et al*. (2004) avastasid inimesel 28 uut transkripti tsentromeeride lähedasest piirkonnast, mis on tekkinud kas eksonite fuseerumisel või lisa eksoni omandamisel mõnest teisest regioonist. Üheteistkümnel neist on ennustatud intaktne ORF ning uuritud 16st transkriptist omasid kõik ekspressiooni testistes ning pooled munasarjas. (She *et al*., 2004)
- Levine *et al*. (2006) leidsid *Drosophila melanogaster´*i genoomist viis geeni, mis on tekkinud mittekodeerivast alast. Nendest neli asuvad X-kromosoomil ning RT-PCR analüüsid näitavad ülekaalukat ekspressiooni testistes. Nende teooria kohaselt on X-kromosoomil paiknevatel geenidel tunduvamalt lihtsam genereerida mutatsioone või aja jooksul fikseeruda võrreldes autosoomsetel kromosoomidel paiknevate geenidega. (Levine *et al*., 2006)

Testistes ekspresseeruvatest geenidest on leitud suur hulk, mis asuvad X-kromosoomis (Begun *et al*., 2007). Begun *et al*. (2007) leidsid, et enamjaolt on käsitletavad geenid noored, eriti võrreldes X-kromosoomist autosoomi läinud geenidega ning antud juhul võiks spermatogeneesis toimuv hüpertranskriptsioon omada kasulikku efekti.

Samas X kromosoomi inaktivatsioon varajases spermatogeneesis võiks anda eelise vanematele retrotransponeeritud geenidele, mis on integreerunud autosoomidesse, võimaldades parema ekspressiooni ning uue funktsiooni tekke testistes (Betran *et al*., 2002; Emerson *et al*., 2004; Vincknbosh *et al*., 2006). Seda fenomeni kutsutakse „X´st välja" (ing k „*out of the X*").

Betran *et al*. (2002) uurisid põhjalikumalt *Drosophila* retrogeenide päritolu, keskendudes 24 *Drosophila´*l leitud retrogeenide paarile. Antud juhul määrati uueks puuduvate intronitega või poolikut polü(A) saba omavad geenid. 50% eellasgeenidest esinesid X-kromosoomil ning nende puhul oli uus retrokoopia integreerunud autosoomi, omades peamiselt ekspressiooni testistes (91%; Betran *et al*., 2002). See on suur määr, arvestades et ainult 10% *Drosophila* geenidest on märgatud ekspressiooni antud koes (Andrews *et al*., 2000). Sarnast fenomeni, kus X kromosoomis esineb tunduvamalt suurem määr retrotransponeeritavaid vanemgeene kui autosoomidel, on detekteeritud lisaks nii inimestes kui ka hiirtes (Emerson *et al*., 2004).

1.2. ENDOGEENSED VIIRUSLIKUD ELEMENDID

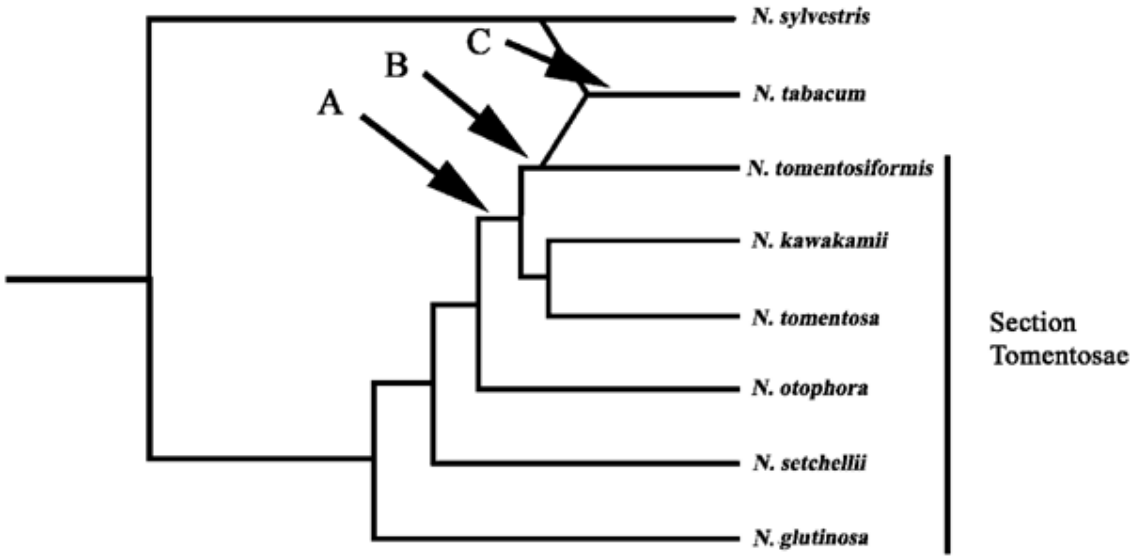
Enogeensed viiruslikud elemendid (EVE, ing k *endogenous viral elements*) on viiruslikku päritolu organismi genoomi integreerunud **järjestused**, mis on läbi põlvkondade fikseerunud. Üldjuhul peetakse nende all silmas laialt levinud retroviiruslikke (ERV ing k *endogenous retroviruses*) elemente, kuid antud töös keskendutakse mitte-reroviiruslikele järjestustele ning just nendele rakendatakse mõistet EVE.

1.2.1. EVE´de avastamine genoomidest

Esimesteks EVE´deks, mis detekteeriti, olid puukentsefaliidi viiruste (ing k *Tick-borne encephalitis virus*; taksonoomia: viirused, ssRNA viirused, (+)ssRNA, *Flaviviridae*, *Flavivirus*) ning leetriteviiruste (ing k *Measles virus*; taksonoomia: viirused, ssRNA viirused, (-)ssRNA, *Mononegavirales*, *Paramyxoviridae*, *Morbillivirus*) järjestuste avastamine ning integreerumise kinnitamine eukarüootsetes **rakuliinides** (Zhdanov *et al*., 1974; Zhdanov ja Parfanovich, 1974, Haspel, *et al*., 1973). Antud rakuliinid, inimese HEp2 ning kana ja hamstri embrüo fibroblasti-rakud, on krooniliselt infekteeritud vastavate viirustega ning toodavad

terviklikke viiruse-spetsiifilisi ribonukleoproteiine, samas kui küpsete virionide tootmine on takistatud. Nende teooria kohaselt aitab rakuliinides esinev latentne onkornaviirus puukentsefaliidi ning leetriviirustel integreeruda terviklikult genoomidesse, toimunud protsessi kinnitati hübridisatsioonikatsetega. (Zhdanov *et al*., 1974; Zhdanov ja Parfanovich, 1974, Haspel, *et al*., 1973)

Esimeseks taimedest detekteeritud mitte-retroviiruslikeks EVE´deks olid geminiviiruse sarnased järjestused (GRD, ing k *geminivirus related DNA*). Nende puhul leiti geminiviiruste *AL1* (*rep*) geeni osasid suurtes kordustes (~340) ühes kindlas regioonis *Nicotiana tabacum´*i (tubakataime) genoomist (Bejarno *et al*., 1996). Hiljem leiti antud elemente lisaks ka *N. tomentosa* ja *N. tomentosidormis* genoomidest (Ashby *et al*., 1997). Kuna elementidel esinesid viirustele vajalikud replikatsiooni-elemendid (TATA-järjestus, DNA lingud, nukleosiidtrifosfaadi seondumissait ning lühikesed inverteeritud kordusjärjestused), pakuti välja viirustelt peremehele toimunud geneetilise materjali ülekannet ning fikseerumist tubakataime meristeemkoesse (Bejarno *et al*., 1996).



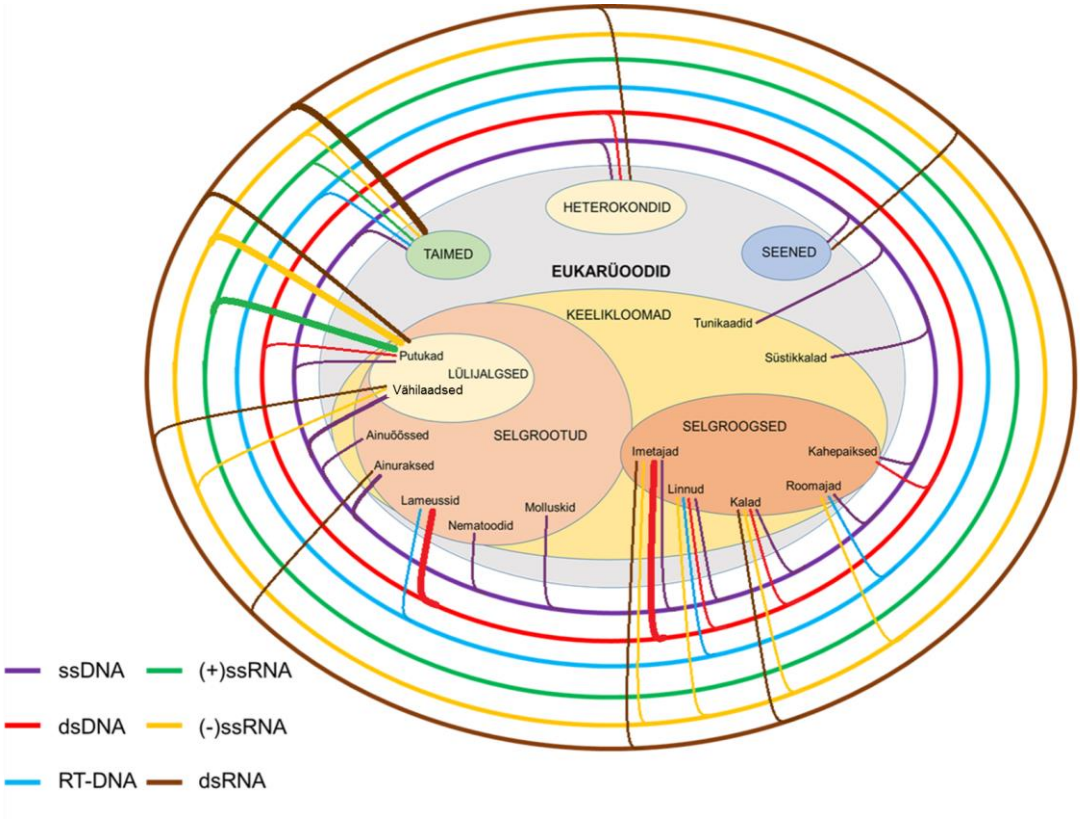
Joonis 4. GRD elementide integreerumine tubakataimede genoomidesse. GRD5 tüüpi element integreerus *Nicotiana* genoomi pärast *N. otophora* divergeerumist (A), arvatavasti tänu geminiviiruse genoomi rekombineerumisega (Ashby *et al*., 1997). Järgnevalt järjestus metüleeriti, amplifitseeriti ning mitmekesistus enne järgmist *Tomentosae* liigiteket. Pärast *N. tomentosiformis* mitmekesistumist (B), tekkis arvatavasti järgmine GRD perekond läbi geminiviiruse uue integreerunud GRD5 ning endogeense Helitron elemendi rekombineerumisele, moodustades GRD3 perekonna. Järgnevalt tekkis uus tubakataimede liik – *N. tabacum*, mis moodustus *N. tomentosiformis* ning *N. sylvestris* genoomide hübridisatsioonil (C). Antud joonis on koostatud Murad *et al*. (2004) poolt.

Murad *et al*. (2004) uurisid tubakataimedes esinevaid GRD´sid põhjalikumalt ning leidsid, et need pärinevad begamoviiruste perekonnast ning on jaotunud kahte selgelt eristatavasse klaasi: perekonnad GRD5 ning GRD3. Nende teooria kohaselt on toimunud kaks eristatavat integratsiooni (joonis 4): esimene pärast *N. otophora* divergeerumist (joonis 4 punkt A; Ashby, *et al*., 1997). Järgnevalt GRD5 metüleeriti, amplifitseeriti ning mitmekesistus, seda kõike enne järgmise liigi teket. Teine integratsioon (joonis 4 punkt B) peaks olema toiminud enne *N. tomentosiformis´*e divergeerumist, kus integreerunud uus GRD5 ning *Helitron* (veereva rõnga replikatsiooni transposoon) ühinesid, moodustades GRD3 perekonna *N. tabacum´*is (joonis 4 punkt C; Murad *et al*., 2004; Murad *et al*., 2002). Seda kinnitavad ka rist-hübridisatsioon (Bejarno *et al*., 1996), PCR (Bejanro *et al*., 1996; Murad *et al*., 2004), fluorestsents *in situ* hübridisatsiooni (FISH; Lim *et al*., 2000) ning genoomsed *in situ* hübridisatsiooni (GISH) analüüsid (Lim *et al*., 2000).

Praeguseks hetkeks on detekteeritud nii tsütoplasmas kui ta tuumas replitseeruvaid viiruslikke järjestusi ((+)ssRNA, (-)ssRNA, dsRNA, ssDNA, dsDNA ning RT-DNA) nii loomade, taimede kui ka seente genoomidest (joonis 5; Frank ja Wolfe, 2009; Horie ja Tomonaga, 2011; Katzourakis ja Gifford, 2010; Aiewsakun ja Katzourakis, 2015; Chu *et al*., 2014). Üldjuhul detekteeritakse üksikuid geene mõnede koopiatega, tihti esinevad neil polü(A) sabad ning asuvad transposoonide läheduses (Katzourakis ja Gifford, 2010; Cui *et al*., 2014; Aswad ja Katzourakis, 2014; Li *et al*., 2015).

Aiewsakun ja Katzourakis (2015) uurisid lähemalt varasemalt detekteeritud EVE´sid. Joonisel 5 on välja toodud Aiewsakun ja Katzourakis (2015) ülevaatetöö baasil koostatud skeem erinevatest viirustüüpidest ning kõrgemate taksonite tasemel eukarüootsetest organismidest, kelle genoomi on kirjeldatud viirused integreerunud. Samuti on lisatud antud artiklist puuduvad, kuid varem detekteeritud ning avaldatud andmed (Murad *et al*., 2004; Kenton *et al*., 1995; Maori *et al*., 2007; Fan ja Li, 2011; Cui ja Holmes, 2012 A; Cui ja Homes, 2012 B; Ballinger *et al*., 2012; Fort *et al*., 2012; Song *et al*., 2013; Arrigada ja Gifford, 2014, Cui *et al*., 2014; Geering *et al*., 2014; Stenglein *et al*., 2014; Umber *et al*., 2014; Bruenn *et al*., 2015; Li *et al*., 2015; Li ja Li, 2015; Mushegian ja Elena, 2015). Skeemil esitatud jooned näitavad viirussugukondade arvukust – kõige peenem joon viitab 1-2 viiruse sugukonna ning kõige paksem 4-5 sugukonna geneetilise materjali ülekandele. Joonisel tuleb välja suur integreerumine taimede, lülijalgsete ning selgroogsete organismide genoomidesse, samas kui lihtsamad organismid – heterokondid, nematoodid, ainuõõssed ning tunikaadid sisaldavad viiruslikke elemente tunduvamalt vähem. Esinevat efekti võib seletata ka viimaste vähene sekveneerimine ning annoteerimine. Kõige arvukamalt on esindatud ssDNA viirused 4

viirusliku sugukonna 17 integratsiooniga. Kõige rohkem viirussugukondi, kelle genoomist on integratsioone toimunud, on dsDNA viirustel, 9 sugukonnaga. Aiewsakun ja Katzourakis (2015) leidsid lisaks, et kirjeldatud integratsioonid on toimunud 4.6-207 MAT.



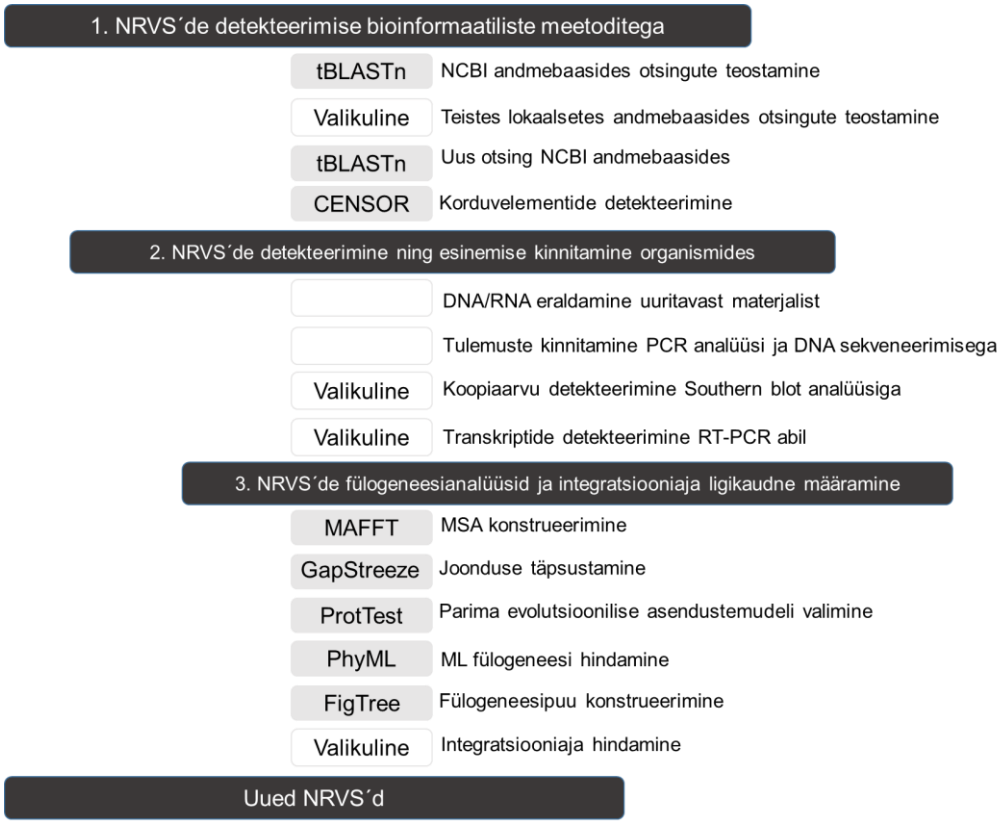
Joonis 5. Viiruslike järjestuste integreerumised kõrgemate eukarüootsete taksonite piires. Värvidega on välja toodud suuremad viirustüübid ning joone paksus näitab viiruste sugukondade arvu, kellel on toimunud integratsiooni tuvastatud. Joonis on koostatud Aiewsakun ja Katzourakis (2015) artikli baasil, millele on lisatud puuduolevad integreerumised (Murad et al., 2004; Kenton et al., 1995; Maori et al., 2007; Fan ja Li, 2011; Cui ja Holmes, 2012 A; Cui ja Homes, 2012 B; Ballinger et al., 2012; Fort et al., 2012; Song et al., 2013; Arrigada ja Gifford, 2014, Cui et al., 2014; Geering et al., 2014; Stenglein et al., 2014; Umber et al., 2014; Bruenn et al., 2015; Li et al., 2015; Li ja Li, 2015; Mushegian ja Elena, 2015). Joonis on koostatud töö autori poolt.

Chu *et al*. (2014) uurisid lähemalt taimede genoomidesse integreerunud viiruslikke järjestusi. Antud töös leiti, et 81% integreerunud viiruslikest järjestustest olid dsRNA päritolu ning kõige vähemal määral oli esindatud (+) ssRNA viiruslikud järjestused (2%). Viiruse sugukondadest olid enamus esindatud mõndade üksikute viirustega, välja arvatud Partitiviridae 26 esindajaga (Chu *et al*., 2014).

Nagu joonis näitab, on toimunud palju ülekandeid erinevatest viirustest mitmetesse taksonitesse, seega on näha, et geneetilise materjali ülekanne viirustelt peremeesorganismidele, ei ole haruldane sündmus.

1.2.2. EVE´de tuvastamise metoodika

EVE´de avastamise ning iseloomustamise tööskeem, mida üldjuhul kasutatakse, koosneb kolmest suuremast etapist (joonis 6):



Joonis 6. Endogeensete mitte-retroviiruslike RNA viiruste sarnaste järjestuste (NRVS, ing k *non-retroviral RNA virus-like sequences*; sünonüüm EVE´le) detekteerimise ja analüüsi skeem. Koosneb kolmest peamisest etapist: (1) NRVS´de detekteerimised bioinformaatiliste meetoditega; (2) NRVS´de deteketeerimine ning esinemise kinnitamine organismides; (3) NRVS´de fülogeneesianalüüsid ja integratsiooniaja ligikaudne määramine. Skeem on koostatud Kondo *et al*. (2015) poolt ning tõlgitud eesti keelde töö autori poolt.

1. NRVS´de detekteerimised bioinformaatiliste meetoditega (joonis 6 etapp 1)

Peamiseks otsingutüübiks, soovides avastada viirusliku nukleiinhappe ülekannet, on BLAST analüüsid, mis on homoloogia otsingute teostamiseks üks levinum viis, suutes detekteerida sarnaseid järjestusi kahe järjestuse võrdluse alusel. Antud otsingu puhul on määratud peamiselt limiteerivaks parameetriks E-väärtus (väiksem kui $1e^{-5}$ peetakse kandidaadiks, välja arvatud Cui ja Holmes, 2012 A – BLAST otsingu E-väärtuse lävend 0.9). Otsingu teostamiseks peab olema kas eelteadmised, missugust viiruslikku järjestust otsida (Bejarno *et al*., 1996 – rist-hübridisatsiooni teke geminiviiruste resistentsete taimede konstrueerimisel; Tanne ja Sela, 2005

– viinamarjades tekkiv reaktsioon kartuliviirus Y (ing k *potato virus Y*) antiseerumi vastu) või tuleks teostada süstemaatilised otsingud (Belyi *et al*., 2010 A – ssDNA viiruste vanuse määramiseks; Liu *et al*., 2010 – dsRNA EVE´de detekteerimine; Liu *et al*., 2011 – *Parvoviridae* detekteerimine eukarüootsetest organismidest; Cui ja Holmes, 2012 A – RNA viiruste detekteerimine putukate genoomidest; Hayward *et al*., 2013 – EVE´de deketeerimine nahkhiirtes) erinevate viiruslike järjestuste baasil.

Leitud järjestused tihti BLAST´itakse üldjuhul uuesti (Liu *et al*., 2010; Liu *et al*., 2011; Cui ja Holmes, 2012 A). See võimaldab detekeerida juba avastatud eukaroüütsetele organismidele sarnaseid, kuid BLAST otsingu jaoks piisavalt erinevaid integratsioone. .

2. NRVS´de deteketeerimise ning esinemise kinnitamine organismides (joonis 6 etapp 2)

Selle etapi eesmärgiks on katseliselt kinnitada bioinformaatiliste meetoditega saadud tulemusi. See on oluline just näitamaks, et algses laboris, kus teostati genoomide või genoomijärjestuste sekveneerimised, ei ole toimunud proovide saastust ning uuritav regioon eksisteerib ka tegelikult organismis. Üldlevinuks EVE´de kinnitamise meetodiks on DNA või transkripti tuvastamine uuritavates organismides, valgu olemasolu üldjuhul katseliselt ei uurita.

EVE´de eksperimentaalse kinnitamise esimeseks etapiks on DNA või RNA eraldamine BLAST analüüsil saadud organismidest. See võib olla teostatud kas kindla koe või terve organismi piires. Järgnevalt teostatakse DNA järjestuse kinnitamiseks PCR reaktsioon, uuritava järjestuse puhastamine agaroosgeelilt ning sekveneerimine (Horie *et al*., 2010; Liu *et al*., 2010; Liu *et al*., 2011). Transkriptide detekteerimiseks eelneb kirjeldatud protsessile pöördtranskriptsioon (Hayward *et al*., 2013). Samuti võib teostada Southern blot analüüsi, määramaks uuritava järjestuse koopiaarvu organismis (Horie *et al*., 2010), ning RT-PCR, määramaks transkriptide esinemist organismides (Liu *et al*., 2010). Viimasele võib anda eelteadmisi ka lisaks ekspressiooniandmebaasidest (näiteks NCBI EST andmebaas) teostatud otsingud, viidates ekspressiooni olemasolule/puudumisele erinevates kudedes (Liu *et al*., 2011). Lisaks võib uurida ORF´ide pikkuseid, sünonüümsete ja mittesünonüümsete asenduste suhet, stop-koodonite esinemist, raaminihke mutatsioone ning insertsioone/deletsioone detekteeritud EVE´des (Liu *et al*., 2010; Liu *et al*., 2011). Need meetodid näitavad, kas organismis hoitakse antud valgujärjestust muutumatuna, takistades mutatsioonide teket ning geeni pseudogeenistumist, mis omakorda võib viidata sellele, kas organism kasutab antud valku ning muutused selles ei pruugi olla kasulikud.

3. NRVS´de fülogeneesianalüüsid ja integratsiooniaja ligikaudne määramine (joonis 6 etapp 3)

Viimaseks etapiks on uuritava järjestuse fülogeneetilised analüüsid (Horie *et al*., 2010; Belyi *et al*., 2010 A; Liu *et al*., 2010; Liu *et al*., 2011; Cui ja Holmes, 2012 A; Hayward *et al*., 2013), milleks joondatakse valgujärjestused ning konstrueeritakse fülogeneetilised puud. Järgnevalt proovitakse määrata insertsiooni toimumise ligikaudne aeg, kasutades selleks kas integratsioone sisaldavate eukarüootsete organismide lahknemise aegu (Belyi *et al*., 2010 A – parvoviiruste ning circoviiruste vanuseks vähemalt 40 – 50 MA; Horie *et al*., 2010) või molekulaarseid dateerimise meetodeid (Hayward *et al*., 2013).

1.2.3. EVE´de funktsiooni uurimine

Kõige põhjalikumalt on uuritud endogeenseid bornaviiruste-sarnaseid (EBL, ing k *endogenous bornavirus-like*) nukleokapsiidi (N; ing k *nucleocapsid*) elemente. Esimesena detekteeris EBLN elemente Horie *et al*. (2010), leides neid primaatide, orava, hiire ning roti genoomidest BLAST otsingu baasil. Antud tulemused kinnitati PCR ning Southern-blot analüüsidega. Imetajatest tuvastatud elemendid jagunesid neljaks klassiks, millest kolm, välja arvatud EBLN-1 elementide klass, omasid kodeerivas regioonis stop-koodoneid või neil puudusid identifitseerivad külgmised järjestused (Horie *et al*., 2010). Fülogeneetilised analüüsid viitasid integratsioonile primaatide eellasse enne reesusmakaagi ning marmoseti lahknemist, ligikaudu 40 millionit aastat tagasi (MAT), samas integratsioon orava genoomi on olnud tunduvamalt tänapäevasem, ligikaudu 20 MAT (Belyi *et al*., 2010 B). (Horie *et al*., 2010)

Lisaks on EBL elemente (seal hulgas nukleokapsiidi, maatriksvalku, RNA-sõltuvat RNA polümeraasi ja glükoproteiini) avastanud süstemaatiliste EVE-de otsingute abil Belyi *et al*. (2010 B), Katzourakis ja Gifford (2010), Horie *et al*. (2013), Gilbert *et al*. (2014) ning Cui *et al*. (2014).

Fujino *et al*. (2014) uurisid lähemalt orava (*Ictidomys tridecemlineatus*) EBLN elemente, mille transkribeeritud mRNA omab 77% järjestuse identsust bornaviiruste nukleoproteiini valgujärjestusega. Täpsemalt uuriti EBLN elemendi mõju *Borna disease virus´*e (BDV) replikatsioonile, mille produkti detekteeriti RT-PCR abil nii metsiku kui ka vangistuses hoitud orvatel ajus, südames ning testistest või munasarjas. Töös leiti, et OL **rakuliini kloneeritud** oravatest pärit **EBLN** kolokaliseerub viraalsete vabrikutega tuumas, inimese EBLN-1 elemendi puhul kirjeldatud seost ei tuvastatud. Lisaks leiti, et orava EBLN, kuid mitte inimese EBLN-1, ekspressioon vähendab viiruste genoomi hulka ning mRNA taset 48h pärast infektsiooni. See

võib viidata BDV replikatsiooni ja transkriptsiooni inhibitsioonile, täpsemalt inkorporeerudes infekteeritud rakkudes viiruslike ribonukleiinhappe-valk kompleksi (RNP; ing k *ribonucleoprotein*) koosseisu, inhibeerides viiruslikku replikatsiooni ning toimides N elemendi inhibiitorina, takistades selle seondumist RNP-ga. (Fujino *et al*., 2014)

Kahjuks antud töö puhul ei õnnestunud autoritel leida orava spetsiifilisi rakuliine ega tekitada eksperimentaalset BDV infektsiooni organismis, seega kirjeldatud funktsiooni ei ole suudetud kindlalt tõestada. Samas tõestati, et organismides detekteeritud elemendi ekspressioon mRNA´na võib viidata kirjeldatud funktsiooni suunas. (Fujino *et al*., 2014)

1.3. VALKUDE STRUKTUURI KAASAMINE ANALÜÜSIDESSE

Võrreldes organismide ning viiruste kohanemisvõimet keskkonnaga, just nukleiinhappe tasemel, siis saadavad mutatsioonikiirused on kordades erinevad: RNA viirustel 10^{-2}-10^{-7} ning eukarüootsetel organismidel 10^{-9} asendust positsiooni kohta aastas (Abroi ja Gough, 2011). Esinev suur mutatsioonikiirus tasakaalustatakse puhastava valiku abil, eemaldades kahjulikud mutatsioonid (Holmes, 2009). See näitab, et uued järjestused võivad tekkida väga kiiresti, kuid struktuuri tasemel toimuvad muutused vajavad selleks tunduvamalt rohkem aega (Caetano-Anolles ja Nasir, 2012). Seega – valkude struktuur on tunduvamalt konserveerunud kui järjestus (Caothia ja Lesk, 1986; Illegard *et al*., 2009). Kaasates struktuure järjestuste seoste uurimisse (sarnasuse identifitseerimiseks või fülogeneetiliste suhete uurimiseks), võiks see saadud tulemusi tunduvalt parandada.

1.3.1. Valkude struktuur sarnasuse identifitseerimises

Evolutsiooniliste suhtete identifitseerimaks kasutatakse tavaliselt kahe järjestuse võrdlusi, mille alusel koostatakse fülogeneetilised puud. Samas on leitud, et alla 30% sarnasusemääraga järjestusi on BLAST otsinguga tunduvamalt raskem identifitseerida. Seda tõestasid Brenner *et al*. (1998), uurides teadaolevate evolutsiooniliste suhetega (struktuuri, järjestuse ja funktsiooni poolest), kuid 20-30% identsusega valke, saadi positiivseid tulemusi ainult pooltel juhtudel. Kirjeldatud probleemide ületamiseks on välja töötatud mitmeid protseduure: matriitsid (Taylor, 1986), peidetud Markovi mudelid (HMM; Krogh *et al*., 1994), PSI-BLAST (Altschul *et al*., 1997), PROBE (Neuwald *et al*., 1997) ning ISS (Park *et al*., 1997). Park *et al*. (1998) leidsid uurides ISS, HMM, PSI-BLAST´i ning paariviisiliste järjestuste võrdlemise meetodite tööefektiivsusi, et uuemad meetodid on tunduvamalt efektiivsemad. Karplus *et al*. (1998)

leidsid, et HMM SAM-T98 meetodil esineb kõige vähem vigu ning see on optimiseeritud leidmaks valkude struktuuride evolutsioonilisi sidemeid superperekonna tasemel.

Murzin *et al*. (1995) lõid SCOP andmebaasi, kuhu kogutakse valgudomeenid, mis esinevad looduses isolatsioonis üksikutena või multidomeense valgu osana (Lo Conte *et al*., 2002), klassifitseerides need struktuuri baasil nelja tasemega hierarhiasse (lisa 1): klass, pakkimise ehk voltumise tase, superperekond ning perekond. Voltumise tasemel esinevad valgud, mille põhistruktuur on pakitud sarnaselt ning superperekonna tasemel on sarnase struktuuri ja väikese identsusega, kuid ühise evolutsioonilise päritoluga valgud (Murzin *et al*., 1995).

Neid kahte nähtust on ära kasutanud valkude ja genoomide struktuurse ja funktsionaalse annotatsiooni andmebaas – SUPERFAMILY (Gough, 2002), kus võetakse teadaolevad SCOP superperekonna valgudomeenid ning otsitakse samasse **superperekonda** kuuluvaid valgudomeene täielikult sekveneeritud genoomidest järjestuse-profiil analüüside abil, kasutades selleks HMM mudeleid (Gough *et al*., 2001; Gough, 2002).

Igale SCOP´i valgustruktuuri superperekonnale on loodud HMM mudelite abil oma profiil (Gough *et al*., 2001; Gough ja Chothia, 2002; Gough, 2002), andes parema võimaluse identifitseerimaks kaugeid homolooge (Karplus *et al*., 1998; Gough, 2002). HMM profiilid moodustavad superperekonna tasemel valgudomeenide mudelite raamatukogusid, mille vastu teostatud otsingud aitavad leida uuritavatele järjestustele homolooge (Gough *et al*., 2001; Gough ja Chothia, 2002; Gough, 2002). Selleks võrreldakse kõiki täielikult sekveneeritud organismide valkude järjestusi HMM mudelite raamatukogu vastu, andes igale vastele skoori ning järjestus määratakse kõrgeima skoori andnud mudeliga samasse superperekonda (Gough *et al*., 2001).

Kaasates struktuursed valgudomeenid homoloogiaotsingutesse, on olnud võimalik tuvastada väga kaugeid evolutsioonilisi suhteid, mida on vanade meetodite (näiteks kahe järjestuse võrdluste) abil tunduvamalt raskem detekteerida.

1.3.2. Valkude struktuuri kaasamine fülogeneesiuuringutesse

Sama põhimõtet saab kasutada ka fülogeneesipuude koostamisel tekkiva ebakindluse lahendamisel, eriti kaugemalt seotud järjestuste puhul (Challis ja Schmidler, 2012; Herman *et al*., 2014). Morrison (2009) leidis, et järjestuste joondused on fülogeneesi uurimise kõige nõrgim lüli, kuna homoloogia, eriti kauge homoloogi kirjeldamiseks puuduvad objektiivsed meetodid. Antud probleem esineb just kaugelt lahknenud organismide ning insertsioonide ja

deletsioonide puhul, kus vale mudeli kasutamine suudab piisavalt tulemusi kallutada (Caetano-Anolles ja Nasir, 2012).

Challis ja Schmidler (2012) uurisid probleemi ning nende lahenduseks oli valkude struktuuri kaasamine fülogeneesipuude koostamisse, kuna see on pikemas ajaskaalas tunduvamalt konserveerunum (Caothia ja Lesk, 1986; Illegard *et al*., 2009) üldjuhul tänu funktsiooni säilitamise vajadusele. Struktuuri kaasamine peaks parandama joondust ja ennustatud evolutsioonilisi vahemaid, eriti kaugelt suguluses olevate järjestuste puhul (Challis ja Schmidler, 2012). Struktuuri kaasamisel on saadud tegelike evolutsiooniliste suhete teadmisega puude puhul reaalsusele lähemaid suhteid kui tavalise, ainult järjestuse joonduse baasil koostatud fülogeneeside puhul (Herman *et al*., 2014). Caetano-Anolles ja Nasir (2012) leidsid, et valkude struktuuride kaasamine fülogeneetilistesse uuringutesse peaks lahendama just horisontaalse geeniülekande tagajärjel tekkivaid probleeme.

Mushegian ja Elena (2015) uurisid Tubaka mosaiigiviiruse 30K (rakust-rakku liikumise) valgu homolooge, nii viiruste enda kui ka erinevate taimede genoomidest leitud homolooge. Antud töös, uurides viiruste 30K valgu fülogeneesi, kaasati ka valgu enda ennustatud struktuurid, mis parandas mitmese järjestuse joondust, näidates tugevalt konserveerunud piirkondi ning andes viiteid uute struktuursete iseärasuste kohta. Samas ei kaasatud valkude ennustatud struktuure viiruste ja EVE´de ühisesse fülogeneesipuu koostamisse.

1.4. KÄRBSELISTE FÜLOGENEES

Lülijalgsete (*Arthropods*) hõimkonda kuulub üle 1 miljoni kirjeldatud liigi, mis koosneb ühest väljasurnud alamhõimkonnast (*Trilobita* ehk trilobiidid, 4000 kirjeldatud liiki) ning neljast antud hetkel eksisteerivate esindajatega alamhõimkonnast: *Myriapoda* (hulkjalgsed, seal hulgas saja- ja tuhatjalgsed; 11 500 liiki), *Chelicerata* (lõugtundlased, seal hulgas ämblikud, lestad, odasabalised; 70 000 liiki), *Hexapoda* (ehk kuulejalgsed, seal hulgas **putukalised**; 948 000 liiki) ning *Crustacea* (ehk vähilaadsed, seal hulgas krevetid ja krabid; 68 000 liiki) (Pisani, 2009). Kõige vanimaks lülijalgsete fossiiliks on trilobiidid ning nad pärinevad varajasest Kambriumi perioodist 519-523 MAT (Gradstein *et al*., 2004), samas varaseimate trilobiitide biogeograafilise jaotus viitab nende mitmekesistumisele enne hiidkontinendi Pannotia jagunemist, 550-600 MAT (Lieberman, 2003).

Kahetiivaliste (*Diptera*; joonis 7) gruppi kuuluvad nii käbselised (*Brachycera*) kui ka sääselised (*Nematocera*), omades kokku ligikaudu 180 sugukonda üle 150 000 kirjeldatud liigiga (Yeates ja Wiegmann, 1999; Yeates ja Wiegmann, 2005). Nad on olulisteks nii inimese kui ka loomade

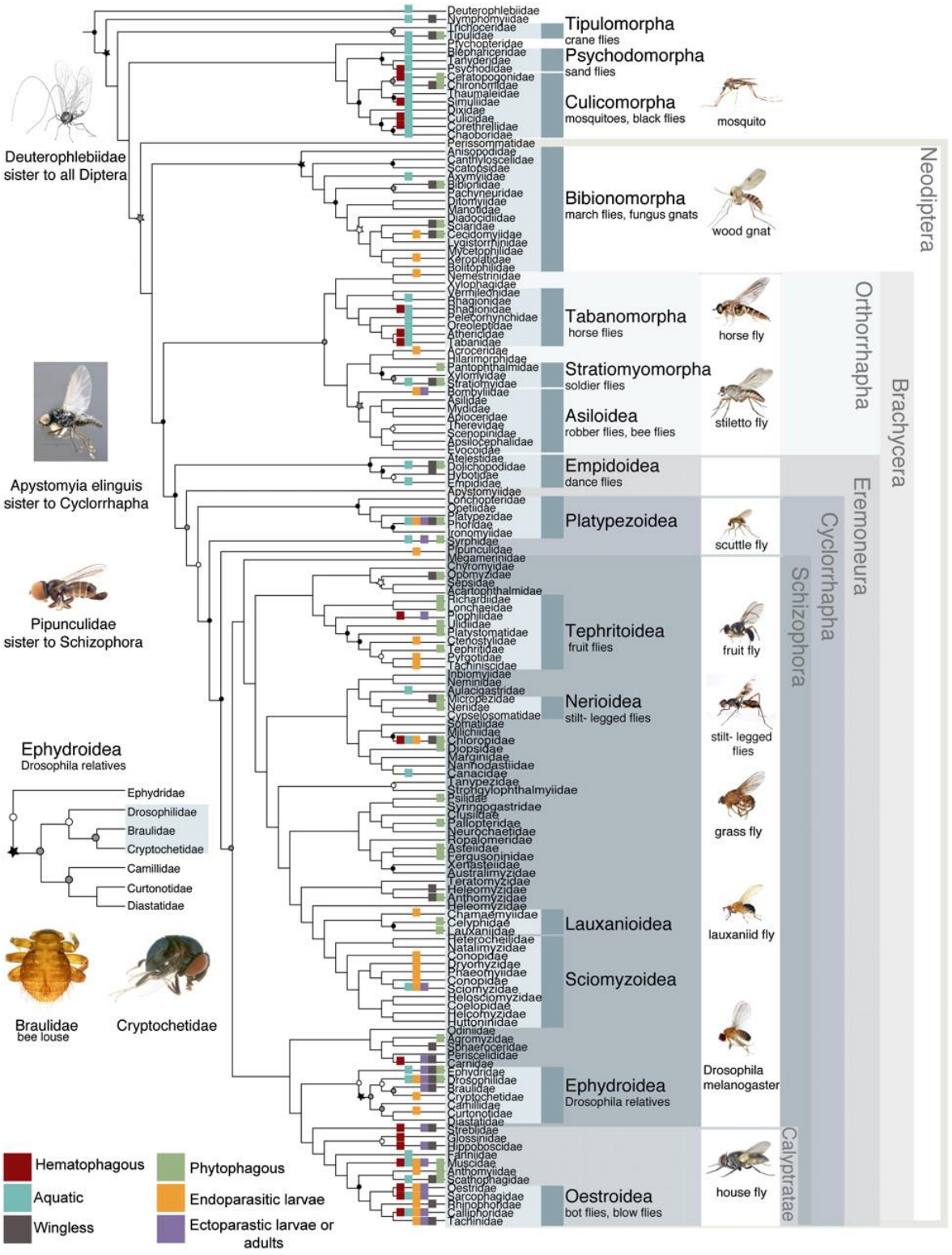
patogeenide vektoriteks (*Culicidae*), vilja ja loomade hävitajateks (*Tephritidae*, *Oestridae*), kiskjad, lagundajad, parasitoidid ning tolmendajad (Labandeira, 2005; Marshall, 2006).

Kahetiivaliste tekke ajaks peetakse Hilis-Juurat (~157 MAT) (Misof *et al*., 2014). Fossiilsete andmete järgi võib pidada kärbseliste eraldumist sääselistest Hilis-Triias Varajase-Juura aega ~187-240 MAT (Mostovski, 2000; Krzeminski ja Evenhuis, 2000; Evenhuis, 1994; Grimaldi ja Cumming, 1999), mida kinnatavad ka molekulaarsete analüüside abil on saadud tulemused (Wiegmann *et al*., 2011; Bertone *et al*., 2008; Wiegmann *et al*., 2003; Misof *et al*., 2014).

Brachycera koosneb neljast monofüleetilisest infraseltsist (joonis 7): *Xylophagomorpha* (1 sugukond *Xylophagidae*), *Tabanomorpha* (8 sugukonda), *Stratiomyomorpha* (3 sugukonda) ning *Muscomorpha* (üle 100 sugukonna) (Wiegmann *et al*., 2003). *Muscomorpha* tekkeks on molekulaarsetel viisidel määratud 198-240 MA (miljonit aastat; ing k *milion years*) (Bertone *et al*., 2008; Wiegmann *et al*., 2003; Brammer ja von Dohlen, 2007; Winterton *et al*., 2007; Evenhuis, 1994) ning see koosneb omakorda neljast sektsioonist, mis on kategoriseeritud peamiselt morfoloogiliste tunnusjoonte baasil: *Heterodactyla*, *Eremoneura*, *Cyclorrhapha* ja *Schizophora* (Yeates ja Wiegmann, 1999). Viimane koosneb mitmest suuremast alamsektsioonist.

Sirelased (*Aschiza*) kuuluvad *Muscomorpha´*de sektsiooni ning koosneb kahest suurest alamsektsioonist (joonis 7): *Syrphoidea* (sugukonnad *Syrphidae* 85-95 MA (Wiegmann *et al*., 2011; Bertone *et al*., 2008) ja *Pipunculidae* 83 MA (Wiegmann *et al*., 2011)) ning *Platypezoidae* (sugukonnad *Opetiidae*, *Ironomyiidae*, *Lonchopteridae*, *Phoridae* 92 MA (Wiegmann *et al*., 2011) ning *Platypezidae* 107-129 MA (Wiegmann *et al*., 2011; Bertone *et al*., 2008; Gaunt ja Miles, 2002)).

Calyptratae alamsektsioon koosneb 18 000 kirjeldatud liigist, mis moodustab ligikaudu 12% kõigist kirjeldatud *Diptera* liikidest (Kutty *et al*., 2010). Sellesse gruppi kuulub suur valik erinevate omadustega kärbselisi (13 sugukonda; joonis 7): kodukärbsed (*Muscidae*), blowflies (*Calliphoridae*), fleshflies (*Sacrophagidae*), tsetse flies (*Glossinidae*) ning botflies (*Oestridae*) (Kutty *et al*., 2010), kes on olulised meditsiinis (*Glossinidae, Muscidae, Oestridae*), kohtuekspretiisis (*Calliphoridae, Sacrophagidae*) ning tõrjevahenditena (*Tachinidae*) (Bertone ja Wiegmann, 2009). Erinevalt väga kirjust seltskonnast, ei ole antud grupp praeguste andmete kohaselt väga vana – *Drosophila* (*Acalyptratae*) – *Calyptratae* lahknemist (joonis 7) peetakse toimunuks 41-80 MAT (Wiegmann *et al*., 2003) ning *Glossinidae* vanuseks on molekulaarsete andmetega määratud 43-44 MA (Misof *et al*., 2014; Wiegmann *et al*., 2011), mida kinnitavad ka fossiilsed andmed, 34 MA (Evanoff *et al*., 2001).



Joonis 7. Kahetiivaliste (*Diptera*) fülogeneesipuu, mis on koostatud kahe alampuu kombineerimisel RaxML programmi abil. Antud fülogeneesipuu baseerub kahel grupil: (1) 42 liigi 14 tuumageeni, täismitokondriaalse genoomi ning 371 morfoloogilise tunnusega ning (2) 202 esindajat, vähemalt igast sugukonnast ühe esindajaga, 5 tuumageeni uurimisega. Ringid näitavad üle 80% väärtusega bootstrap väärtusi (must – 95-100%; hall – 88-94%, valge 80-88%). Tähekestega on märgitud hilisemad pärast analüüsi teostatud boostrap väärtuste parandamised (must – 95-100%, hall 88-94%, valge – 80-88%). Värvilised ruudukesed näitavad teatud omaduste esinemist vähemalt ühes kärbseliste liigis antud sugukonna piires. Nende tähendused on märgitud joonise all vasakul nurgas. Joonis on koostatud Wiegmann *et al*. (2011) poolt.

Acalyptratae gruppi kuuluvad üle poole seltsi sugukonnalisest mitmekesisusest (~62 sugukonda), samas koosneb vaid 20% kärbseliste kõikidest liikidest (joonis 7). Kõige liigirikkamad sugukonnad antud grupis on *Tephritidae, Lauxaniidae, Agromyzidae, Chloropidae, Drosophilidae* ja *Ephydridae*, mis moodustavad üle 50% liikidest grupis (Bertone ja Wiegmann, 2009). Antud alamsektsiooni kõige parimini uuritud sugukond on *Drosophilidae*, mille vanuseks on saadud 66-83 MA (Misof *et al*., 2014; Gaunt ja Miles, 2002).

2. EKSPERIMENTAALNE OSA

2.1. TÖÖ EESMÄRGID

Käesolev töö jätkab bakalaureusetööd (Kirsip, 2013), kus uuriti viiruste ja peremeesorganismide vahelist geneetilise materjali ülekannet. Täpsemalt keskenduti tubaka mosaiigiviiruse kattevalgule (TMV-CP_{viral}), mida SUPERFAMILY andmebaasi alusel leidus paljudes viirustes ja mõnedes organismides, sealhulgas *Drosophila* kärbseliste genoomis (FlyBase versioon FB2015_01 geen *CG15772*). Lisaks tuvastati integreerunud TMV-CP´d kolmest teisest kärbselistest: *Glossina morsitans morsitans*, *Musca domestica* ning *Ceratitis capitata*, tavalise BLAST analüüsi abil, määrates integratsioon *Schizopohora* kärbseliste eellasesse 60-250 MAT (Kirsip, 2013).

Magistritöö ülesandeks on uute andmete baasil kinnitada toimunud integratsiooni suunda ning täpsustada sündmuse toimumise aega. Selleks uuritakse lähemalt enne *Schizophora´*sid lahknenud kärbselisi – sirelasi (*Aschiza*). Tulemuste usaldusväärsuse hindamiseks uuritakse kärbseliste (*Brachycera*) üldist fülogeneetilist katvust genoomi ja transkriptoomi sekveneerimise andmetega. Samuti katsetatakse erinevaid järjestusanalüüsi meetodeid (HMM mudelid, eelasjärjestused ning struktuuri kaasamine) EVE´de deteketeerimiseks ning püütakse hinnata nende efektiivsust võrreldes tavapraktikaks saanud BLAST analüüsiga.

Töö teiseks eesmärgiks on uurida horisontaalsel geeniülekandel tekkinud *uue* geeni funktsiooni organismides, keskendudes kärbselistes annoteeritud TMV-CP_{fly} funktsiooni määramisele bioinformaatiliste meetoditega:

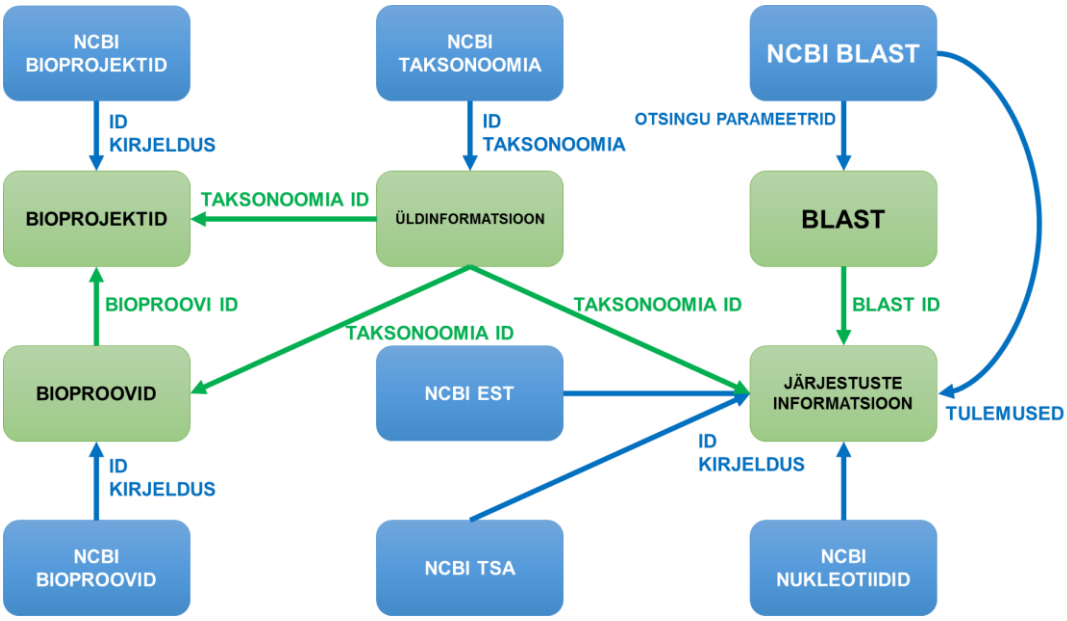
- Tuumagenoomi ning transkriptoomi proovide andmete abil.
- Geenieksрessiooni andmete abil.
- Valk-valk, geen-miRNA ja geen-transkriptsioonifaktor interaktsioonide konserveeruvuse abil.

2.2. MATERJALID JA METOODIKA

Uuritavaid viiruslikke järjestusi nimetatakse käesolevas töös TMV-CP_{viral} ning üldiselt kärbselistesse integreerunud geneetilist materjali TMV-CP_{fly}. *Drosophila melanogaster* organismis eksisteeriva geeni käsitlemisel kasutatakse geeninime *CG15772* (FlyBase ID FBgn0029799, versioon FB2015_01; dos Santos *et al*., 2015).

2.2.1. Andmebaasi koostamine

Tuvastamaks TMV-CP geeni genoomides, mille sekveneerimine ei ole veel lõpetatud ning uurimaks TMV-CP_{fly} ekspressioonimustrit loodi andmebaas, mis sisaldab käesoleva tööga seotud organismide taksonoomia, bioprojektide ning bioproovide andmeid. Samuti lisati suuremahulised BLAST programmi läbitöötamise tulemused ning sellest tulenevad järjestused koos iseloomustavate andmetega. Andmebaasi koostamise eesmärgiks oli paremini hallata kärbseliste spetiisifilisi andmeid, suuremahulisi BLAST tulemusi ning nendest tulenevate seoste hõlpsamat käsitlemist.



Joonis 8. Andmebaasi tööskeem. Siniste kastikestega on märgitud andmete algsed NCBI andmebaasid, millest päring teostati, ning rohelisega antud töös koostatud lokaalse andmebaasi tabelid, kuhu andmed sisestati. Siniste nooltega on näidatud andmete liikumise suund NCBI andmebaasist koostatud lokaalsesse andmebaasi ning välja on toodud tähtsamad andmed, mis sisestati tabelitesse. Roheliste nootlega on välja toodud andmebaasi tabeleid siduvad suhted koos tähtsamate andmetega, mis antud tabeleid seovad.

Andmebaas on koostatud Sybase SQL Anywhere 16 programmi abil SQL keeles. Tööskeem ja koostatud tabelite vahelised seosed on välja toodud joonisel 8. Järgnevalt kirjeldatakse koostatud tabeleid lähemalt, keskendudes nende olulisusele käesolevas töös.

ÜLDINFORMATSIOON ORGANISMIDE KOHTA (joonis 8)

Antud tabel sisaldab organismispetsiifilisi andmeid, peamiselt koosnedes organismide levinud nimetuste, NCBI ID ning taksonoomia andmetega. Tabeli eesmärgiks on siduda teiste tabelite andmed läbi organismide. Andmed on võetud NCBI taksonoomia andmebaasist NCBI koduleheküle (Sayers *et al*., 2009) abil kasutades limiteerivaks märksõnaks „*Diptera*“, 12.05.2015 seisuga.

BIOPROJEKTIDE ANDMED (joonis 8)

Antud tabel sisaldab bioprojektide ID koode, organismide taksonoomia ID koode ning bioprojekti iseloomustavaid andmeid. Tabeli eesmärgiks oli kirjeldada kõiki transkriptoomi ning genoomi projekte, mis on NCBI andmebaasi sisse kantud kärbseliste taksonoomia piires. Andmed on võetud NCBI bioprojektide andmebaasist (Sayers *et al*., 2009) kasutades limiteerivaks otsingusõnaks „*Diptera*“, 12.05.2015 seisuga.

BIOPROOVIDE ANDMED (joonis 8)

Antud tabel sisaldab bioproovide ning taksonoomia ID koode ning neid iseloomustavaid kirjeldusi, mis on NCBI bioproovidesse sisse kantud. Tabeli eesmärgiks oli kokku koguda kärbseliste analüüsides kasutatud bioproovid, tekitada seos BLAST analüüside tulemustega ning kirjeldada nende omadusi: sugu, vanus, kude, asukoht, aeg. Andmed on võetud NCBI bioproovide andmebaasist (Sayers *et al*., 2009) kasutades taksonoomilise piirangu limiteerivaks otsingusõnaks „*Diptera*“, 12.05.2015 seisuga.

BLAST (joonis 8)

Antud tabel sisaldab BLAST otsingu algjärjestuse ja otsingut teostatud andmebaasi andmeid ning otsingu parameetreid. Tabel koostati teostatud suuremahuliste BLAST otsingute ning

nende tulemuste paremaks haldamiseks. Otsingu vasteid iseloomustavad andmed on lisatud BLAST´i tulemuste tabelisse.

Otsingud teostati NCBI kodulehel asuva BLAST (BLAST+ 2.2.30+, uuendatud 6.10.2014; Altschul *et al*., 1997) perekonna TBLASTN (valgujärjestuse baasil teostatud otsingud transleeritud nukleotiidide andmebaaside vastu) programmi abil ajavahemikus 3.02.2015-11.02.2015 ning 12.05.2015. Tulemusi limiteerivaks lävendväärtuseks seati E-väärtus $1e^{-5}$ ning kasutati madala kompleksusega regioonide (ing k l*ow complexity regions*) filtrit. Otsitavad algjärjestused:

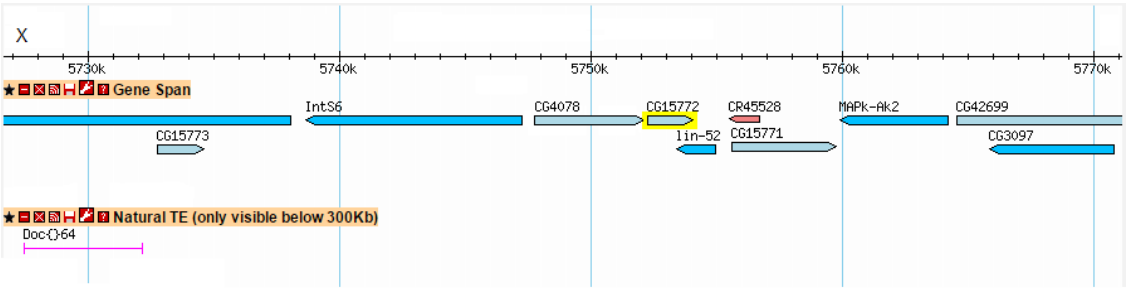
- *Drosophila melanogaster* FlyBase´ist (versioon FB2015_01, uuendustega 24.02.2015; *D. melanogaster* genoomi versioon 6.01; dos Santos *et al*., 2015) **valgujärjestused** (joonis 9). Välja on toodud geeninimed, kromosoom ning geeni nukleotiidse järjestuse koordinaadid.
    - ***IntS6*** (CG3125) - X: 5738663 .. 5747254 [-]
    - ***CG4078*** - X: 5747747 .. 5752077 [+]
    - ***CG15772*** - X: 5752249 .. 5754041 [+]
    - ***lin-52*** (CG15929) - X: 5753390 .. 5754958 [-]
    - ***CG15771*** - X: 5755574 .. 5759751 [+]
- SUPERFAMILY andmebaasist (versioon 1.75; Gough, 2002) *superfamily* taseme SCOP ID 47195, organism *Drosophila melanogaster*, **valgujärjestuse** ID ***FBgn0029799***, regioon 112-259 AH (aminohapet).
- NCBI viiruste täielike genoomide andmebaasist (Sayers *et al*., 2009) *Peanut clump virus* (taksonoomia: viirused, ssRNA viirused, (+)ssRNA viirused, *Virgaviridae*, *Pecluvirus*):
    - Kattevalk (CP; ing k *coat protein*):
        - Nukleotiidne järjestus: NCBI järjestuse ID NC_003668.1, regioon 391 .. 1014.
        - Valk: NCBI järjestuse ID NP_620028.1, regioon 13 .. 185.
- PDB andmebaasist (ing k *RCSB protein data bank*; Berman *et al*., 2000) tubaka mosaiigiviiruse (ing k *Tobacco mosaic virus*; taksonoomia: viirused, ssRNA viirused, (+)ssRNA viirused, *Virgaviridae*, *Tobamovirus*)
    - Kattevalgu struktuuri järjestus: (PDB ID) 1EI7.

Otsingu teostamisel seatud limiteerivaks taksonoomia ühikuks:

- Putukalised (ing k *insecta*, NCBI TaxID 50557).

- Kahetiivalised (ing k *diptera*, NCBI TaxID 7147).
- Kiletiivalised (ing k *hymenoptera*, NCBI TaxID 7399).
- Viirused (ing k *viruses*, NCBI TaxID 10239).

Kasutatud andmebaasid: *nucleotide collection* (nr/nt), *reference RNA sequences* (refseq_rna), *reference genomic sequences* (refseq_genomic), *NCBI Genomes* (chromosome), *expressed sequence tags* (EST), *genomic survey sequences* (gss), *high throughput genomic sequences* (HTGS), *whole-genome shotgun contigs* (wgs) ning *transcriptome Shotgun Assembly* (TSA).



Joonis 9. Uuritava geeni TMV-CP$_{fly}$ ümbritsev ala *Drosophila melanogaster´*i genoomi baasil. Uuritav geen on märgitud kollasega. Üleval vasakul on märgitud kromosoom (X) ning regiooni ligikaudsed koordinaadid, lisaks on välja toodud läheduses asuv transponeeruv element (punane). Joonis on võetud FlyBase koduleheküljelt (versioon FB2015_01; dos Santos et al., 2015) ning täiendatud töö autori poolt.

BLAST´I TULEMUSED (joonis 8)

Antud tabel sisaldab BLAST´i vasteid iseloomustavaid andmeid: järjestuse ID koodi, teostatud BLAST otsingu ID koodi, bit-skoori, skoori, E-väärtust, vaste pikkust, asukohta, lugemisraami, *gap´*de arvu ning vaste tulemust järjestusena. Lisaks loodi alamtabel, mis koosneb järjestuse spetsiifilisest informatsioonist, sisaldades järjestuse pikkust ning sellega seotud organismi NCBI taksonoomia ID koodi. Tabel koostamise eesmärgiks oli iseloomustada BLAST tulemusi ning hallata suuremahulisi andmeid.

BLAST´i vastete andmed on võetud BLAST´imisel saadud tulemustest, järjestusi iseloomustavad andmed on võetud NCBI nukleotiidide, EST ja GSS andmebaasidest (Sayers *et al.*, 2009) järjestuse ID koodi baasil mai 2015.

2.2.2. TMV-CP leidumine kärbselistes

Laiendamaks TMV-CP$_{fly}$ leiduvust, eriti mitte-täielikult sekveneeritud organismides, teostati BLAST otsingud punktis 2.2.1 toodud parameetrite järgi SUPERFAMILY (Gough, 2002) geeni

FBgn0029799 valgujärjestuse baasil. Analoogne otsing teostati lisaks ka *Drosophila melanogaster´*i baasil käsitletava elemendi ümbritsevatele geenidele (FlyBase geeninimedega *IntS6*, *CG4078*, *lin-52* ja *CG15771*; dos Santos *et al*., 2015; joonis 9) TBLASTN alamprogrammi abil. Läheduses asuvatest geenidest (FlyBase ID) *CR45528* ja *MAPk-Ak2* jäeti antud otsingust välja, kuna esimene omab suurt ülekattuvat regiooni geeniga *CG15771* ning teine on väga laialt levinud geen eukarüootsete organismide genoomides (dos Santos *et al*., 2015). Antud tulemuste baasil saaks täpsemalt hinnata toimunud integratsiooniaega, eriti just nende organismide abil, kelle genoom on kvaliteetselt sekveneeritud, kuid uuritavad TMV-CP_{fly} geeni ei ole detekteeritud.

BLAST´i tulemuste alusel loodi kokkuvõttev tabel (tabel 1, lisa 2) kärbselistes esinevatest homoloogsetest geenidest *Drosophila melanogaster´*i baasil (uuritav *CG15772* ning seda ümbritsevad geenid *IntS6*, *CG4078*, *lin-52*, *CG15771*). Iseloomustamaks kui hästi on liik sekveneeritud ja annoteeritud, lisati tabelisse mRNA´de ja valkude arv, mis on NCBI vastavate andmebaaside (Sayers *et al*., 2009) abil detekteeritud või uuritud ning samuti toodi välja kärbseliste sugukondade, perekondade ning liikide tasemel bioprojektide arvud vastavate gruppide piires, 12.05.2015 seisuga.

EELASJÄRJESTUSTE KONSTRUEERIMINE KÄRBSELISTELE NING VIIRUSTELE

Leidmaks toimunud esmasele integratsioonile lähedasemaid järjestusi organismidest, loodi BLAST´i analüüsil detekteeritud kärbsliste alusel (kärbselised kokku, *Acalyptratae* ja *Calyptratae* kärbselised eraldi) ning *Virgaviridae* viiruste alusel antud gruppidele tõenäolised eellasjärjestused, mida omakorda BLAST´iti leidmaks teistest, eelnevalt detekteerimata organismide genoomidest ja transkriptoomidest TMV-CP_{fly} tõendeid. Kasutatud järjestuste informatsioon on välja toodud lisas 13.

Algjärjestused:

- Viiruslikud – NCBI täielikult sekveneeritud viiruslike genoomide andmebaasist (Sayers *et al*., 2009) *Virgaviridae* sugukonna kõikide esindajate ning *Potyviridae* sugukonna bümoviiruste perekonna kõikide esindajate kattevalkude aminohappelised järjestused.
- Kärbseliste järjestused. BLAST otsingu *FBgn009799 Diptera´*de tasemel jäänud tulemused, mis filtreeriti korduste vähendamiseks järjestuse NCBI ID alusel. Lisaks määrati limiteerivateks kriteeriumiteks järjestuse pikkus (minimaalne pikkus 100 AH) ning otsingu vaste maksimalselt e-väärtuseks $1e^{-10}$.

Antud järjestused joondati vastavalt gruppidele MEGA programmi abil (versioon 6.06; Tamura *et al*., 2013) Muscle alamprogrammi abil, kasutades vaikimisi määratud parameetreid.

ProtTest programmiga (versioon 3.2; Darriba *et al*., 2011) ennustati parim fülogeneesi koostamise mudel (mudelite valik: JTT, LG, Dayhoff, WAG ning aluspuu koostamise meetodiks Maximum Likelihood (ML), ülejäänud parameetrid olid vaikimisi seatud). Parima mudeli valimiseks kasutati AIC väärtusi.

Fülogeneesipuu koostamine toimus MEGA programmi (versioon 6.06; Tamura *et al*., 2013) abil, ML fülogeneesipuu koostamise alamprogrammiga, kasutades ennustatud parimat mudelit koos vaikimisi määratud parameetritega, bootstrap väärtuste lisaarvutamisega (500).

Eelasjärjestused konstrueeriti FastML programmi abil (Ashkenazy *et al*., 2012) parameetritega: asenduste mudel LG, fülogeneetilise puu konstrueerimine ML meetodiga, kasutades optimiseetitud harude pikkuseid, gamma jaotust ning arvutades *joint reconstruction´*it. Indelite rekonstrueerimiseks kasutati samuti ML meetodit, eelistades ansestraalset indelit 0.5 tõenäosusliku limiteeriva parameetrina.

Konstrueeritud järjestuste baasil teostati BLAST otsingud (TBLASTN, versiooniga 2.2.31+; Sayers *et al*., 2009) putukate taksonoomia piires. Teiseks limiteerivaks parameetriks oli maksimaalne E-väärtus $1e^{-5}$. Otsingud teostati nt/nt, RefSeq-RNA, Ref-Seq-Genomics, Chromosomes, EST, GSS, HTGS, wgs, TSA andmebaaside vastu 2.05.2015. Tulemused lisati tabelisse 1 (lisa 2).

PROFIIL-HMM KASUTAMINE JÄRJESTUSTE ANALÜÜSIDES

Profiil-HMM kasutati uurimaks, kas varasemates töödes BLAST analüüsist efektiivsemaks määratud (lähemalt punktis 1.3.1) meetod aitab identifitseerida uusi järjestusi EVE´de kirjeldamisel.

Algjärjestusteks kasutati NCBI täisgenoomide andmebaasist (Sayers *et al*., 2009) tubaka mosaiigiviiruse kattevalku ning SUPERFAMILY andmebaasist (versioon 1.75; Gough *et al*., 2001) superperekonna 47195 *Drosophila melanogaster* geeni *FBgn0029799*. Lisaks kasutati (hmmsearch otsingu jaoks) eelnevalt koostatud tobamoviiruste mitmese järjestuse joondust (MSA; ing k *multiple sequence alignment*), mille järjestused võetud NCBI täisgenoomide andmebaasist (Sayers *et al*., 2009), ning varasemalt BLAST otsinguga detekteeritud (E-väärtuse lävend $1e^{-10}$ ning järjestuse minimaalne pikkus 100 AH) kärbseliste valgujärjetuste baasil koostatud MSA´d. Järjestused joondati MEGA programmi (versioon 6.06; Tamura,

2013) Muscle alamprogrammiga ette määratud parameetritega ning joondused täpsustati Muscle (versioon 3.8.31; Edgar, 2004) *refine* käsuga.

Otsingud on teostatud HMMER programmi (versioon 1.9; Finn *et al*., 2011) vahendusel 10.05.2015. Kasutatavad andmebaasid:

- phammer – valgujärjestuse otsing valgujärjestuse andmebaasi vastu. Andmebaasideks NR, RefSeq, UniProtKB, Pfam ning struktuuride baasil SwissProt ja PDB.
- hmmscan – valgujärjestuse otsing profiil-HMM andmebaaside vastu. Andmebaasideks Pfam, TIGRFAM, Gene3D ja SUPERFAMILY.
- hmmsearch – valgujärjestuste joondus või profiil-HMM otsing valgujärjestuste andmebaaside vastu. Andmebaasideks NR, RefSeq, UniProtKB, Pfamseq, SwissProt ja PDB.

Madala E-väärtusega (suurem kui $1e^{-8}$) detekteeritud valkudele teostati LOMETS programmi (versioon 4.0, viimane uuendus 07.09.2014; Wu ja Zhang, 2007) abil struktuuride ennustus määramaks vaste usaldusväärsust alternatiivse meetoditega.

STRUKTUURI KASUTAMINE FÜLOGENEESIPUU KONSTRUEERIMISEL

Eelnevalt on leitud (lähemalt punktis 1.3.2.), et struktuuri kaasamine MSA ning fülogeneesi konstrueerimisel, võib aidata just kaugemalt seotud olevate organismide lahknemiste probleemide lahendamisel. Antud juhul prooviti fülogeneesipuid täpsustada kasutades teadaolevaid, antud töös käsitlevate viiruste, kattevalkude struktuure.

Viiruslikud andmed on võetud SUPERFAMILY (versioon 1,75; Gough *et al*., 2001) andmebaasist superperekonnale 47195 vastavate NCBI andmbaasi kuuluvad viiruslikud järjestused (11.12.2014) limiteeriva E-väärtuse parameetriga (maksimaalne väärtus $1e^{-8}$). Eukarüootsed järjestused on võetud NCBI BLAST´imiste tulemusena *Drosophila melanogaster* geeni *FBgn0029799* valgujärjestuse suhtes, eristavaks tunnuseks järjestuse NCBI ID, limiteerivateks piiranguteks minimaalne järjestuse pikkus 100 AH ning BLAST otsingu limiteerivaks parameetriks E-väärtus $1e^{-10}$. Kasutatud järjestuste informatsioon on välja toodud lisas 13.

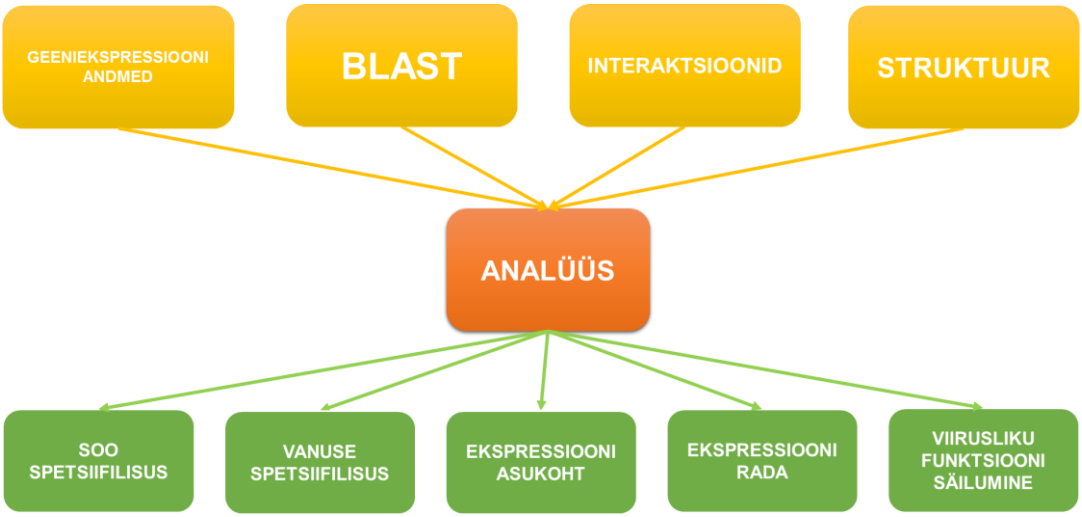
Järjestused sisestati programmi Promals3D (Pei *et al*., 2008) koos teadaolevate valkude struktuuridega (PDB ID: 1EI7, 1VTM, 1RMV, 3PDM, 1CGM, 4GQH; Berman *et al*., 2000). Järgnevalt kasutati Jalview programmi (versioon 2.8.2, viimase uuendusega 15.12.2014;

Waterhouse *et al*., 2009) kõdususe eemaldamiseks (ing k *reduced redundancy*; väärtusega 98%) ning joondust täpsustati Muscle (versioon 3.8.91; Edgar, 2004) *refine* käsuga.

MEGA programmiga (versioon 6.06; Tamura *et al*., 2013) koostati Neighbour Joining (NJ) fülogeneetiline puu, kasutades boostrap meetodit, Poissoni asenduste mudelit, *uniform rates* ning paariviisilist *gap´*ide ja puuduva info deleteerimist.

2.2.3. Uuritava geeni funktsiooni määramine

Antud töö teiseks eesmärgiks oli määrata bioinformaatiliste meetoditega TMV-CP_{fly} geeni funktsiooni ning ekspresseerumise spetsiifika organismides. Selleks uuriti kiipide geenieksprессiooni andmeid *Drosophila melanogaster* organismi piires. See võiks anda esmaseid viiteid uuritava geeni soospetsiifilisuse, vanuse ning ekspressiooni asukoha kohta. Antud andmeid püütakse täiendada ning laiendada teiste organismide genoomide ja transkriptoomide abil, kasutades selleks BLAST otsinguid ning vastavate tulemuste kirjeldatud bioproovide andmeid.



Joonis 10. Uuritava geeni funktsiooni määramise tööskeem. Kollasega on märgitud erinevad andmetüübid, mida kasutatakse analüüsimisel ning rohelistega on märgitud need omadused, mida soovitakse antud uuringust teada saada.

Lisaks vaadati TMV-CP_{fly} **valgu** teadaolevaid interaktsioone teiste valkudega, kasutades selleks varem teostatud valk-valk interaktsioonide analüüside tulemusi, püüdes selle abil määrata, millises tingimustes võiks uuritav valk toimida. Lisaks uuriti TMV-CP_{fly} **geeni** interaktsioone mikroRNA´dega. Täpsemalt mikroRNA´sid, mis võiksid seonduda uuritava

geeniga ning uurida nende ülesandeid, leidmaks seoseid erinevate uurimistasemete tulemustes. Lisaks vaadati ka transkriptsioonifaktorite seondumist TMV-CP_{fly} **geenile**, andes viiteid geeni ekspresseerumisele.

Üldine tööskeem on välja toodud joonisel 10.

GEENIEKSPERSSIOONI (CHIP) ANDMED

Geenikespressiooni andmete uurimise eesmärgiks oli tuvastada TMV-CP_{fly} ekspresseerumise spetsiifilisust (sugu, vanus, kude). Antud andmed eksisteerivad vaid *Drosophila melanogaster´*i kohta, seega geenikespressiooni tulemused on ainult antud organismi põhised. Töö eesmärgiks oli uurida erinevaid andmebaase ning kokku koguda eelnevalt normaliseeritud andmed, katogoriseerides koe ning soo ja/või arengustaadiumi baasil.

- Bgee andmebaas (Bastian *et al*., 2008; 22.04.2015) on loodud geenikespressiooni mustrite vaatamiseks ja võrdlemiseks erinevate loomaliikide vahel. Otsing on teostatud märksõnaga „*CG15772*“ ning liigiga „*Drosophila melanogaster*“ liigi piires
- Genevestigator andmebaas (Hruz *et al*., 2008) on käsitsi kureeritud ja hästi kirjeldatud geenikespressiooni andmebaas, mis suudab visualiseerida antud andmeid haiguste, ravimite, kudede, rakuliinide ning genotüüpide kontekstis.
    - Genevisible andmebaasi (23.04.2015) otsingu märksõnaks „*CG15772*“ ning liigiks „*Drosophila melanogaster*“.
    - Genevestigator programm (versioon 4-36-0, uuendatud 23.03.2015), otsingumärksõnaks „*CG15772*“, liik *Drosophila melanogaster* ning platvormiks „Affymetrix Drosophila Genome 2.0 Array“, 23.04.2015 seisuga.
- FlyMine (Lyne *et al*., 2007; versioon 40.0, uuendatud 9.09.2014) on andmebaas, mis sisaldab kahetiivaliste organismide genoomseid ning proteoomseid andmeid. Otsing teostati 22.04.2015 märksõnaga „*CG15772*“.

BLAST´I ABIL EKSPRESSIOONIANDMETE ANALÜÜS

Geenikespressiooni andmetele lisaks uuriti BLAST otsingu baasil transkriptide tõendeid TMV-CP_{fly} ekspresseerumisest uuritavas organismis. BLAST otsing aitas laiendada organismide gruppi, kellel on tuvastatud tõendeid TMV-CP_{fly} geenikespressioonist. Lokaalsest andmebaasist võeti välja varem BLASTitud andmetest geeni *CG15772* valgujärjestuse vasted EST ning TSA andmebaasi (Sayers *et al*., 2009) alusel.

INTERAKTSIOONIDE ANDMED

Erinevate interaktsioonide andmete kasutamise eesmärgiks oli saada lisainformatsiooni uuritava geeni või valgu funktsioonist või ülesannetest ning ekspressioonitingimustest.

TMV-CP_{fly} (VALK) – VALK INTERAKTSIOONID

Kõigepealt uuriti *Drosophila* proteoomika andmebaase, teada saamaks, kas uuritavat geeni transleeritakse valguks. Selleks teostati Peptide Atlasest (koosneb mitmete organismide suurtest valkude mass-spektromeetria eksperimentide tulemustest; Desiere *et al*., 2006) otsing geeni *CG15772* baasil, 24.05.2015 seisuga.

Uurides TMV-CP_{fly} **valgu** hinnatud interaktsiooni teiste valkudega, võib see anda viiteid oluliste protsesside kohta, kus antud valk toimib. Selleks on teostatud erinevates andmebaasides otsing *Drosophila melanogaster* geeni *CG15772* baasil 21.04.2015. Teisi kärbseliste TMV-CP_{fly} interaktsioone ei ole kasutatud, kuna vastavad andmed puuduvad.

- DroID (*The Drosophila Interactions Database*; versioon 2014_10 uuendatud 2.10.2014 Yu *et al*., 2008) on spetsiaalselt *Drosophila* mudelorganismile loodud geeni ja valkude interaktsioonide andmebaas.
- BioGRID (versioon 3.3.123 uuendatud 1.04.2015; Stark *et al*., 2006) on avalik andmebaas, mis sisaldab geneetilist ja valkude interaktsioonide andmeid. Antud andmebaas sisaldab kärbselistest *Drosophila melanogaster*, *Drosophila persimilis*, *Drosophila yakuba* ning *Drosophila sechellia* interaktsioonide andmeid, kuid TMV-CP_{fly} valgu interaktsioonide andmed esinevad ainult *D. melanogaster* organismi jaoks, teistel kas andmed puuduvad või on alles ülevaatamise järgus.
- IntAct (versioon 4.1.5-SNAPSHOT; Orchard *et al*., 2014) andmebaas sisaldab molekulaarsete interaktsioonide andmeid, mis on saadud kirjanduste või kasutajate enda lisatud andmetest.
- Mentha (Calderone *et al*., 2013) on käsitsi kureeritud valk-valk interaktsioonide andmebaas.
- MINT (Licata *et al*., 2012) on molekulaarsete interaktsioonide andmebaas, mis sisaldab vaid eksperimentaalselt kinnitatud valk-valk interaktsioone.
- PSICQUIC View (versioon 1.4.5; Del-Toro *et al*., 2013) koondab enda alla erinevate meetodite baasil erinevates andmebaasides sisaldavaid valk-valk interaktsioonide andmed.

TMV-CP_{fly} (GEEN) – mikroRNA INTERAKTSIOONID

Uurides mikroRNA´sid, mis seonduvad TMV-CP_{fly} **geeniga**, võib leida seoseid uuritava geeni ekspressioonimustrite kohta. Selleks uuriti andmebaase DrioID (versioon 2014_10 uuendatud 2.10.2014 Yu *et al*., 2008) ning TargetScan (versioon 6.2, uuendatud juuni 2002; Grimson *et al*., 2008), kus teostati otsing *D. melanogaster´*i geeni *CG15772* baasil 21.04.2015.

TMV-CP_{fly} (GEEN) – TRANSKRIPTSIOONIFAKTOR INTERAKTSIOONID

Lisaks uuriti transkriptsioonifaktoreid, mis seonduvad *D. melanogaster´*i *CG15772* **geeniga**. See võib sarnaselt mikroRNA seondumistega anda viiteid tingmustele, millal on uuritava geeni ekspressioon aktiivne. Selleks uuriti andmebaasi DrioID (versioon 2014_10 uuendatud 2.10.2014 Yu *et al*., 2008). Otsing teostati *CG15772* geeni baasil 21.04.2015.

VALGU STRUKTUURSE INFO KASUTAMINE FUNKTSIOONIDE UURIMISEL

Uurimaks, kas kärbselistes eksisteeriv TMV-CP_{fly} valk võiks olla säilitanud RNA sidumise ning filamentse struktuuri loomise funktsiooni, võrreldi TMV-CP_{viral} ning TMV-CP_{fly} ennustatud valgustruktuure, eriti just oluliste interaktsiooniregioonide konserveeruvust. Lisaks võrreldi kirjandusest eksperimentaalselt saadud tubaka mosaiigiviiruse RNA-VALK ning VALK-VALK interaktsioonideks olulisi aminohappeid eksperimentaalselt saadud TMV-CP_{viral} struktuuriga ning kärbselistes ennustatult saadud TMV-CP_{fly} valgustruktuuriga.

KASUTATUD STRUKTUURID

Kasutatud valkude struktuurid (PDB ID; Berman *et al*., 2000): 1EI7, 1VTM, 1CGM, 1RMV, 3PDM ning 4GQH (Bhyravbhatla *et al*., 1998; Pattanayek ja Stubbs, 1992; Wang ja Stubbs, 1994; Wang ja Stubbs, 1993; Tewary *et al*., 2011; Li *et al*.2013).

LOMETS programmi (versioon 4.0, viimane uuendus 07.09.2014; Wu ja Zhang, 2007) kasutati valkude struktuuride ning parima 10 valgustruktuuri mudeli ennustamiseks. LOMETS kasutab ennustamiseks erinevaid programme (täpsemalt lisas 3). Algjärjestused:

- *Drosophila melanogaster* geen *FBgn0029799* (SUPERFAMILY andmebaas, versioon 1.75, SF 47195, regiooni osa; Gough *et al*., 2001).

- *Glossina morsitans morsitans* (NCBI Accession ID CCAG010017880.1, regioon 9646-10065; Sayers *et al*., 2009).

STRUKTUURI KONSERVEERUVUSE ANALÜÜS

The ConSurf Server (Landau *et al*., 2005) – valkude MSA alusel konserveerumisastmega ühise struktuuri loomine eelnevalt joondatud järjestuste baasil. Ette antud valkude struktuurideks oli viiruste puhul 1VTM (PDB ID; Berman *et al*., 2000), kärbseliste puhul LOMETS´a programmiga (Wu ja Zhang, 2007) ennustatud struktuurid. Arvutuslikuks meetodiks määrati Maximum Likelihood ning evolutsiooniliseks asenduste mudeliks LG.

INTERAKTSIOONIDEKS OLULISTE AMINOHAPETE OTSINGUD KIRJANDUSE BAASIL

Kirjandusest võetud TMV-CP RNA-valk ning valk-valk interaktsioonideks vajalikud aminohapped ning nende esiletoomine antud struktuurides. Võrreldi antud aminohappeid TMV-CP ja kärbselistes esineva valgu vahel. Tulemusi näidati eelnevalt koostatud MSA baasil BioEdit programmi (versioon 7.2.5; Hall, 1999) abil, millele on lisatud andmekaeve tulemused.

STRUKTUURI KONSERVEERUVUSE NÄITAMINE JOONISENA

ConSurf´i (Landau *et al*., 2005) tulemused avati PyMol (versioon 1.1; https://www.pymol.org/) programmiga ning kasutades ConSurf´i loodud python skripti, värviti PyMol programmis valgustruktuurid konserveeruvuse alusel, mille arvutas ConSurf programm.

STRUKTUURIDE VÕRDLUSED

Swiss-PdbViewer programmi (versioon 4.1.0; Guex ja Peitsch, 1997) kasutati valkude struktuuride koosvaatamiseks, kasutades automaatset sobitavust (*automatic fit*), tavalist sobitamist (*fit*) ning maagilist sobitamist (*magical fit*), saades parima kahe struktuuri sobivuse. Selle määramiseks kasutati minimaalseimat RMS väärtust, mida oli võimalik saavutada.

2.3. TULEMUSED

2.3.1. TMV-CP kärbselistes

Uurides TMV-CP$_{fly}$ esinemist organismides, on võimalik täpsustada toimunud integratsiooni, mis on eelnevalt hinnatud 60-250 MAT *Schizophora* kärbseliste eellasesse (Kirsip, 2013) ning laiendada detekteeritud peremeesorganismide ringi. Lisaks uurides TMV-CP$_{fly}$ geeni ümbritsevat ala sekveneeritud organismide genoomides, saab sünteensuse alusel hinnata toimunud integratsioonisündmuste kordi. Lisaks oli teiseks uurimiseesmärgiks testida alternatiivsed EVE´de tuvastamise meetodeid ning uurida, kas strktuuri kaasamine võib fülogeneesi parandada, täpsemalt sügavamate lahknemiste juures.

BLAST ANALÜÜS EVE´DE JA ÜMBRITSEVATE GEENIDE TUVASTAMISEKS

Uurimaks TMV-CP$_{fly}$ esinemist erinevatest organismidest, teostati genoomi ning transkriptoomi baasil BLAST analüüsid *Drosophila melanogaster CG15772* valgujärjestuse alusel. Lisaks uuriti läheduses asuvaid geene, teada saamaks, kas ümbritsev regioon võiks olla sekveneeritud ning vastaval juhul regiooni sünteensuse määra. See võiks näidata, kas on toimunud üks intergratsioon kärbseliste eellasesse, mitu integratsiooni erinevatesse regioonidesse või toimunud mitmeid ümberkorraldusi. Seda saab uurida hetkel vaid genoomi (või genoomse) informatsiooni baasil täisgenoomidel. Tuvastatud transkriptid annavad viiteid eksisteeriva transkriptsiooni kohta. Lisaks uurides nii genoomi kui ka transkriptoomi andmeid, peaks olema võimalik laiendada organismide ringi, kes uuritavat TMV-CP$_{fly}$ geeni omada võiks, kuna erinevatel organismidel võib olla kas ainult transkriptoomi või teatud genoomi piirkondi sekveneeritud.

Tabelis 1 (lisa 2) on välja toodud organismid nii sugukonna (antud taseme tulemused on esitatud joonisel 11), perekonna kui ka liigi tasemel, kellel on uuritav TMV-CP$_{fly}$ ja/või seda ümbritsevaid geene (FlyBase ID: *IntS6*, *CG4075*, *lin-52*, *CG15771*) detekteeritud. Samuti on välja toodud organismide piires NCBI andmebaaside (Sayers *et al*., 2009) abil tuvastatud annoteeritud mRNA´de ja valkude ning kärbseliste sugukondade, perekondade ja liikide puhul bioprojektide arv iseloomustamaks nende organismide sekveneeritust ja annoteeritust. Tabelis 1 on tumedamas kirjas välja toodud BLAST analüüsi käigus tuvastatud, kuid eelnevalt detekeetrimata, tulemused.

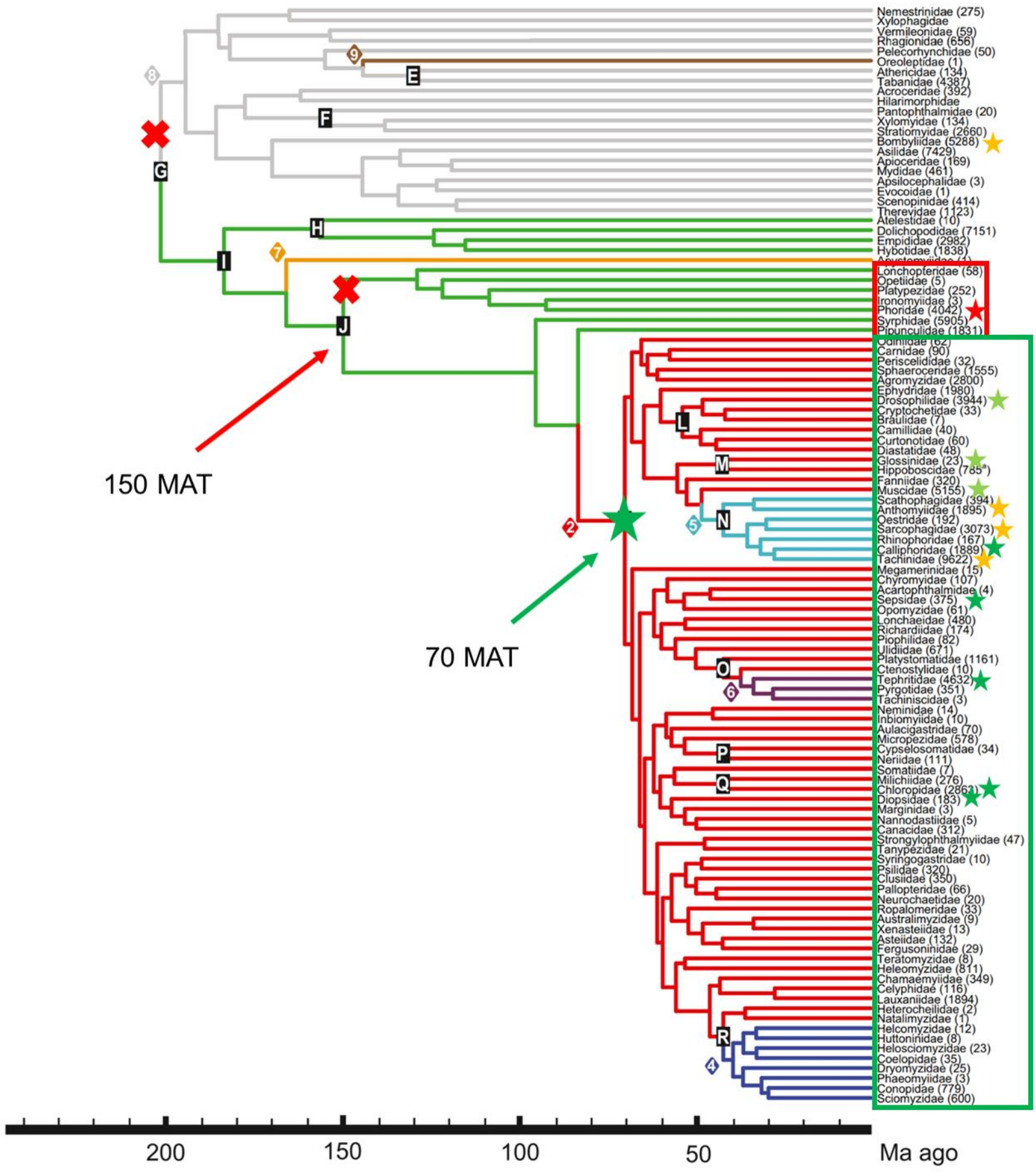
Bombyliidae on kärbselised, kes kuuluvad *Asiloidea* superperekonda ning on sirelastest varasemalt lahknenud. *Phoridae* sugukond kuulub sirelaste (*Aschiza*) *Platypezoidea* superperekonda. Kummaski sugukonnas on üks esindaja, kellelt on leitud vähemalt üks viiest

käsitlevast geenist. Mõlemal organismil puuduvad tõendid uuritava TMV-CP_{fly} geeni omamise kohta (tabel 1, lisa 2; joonis 11), kuigi *Phoridae´*del on TMV-CP_{fly} potentsiaalsed naabergeenid leitud genoomi (mitte transkriptoomi) andmetest.

Calyptratae kärbseliste alla kuuluvatest sugukondadest *Tachinidae*, *Sacrophagidae* ning *Anthomyiidae*, kellel kõigil on üks esindaja, puudub antud uuritav TMV-CP_{fly} geen (vähemalt hetkel esinevate andmete põhjal; tabel 1, lisa 2; joonis 11). Samas ülejäänud uuritavate geenide suhtes on tõendeid ainult transkriptide põhjal, mis on kohati puudulikud. Ülejäänud *Calyptratae* kärbselistel: *Muscidae*, *Calliphoridae* (mõlemad ühe esindajaga) ning *Glossinidae* (kuue esindajaga) on deteketeeritud uuritavat ning seda ümbritsevaid geene, seda nii genoomsete kui ka transkriptoomsete andmete abil.

Acalyptratae kärbsliste sugukondades *Sepsidae*, *Chloropidae* (mõlemal üks liik) ning *Diopsidae* (kahe liigiga) on uuritavat ning seda ümbritsevaid geene deketeeritud kõigil, peamiselt transkriptide, kuid osadel ka genoomsete tõendite alusel (tabel 1, lisa 2; joonis 11). *Tephritidae* sugukond on esindatud kolme perekonnaga: *Rhagoletis* (uuritav geen puudub), *Ceratitis* (uuritav ning ümbritsevad geenid olemas nii genoomsete kui ka transkriptoomsete tõenditega) ning *Bactrocera*, kellel kahel liigil on nii genoomide kui ka transkriptoomide tõendeid, ühel ainult genoomi ning kahel liigil puuduvad uuritava geeni tõendid (ülejäänud geenid on esindatud mRNA eksisteerimisega). *Drosophilidae* sugukond on kõige suuremal määral 32 esindajaga. Need, kellel antud geen puudub, on ülejäänud naabergeenid, nii palju kui neid leidub, tõendatud transkriptidega, samas need, kellel uuritav geen on tuvastatud, on see tõendatud peamiselt genoomsete andmetega, osadel juhtudel lisaks ka transkriptidega. Paksemas kirjas on märgitud need *Drosophila* liigid, kelle genoom on täielikult sekveneeritud (Flybase versioon FB2015_01; dos Santos *et al*., 2015).

Kõkkuvõtvalt on uuritavat TMV-CP_{fly} geeni tuvastatud kärbselistest ning kui arvestada kõiki täielikult sekveneeritud kärbselisi, siis uuritavat geeni ei detekteeritud ainult ühest organismist – *Megaselia scalaris* (enne *Schizophora´*de divergeerumist lahknenud sirelaste grupp) (joonis 11). Ülejäänud organismidel, kellel leiduvad ülejäänud TMV-CP_{fly} ümbritsevad geenid, kuid uuritavat TMV-CP_{fly} puudub, on hetkeseisuga teostatud genoomi või transkriptoomi uuringud alles väga alguse järgus ning suure tõenäosusega pole lihtsalt veel uuritavat TMV-CP_{fly} geeni/mRNA´d veel tuvastatud.



Joonis 11. BLAST analüüsil detekteeritud TMV-CP_{fly} leiduvus organismiti. Rohelise tähekesega on märgitud need kärbseliste sugukonnad, kellel on uuritavat geeni detekteeritud (heledama rohelisega on tuvastatud juba bakaulaureusetöös (Kirsip, 2013) ning tumadama rohelisega on uued tuvastatud sugukonnad) ning punasega need, kelle praeguste andmete seisuga see puudub. Oranzi tähekesega on märgitud need kärbseliste sugukonnad, kelle uuritavat TMV-CP_{fly} geeni ei detekteeritud, kuid ülejäänud naabergeenid on tõendatud ainult transkriptoomsete andmetega. Suure rohelise tähekesega on välja toodud organismide grupp, kellel uuritavat TMV-CP_{fly} geen olemas on, ning välja on toodud nende ligikaudne tekkeaeg (70 MAT). Punase ristiga on välja toodud need organismidegrupid, kellel uuritavat geeni ei leidu. Suure punase kastiga on välja toodud sirelased (*Achiza*) ning rohelise kastiga *Schizophora* kärbselised. Antud fülogeneesiuuringu baasil saab hinnata toimunud integratsiooniks 70-150 MA. Fülogenees on koostatud Wiegmann *et al*. (2011) poolt ning see on vähendatud uuritavate gruppide piires. Joonist on täiendatud töö autori poolt.

EELASJÄRJESTUSE BLAST ANALÜÜS

Kuna viiruslike järjestuste sarnasus kärbseliste omaga ei olnud BLAST analüüsiga detekteeritav, vaid oli tuvastatav ainult läbi HMM´ide SUPERFAMILY andmebaasi abil (Kirsip, 2013; mai, 2015 teostatud uued BLAST analüüsid), kasutati eellasjärjestust leidmaks teisi homolooge. Selleks rekonstrueeriti *Virgaviridae* ning kärbseliste (kõigi kärbseliste, *Acalyptratae* ning *Calyptratae* gruppide) eellasjärjestused. Antud järjestusi kasutati BLAST otsingutes, leidmaks uusi organisme, kus on TMV-CP_{fly} deteketeeritud kas genoomsete või transkriptoomsete andmete abil. Uued tulemused lisati tabelisse 1 (lisa 2) ning on märgitud geeni *CG15772* tulpa rohelises kirjas (*Drosophila pseudoobscura pseudoobscura* ning *Lucilia cuprina* transkriptid).

PROFIIL-HMM KASUTAMINE JÄRJESTUSTE ANALÜÜSIDES

Varasemalt on näidatud, et HMM meetodi kasutamine suudab tuvastada väiksema identsusega järjestusi, kui tavaline kahe järjestuse võrdlusel baseeruv BLAST analüüs. Uurimaks profiil-HMM otsingumeetodi kasutatavust uute EVE´de deteketeerimisel, jooksutati viirusliku kattevalgu (TMV-CP_{viral}) ning kärbseliste *CG15772* valgujärjestused HMMER programmist läbi, kasutades erinevate meetodite baasil otsinguid. Tulemused on skemaatiliselt välja toodud joonisel 12.

TMV-CP_{viral} ja tobamoviiruste MSA kasutamine otsingute algjärjestusena eesmärgiks oli määrata meetodi detekteerimisvõimet ning identifitseerida viiruslike järjestuste baasil uusi, BLAST analüüsiga detekteerimata, EVE´sid. Antud analüüsi tulemused on välja toodud joonisel 12. phammer (valgujärjestus valgujärjestuste andmebaaside vastu) otsing tuvastas ainult varasemalt detekteeritud viiruslikke järjestusi. hmmscan (valgujärjestuse profiil-HMM andmebaaside vastu) detekteeris TMV-CP valgudomeeni kolmes erinevas andmebaasis: pfam, Gene3D ja SUPERFAMILY. Kõik need sisaldavad erinevatest organismidest tuvastatud ning valgudomeenidesse klassifitseeritud andmeid, seega tõenäosus, et nende abil on võimalik detekteerida viiruslikku järjestusi eukarüootsetest organismidest, on suur. Hmmsearch (MSA valgujärjestuste andmebaaside vastu) detekteeris madala E-väärtusega ($5.6e^{-3}$) kärbselisi ning piiripealse E-väärtusega ($1e^{-4}$) kolme mesilase järjestust:

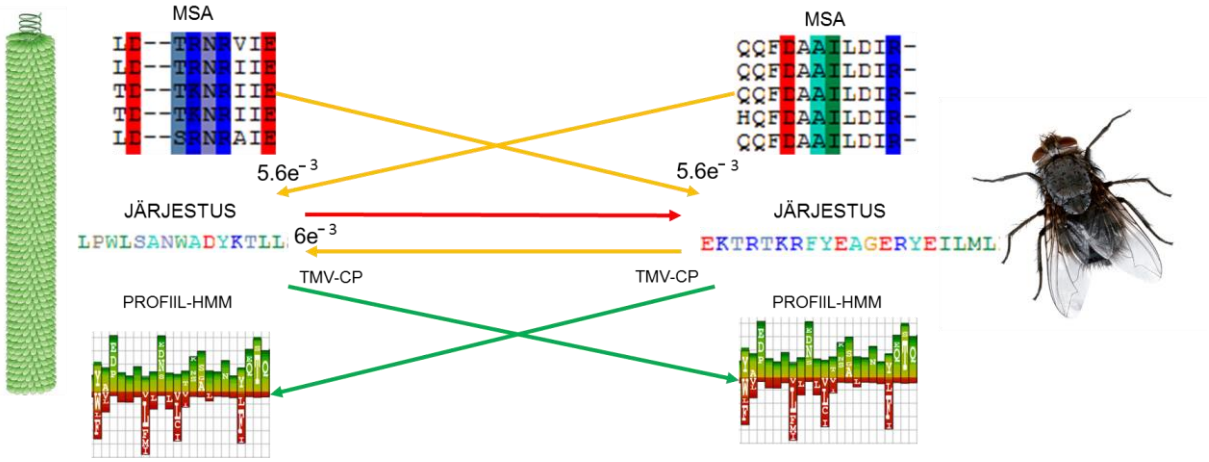
- *Apis dorsata* valgujärjestuse koodid XP_006617322.1 ja XP_006617321.1, millele vastab üks geen, NCBI ID LOC102673861 (*Apis dorsata* genoomiversioon 1.3).

- *Apis mellifera* valgujärjestuse kood XP_006557954.1, millele vastab BeeBase (Munoz-Torres *et al*., 2011) ID GB52320 (*Apis mellifera* genoomi versioon 3.2).

Lisaks detekteeriti kõrge E-väärtusega (~1e^{-22}) *Nicotiana tabacum´*i viiruslikku valku (NCBI ID CAB03628.1), mis on suure tõenäosusega viirusega infekteerunud lehe proovist saadud, kuna järjestuse andmebaasi annotatsiooni põhjal on koe tüübiks nakatunud leht.

Kolmele kirjeldatud mesilase järjestustele ennustati valgustruktuuri LOMETS programmi abil. Kõige kolme puhul esines ennustatud struktuurides üle 50 Z-skooriga mõned struktuurid, mis kõik vastasid tubaka mosaiigiviiruse kattevalgule. Antud Z-skoori võiks pidada piiripealseks – madalama Z-skooriga ennustusi ei soovitata eriti pidada õigeteks ning tunduvamalt kõrgemate skooridega (90 kanti) võiks pidada suure tõonäosusega õigeteks ennustusteks. Kuna antud järjestuse puhul esineb liiga suur ebakindlus, oleks soovitatav see edasistest analüüsidest välja jätta.

Kasutades algjärjestusteks kas ***CG15772*** **valgujärjestust** või antud järjestuse baasil koostatud eukarüootsete organismide **(kärbseliste) MSA´d**, detekteeriti peamiselt juba varem kinnitatud organismide geene (Kirsip, 2013), lisaks ka madala E-väärtusega (6e^{-3}) tobamoviirust. hmmscan programmi abil jooksutatud taksonoomilise piiranguteta otsing andis viiruste baasil teostatud otsingule identsed vasted, seega saab teoorias selle meetodiga detekteerida EVE´sid või annoteerida geene, kasutades eukarüootsete organismide valgujärjestusi. hmmsearch detekteeris sarnaselt viiruslike MSA baasil piiripealse E-väärtusega *Apis* perekonna esindajate järjestusi (1e^{-5}) ning madala E-väärtusega tobamoviiruseid (5.6e^{-3}). Antud tulemusi kirjeldab joonis 12.



Joonis 12. HMM analüüside tulemused kärbseliste ja viiruslike organismide baasil. Välja on toodud kolm otsingutüüpi: MSA valgujärjestuste andmebaasi vastu, valgujärjestus valgujäjestuste andmebaasi vastu ning valgujärjestus profiil-HMM andmebaaside vastu. Roheliste nooltega on märgitud positiivsed otsingutulemused ning punastega negatiivsed. Oranžide nooltega on välja toodud väga madala E-väärtusega tuvastatud tulemused. Joonis on koostatud teostatud HMM analüüside baasil, 23.05.2015.

TEADAOLEVATE STRUKTUURIDE ABIL TEOSTATUD FÜLOGENEESIANALÜÜSID

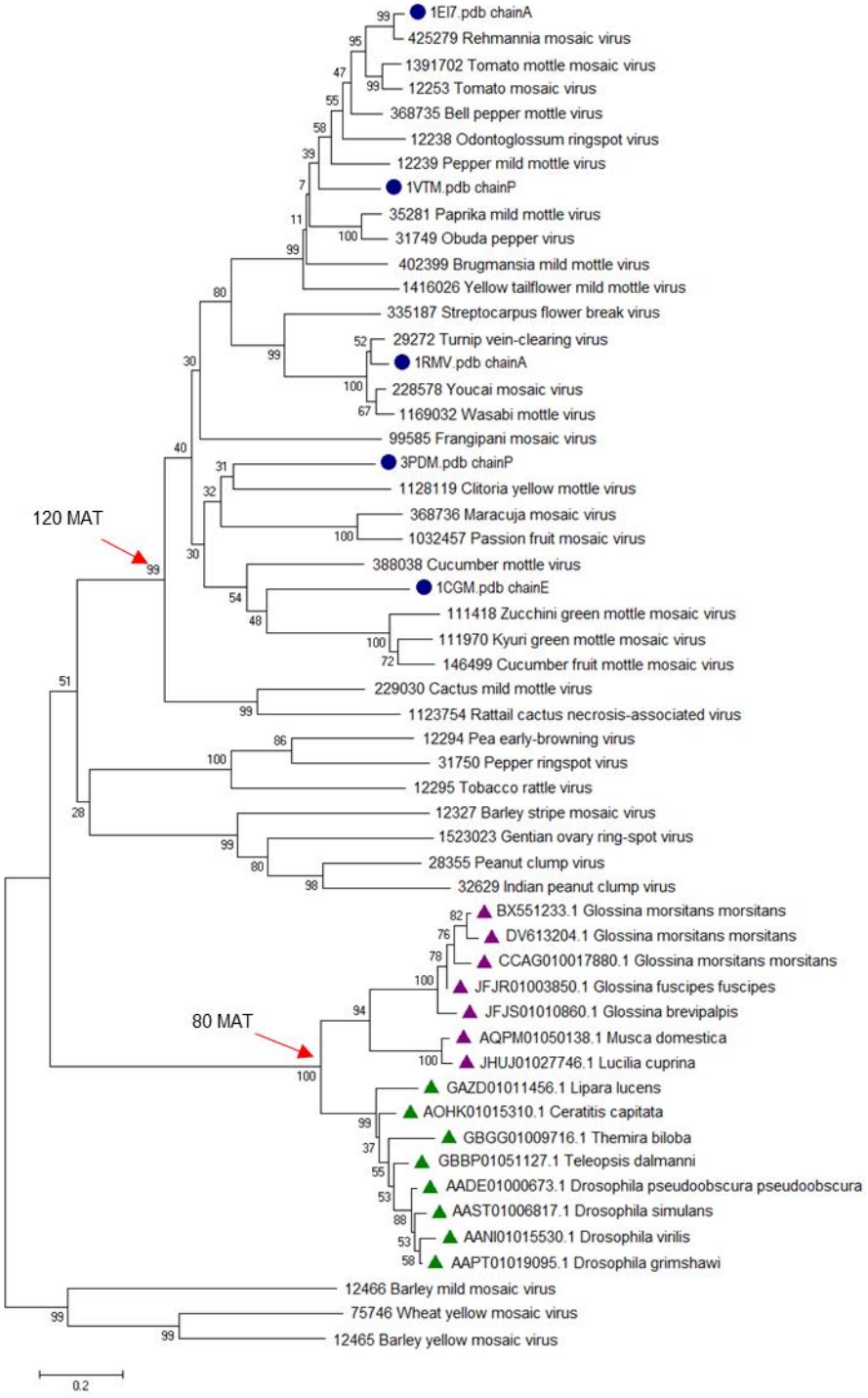
Eelnevalt on näidatud, et strutkuuri kasutamine MSA loomisel, võib aidata kaugelt lahkenud organsimide fülogeneesis tekkinud probleeme lahendada (lähemalt punktis 1.3.3). Uurides kirjeldatud viisi kasulikkust EVE´de uurimisel, kasutati TMV-$CP_{viral\ ja\ fly}$ MSA loomiseks teadaolevaid *Virgaviridae* tobamoviiruste struktuure.

TMV-CP geeni sisaldavate viiruste (*Virgaviridae* perekonnad tobamo-, tobra-, peklu- ning hordeiviirused ning *Potyviridae* bümoviirused) ning kärbseliste (BLAST otsingute tulemuste) järjestuste baasil konstrueeriti uus fülogeneesipuu (joonis 13), mille ehitamiseks kasutati teadaolevaid TMV-CP_{viral} valgu struktuure. Antud puul on struktuurid märgitud sinise täpiga ja kärbselised kolmnurkadega (rohelised *Acalyptratae* ja lillad *Calyptratae*) ning on juuritud *Potyviridae* esindajate baasil.

Antud puu puhul jäävad kärbselised viiruste keskele, esinevad perekonnad on monofüleetilised (seal hulgas kärbseliste jaotuvus) ning need lahknemised on toetatud tugeva boostrap väärtustega, samas sügavamad lahknemised on madalate boostrap väärtustega. Analoogselt samade andmetega, kuid struktuuri arvestamata, koostati teine fülogeneesipuu (lisa 4), võrdlemaks kirjeldatud meetodi efektiivsust. Struktuuri arvestamata fülogeneesipuul esinevad samuti viiruste perekonnad ning kärbselised monofüleetiliselt, mis on teostatud kõrgete boostrap väärtustega. Kaugemate lahknemiste puhul võib struktuuri mittearvestamisel saada isegi kvaliteesema tulemuse nende andmete baasil. Samas tuleks arvesse võtta ka seda, et teadaolevad struktuurid ei kata ühtlaselt käsitletud viiruste sugukondi, mis kindlasti kallutab tulemusi.

KOKKUVÕTE TMV-CP_{fly} ESINEMISEST KÄRBSELISTES

Uurides TMV-CP_{fly} ning seda ümbritsevaid geene, on teada saadud, et toimunud võiks olla üks integratsioon *Schizopora* eelastesse. Nendest eelnevalt lahknenud sirelastel (*Aschiza*), kelle vanuseks on saadud 83-129 MA (Gaunt ja Miles, 2002; Bertone *et al*., 2008; Wiegmann *et al*., 2011), uuritavat geeni ei eksisteeri, vähemalt käesolevate andmete baasil (joonis 11, punaste kastiga märgitud kärbseliste sugukonnad). Samuti ei ole TMV-CP_{fly} detekteeritud ka veel varem lahknenud *Bombyliidae* kärbselistes. Seega võiks hinnata integratsioonisündmuse toimumiseks 80-129 MAT. Teades, et tobamoviiruste vanuseks hinnatakse 120 MA (Stobbes *et al.,* 2012) ning TMV-CP_{viral} eksisteerib ka teistes *Virgaviridae* viirustes, mis teeks TMV-CP_{viral} veel vanemaks, on tõenäoline ülekande suund viirustelt peremeesorganismidesse.



Joonis 13. TMV-CP viiruslike ning eukarüootsete järjestuste baasil konstrueeritud fülogeneetiline NJ puu (Poissoni korrektsiooni mudeli baasil paariviisiliste *gap´*ide deleteerimisega), mis on juuritud bümoviiruste baasil. All on märgitud aminohapete asenduste arvu positsiooni kohta (0.2). Antud analüüsis osales 54 järjestust 217 informatiivse positsiooniga. Siniste täppidega on märgitud PDB (Berman et al., 2000) struktuuride järjestused, kolmnurkadega eukarüootset päritolu järjestused (rohelised kolmnurgad *Acalyptratae* kärbselised ning lillad kolmnurgad viitavad *Calyptratae* kärbselistele). Tobamoviiruste teket peetakse 120 MAT (Stobbes *et al*., 2012) ning *Acalyptratae* ja *Calyptratae* kärbseliste lahknemist 80 MAT (Wiegmann *et al*., 2003).

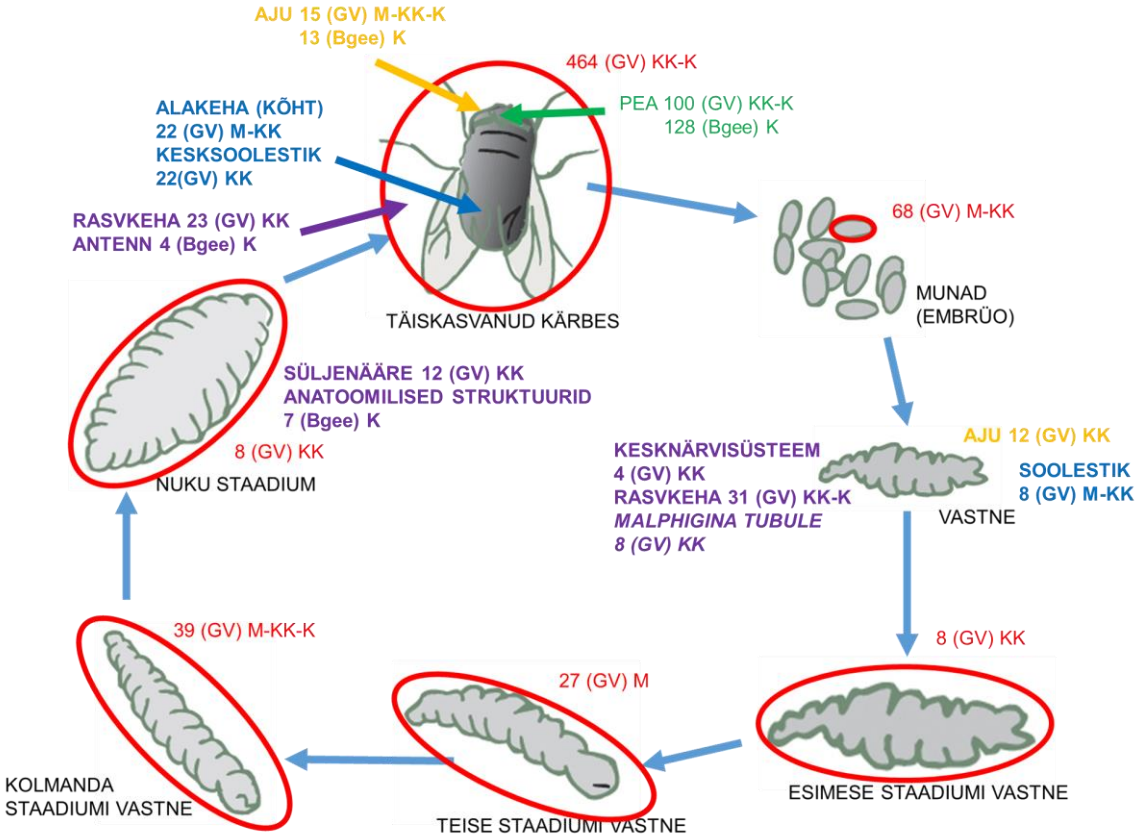
EVE´de deteketeerimise uuritud alternatiivsetest meetoditest osutus kõige efektiivsemaks HMM analüüsid, tuvastades viirusliku järjestuse alusel eukarüootsetele organismidele kuuluvaid järjestusi. Antud tulemusi BLAST analüüsides ei tuvastatud.

2.3.2. TMV-CP ekspressiooni andmed kärbselistes

Antud töö üheks eesmärgiks oli uurida TMV-CP_{fly} funktsiooni ning ekspressioonimustrit kärbselistes. Selleks uuriti erinevate geenieksрessiooni andmebaaside abil *Drosophila melanogaster* geeni *CG15772,* keskendudes peamiselt soospetsiifilisuse, vanuse ning koe tüübi identifitseerimisele. Lisaks uuriti TMV-CP_{fly} valgu interakteerumist teiste valkudega ning uuritava geeni interakteerumist mikroRNA´de ning transkriptsioonifaktoritega. Antud seosed kirjeldavad TMV-CP_{fly} ekspressioontingimusi, mis omakorda võiks anda vihjeid geeni mõjust organismile.

GEENIEKSPRESSIOONI TULEMUSED

Bgee, Geneinvestigator ning FlyMine ekspressioonide tulemuste koosuurimisel (tabel 2, lisa 5; keskendudes keskmisele või kõrgemale ekspressiooni määrale; kokkuvõte joonisel 14) on näha TMV-CP_{fly} ekspressiooni nii emas- kui ka isasorganismis erinevatel arenguetappidel (embrüo, larva, nukk ja täiskasvanud staadiumid). Igas arenguetapis on kõige rohkemalt uuritud ekspressiooni terve organismi piires (joonisel 14 punaselt välja toodud informatsioon). Enim on erinevate kudede geenieksрessiooni TMV-CP_{fly} puhul uuritud täiskarvanud staadiumis. Kõrgelt ekspresseerunud kudedest on arvukalt esindatud aju (13 *probeset* Bgee) ning pea (100 *probeset* GV ning 128 *probeset* Bgee). Antud juhtudel on probeset defineeritud kui ühel kiibil teostatud analüüs. Keskmise või keskmisest kõrgema ekspressioonitasemetega paistavad silma täiskasvanud staadiumis soolestik ja rasvkeha ning kui arengustaadiume ei arvestata või ei ole uuringu piires välja toodud, siis on kõrgelt ekspresseerunud aju, rasvkeha, pea, munasari ning keskmise ekspressioonitasemega silm, alakeha, *Malphigian tubule*, kesksool, süljenäärmed, testis ning emasorganismidel esinev *spermathecum*. Seega on TMV-CP_{fly} transkriptid ekspresseeritud mitmetes kudedes ja mitmes erinevas arengustaadiumis.



Joonis 14. Kärbseliste geenieskpressiooni tulemused, mis on välja toodud staadiumite kaupa. Joonisel on kujutatud suurmate eksperimentide arvuga tulemused. Punane tekst väljendab terve organismi piires tehtud katset, roheline pea, oranž aju, sinine soolestik ja alakeha ning lillad ülejäänud kudede piires teostatud proovid. Igale proovile toodud proovide arv, andmebaas ning ekspressiooni tase (M – madal, KK – keskmine, K – kõrge). Joonis on koostatud töö autori poolt. Pildid on võetud *Digital Insect Collection* koduleheküljelt *Common green bottle fly* alamlehelt (7.05.2015).

BLASTIMISE TULEMUSED EKSPRESSEERUMISMUSTRI AVASTAMISEKS

BLAST´i transkriptoomi andmete otsingud võimaldavad laiendada uuritavate peremeesorganismide ringi ning kasutades transkripte kirjeldavaid bioproove, on võimalik kaudselt määrata mRNA ekspressioonimustrit. Selleks uuriti TMV-CP_{fly} transkriptide olemasolu andmebaasides, kasutades BLAST´imise tulemusi täiendavaid bioproovide andmed, milles keskenduti soole, vanusele ning koele. Tabel 3 (lisa 6) sisaldab EST ning tabel 4 (lisa 7) TSA andmebaasi tulemusi. Kokkuvõtvalt on näha geenieksprессiooni esinemist meesorganismides, samas kui emasorganisme sisaldavaid bioproove ei tuvastatud, seega emasorganismides TMV-CP_{fly} ekspresseerumise kohta ei saa antud tulemuste baasil järjeldusi teha. Täpset aega, millal antud geeni ekspressioon toimub, ei suudetud samuti käesolevate

andmete baasil määrata. Kudedest esinesid ülekaalukalt testis, pea, rasvkeha ning seedekulgla (seda nii *Drosophila melanogaster* bioproovide ja ekspressiooniandmete kui ka teiste organismide bioproovide andmete järgi).

INTERAKTSIOONIDE TULEMUSED

Saamaks teada, mis radades TMV-CP_{fly} valk võiks toimida, on kasulik uurida valk-valk interaktsioone, keskendudes just usaldusväärsetele interaktsioonidele. Lisaks võib geenieksprессiooni toimumise hetke kohta anda viiteid TMV-CP_{fly} geeniga seonduvad transkriptsioonifaktorid ning mikroRNA´d, viidates tingumustele, millal toimuvad geeni ekspressioonis muudatused.

TMV-CP_{fly} (VALK) – VALK INTERAKTSIOONID

Proteoomika andmebaas Peptide Atlas (Desiere *et al*., 2006) näitas, et uuritavat geeni *CG15772* transleeritakse valguks, ekspressiooniga peas ning antud valku on detekteeritud nii tsütoplasma kui ka membraani fraktsioonis Brunner *et al*. (2007) eksperimendi tulemuste baasil.

Erinevate valk-valk interaktsioonide uurimistest *Drosophila melanogaster´*i puhul on leitud 11 valku, mis interakteeruvad *CG15772* valguga (tabel 5, lisa 8). Nendest on kõige kõrgema interaktsiooni skoori saanud (interaktsioonidest lähemalt lisas 9) algse analüüsi teostanud autorite ,Giot *et al*. (2003), poolt *l(3)03670*, skooriga 0.98 (skaalal 0-1). Algse autori poolt teised kõrge skoori on saanud *DNA polümeraas δ* (0.61) ning *CG1316* skooriga 0.56. Teiste andmebaaside (IntAct, Mentha, DroiD) skoorid varieeruvad vähem ning ei kirjelda nii üheselt tugevaid suhteid käsitlevate valkude vahel. BioGRID andmebaas sisaldab ka teiste *Drosophila´*de valkude interaktsioonide andmeid, kuid TMV-CP_{fly} interaktsioone ei ole praeguseks hetkeks üle kureeritud ning neid andmeid ei ole võimalik kätte saada.

TMV-CP_{fly} (GEEN) – miRNA INTERAKTSIOONID

Ennustatult interakteerub uuritava TMV-CP_{fly} geeniga 6 mikro-RNA´d (tabel 6, lisa 10). mir-124 ning mir-4 seondumiskohad *Drosophila* kärbeste TMV-CP_{fly} geenil on tugevalt konserveerunud, esinedes kõigis 12 täielikult sekveneeritud liigis (*D. simulans, D. sechellia, D. melanogaster, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. mojavensis, D. virilis, D. grimshawi*). Mõlemad mikroRNA´d ekspresseeruvad

embrüonaalses ning larva esimeses staadumis. mir-124 on seotud närvirakkude arengu kontrollimisega. Ülejäänud mikroRNA´del toimub seondumised TMV-CP_{fly} geeni piirkonda ainult osades *Drosophila* liikides.

TMV-CP_{fly} (GEEN) – TRANSKRIPTSIOONIFAKTORI INTERAKTSIOONID

DroID andmebaasi järgi on TMV-CP_{fly} geenil *Drosophila melanogaster* organismis kahe transkriptsioonifaktori seondumiskohad: *Cad* (FlyBase geen FBgn0000251) ja *Kr* (FlyBase geen FBgn0001325). *Cad* transkriptsioonifaktor (geen *caudal*) on peamiselt seotud kärbselise embrüonaalse arenguga: anterioorse/posterioorse telje spetsifikatsioonis, *Malphigian tubule´*i, tagasoole morfogeneesis, blastodermi segmentatsioonis, suguelundite plaadi formeerumises ning arengus (lähemalt lisas 11). *Kr* transkriptsioonifaktor (geen *Kruppel*) on samuti seotud peamiselt kärbseliste embrüonaalse arenguga: tiibade plaadi, silma, anterioorse ja posterioorse telje arengus ning *Malphigian tubule´*i morfogeneesis (lähemalt lisas 11).

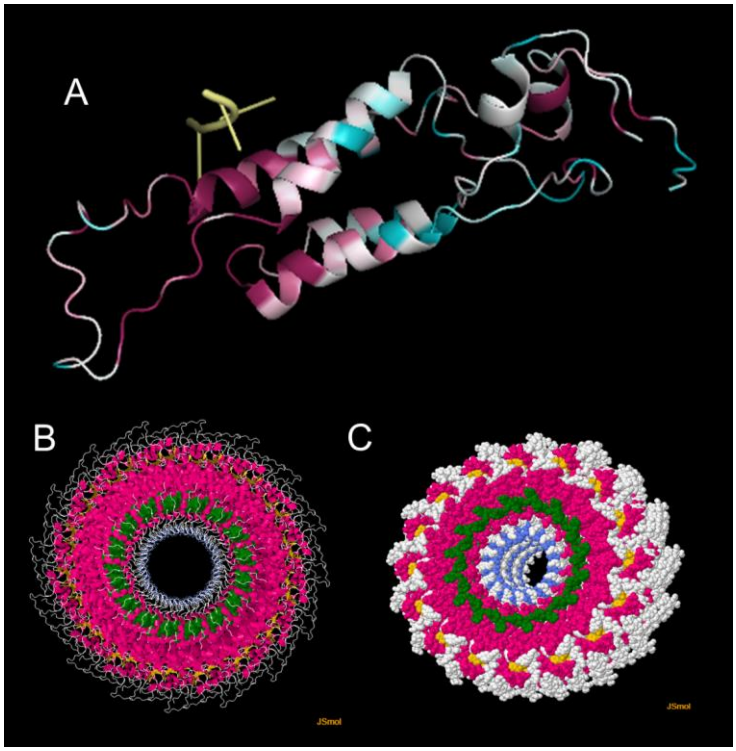
VALGU STRUKTUURSE INFO KASUTAMINE FUNKTSIOONIDE UURIMISEL

Võrreldes viiruste TMV-CP_{viral} ja kärbseliste ennustatud TMV-CP_{fly} valgustruktuure, rõhudes konserveeruvusele, võib proovida teha järeldusi kärbselistes eksisteeriva TMV-CP_{fly} valgu kohta. Keskendudes just neile omadustele, mida antud kattevalk viirustes omab – RNA sidumine ning filamentse struktuuri loomine, on võimalik teha järeldusi, kas kärbselistes eksisteeriv TMV-CP_{fly} on säilitanud oma algse, viirusliku funktsiooni, või on omandanud täiesti uue funktsiooni.

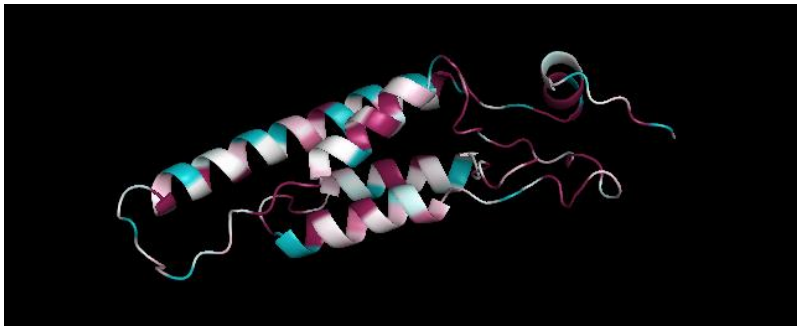
STRUKTUURIANALÜÜSIDE TULEMUSED

Järgnevad joonised näitavad ConSurf programmi (Landau *et al*., 2005) abil määratud konserveeruvust vastavalt viirustel ja kärbselistel (joonis 15 pilt A ja joonis 16). Mida lillam regioon on, seda konserveerunum ning mida sinisem, seda varieerunum piirkond on. Joonisel 15 on lisaks näidatud ka filamentset struktuuri ning üksikute subühikute paikemist suuremas stuktuuris. Samuti on välja toodud RNA paiknemine (pilt A – kollane, pilt B ja C - roheline). Viirustel on näha RNA seondumispiirkonnas esinev konserveeruvus (subühikul nii all kui üleval), mis viitab suurema struktuuri moodustumisele ning tänu sellele peab siduma RNA´d mõlemal pool. Lisaks esineb veel mõningad konserveerunumad piirkonnad, mis võivad olla

seotud valk-valk interaktsioonidega. Kärbselistel esineb tunduvamalt väiksem konserveeruvus ning see ei ole lokaliseerunud põhistruktuuride (ing k *core structures*) juurde, vaid varieerunumatesse lingude piirkondadesse.

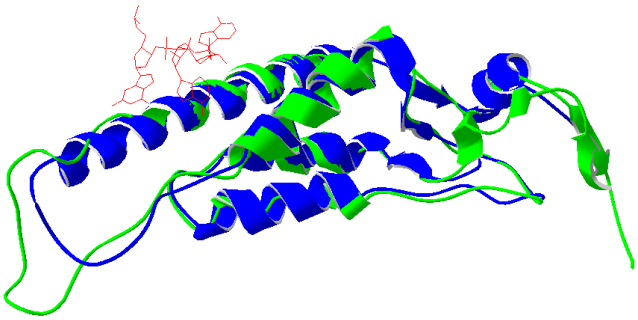


Joonis 15. Viirusliku TMV-CP_{viral} struktuur. Pildil A on välja toodud ConSurf´i programmi 1VTM tulemused, mis on Värvitud PyMol programmi abil ConSurf´i määratud konserveeruvuse alusel. Mida punasem, seda konserveerunum järjestus, mida sinisem, seda vähem conserveerunum järjestus. RNA on värvitud kollaselt. Joonis on koostatud PyMol (versioon 1.1; https://www.pymol.org/) programmiga. Piltidel B ja C on välja toodud PDB (Berman et al., 2000) andmebaasis loodud 1VTM filamentse struktuuri pildid, kus RNA on välja toodud rohelisena, alpha-heeliksid roosadena, beeta-lehed kollasena ning lingud valge ja sinisena.



Joonis 16. ConSurf´i programmi kärbseliste (nii *Calyptratae* kui ka *Acalyptratae*) tulemus. Värvitud on PyMol programmi abil ConSurf´i määratud konserveeruvuse alusel. Mida punasem, seda konserveerunum järjestus, mida sinisem, seda vähem conserveerunum järjestus. Joonis on koostatud PyMol (versioon 1.1; https://www.pymol.org/) programmiga.

Joonisel 17 on välja toodud viiruste teadaoleva kattevalgu struktuuri (PDB ID 1VTM) võrdlus ennustatud kärbselisete omaga. Joonisel on näidatud ka RNA paiknemine struktuuride suhtes (punane). Näha on, et peamised neli alpha-heeliksit ning lingude üldine ehitus on üldiselt sarnased. Esinevad väikesed erinevused võivad olla tulenenud programmide iseärasustest. Üldjuhul põhistruktuurid ning motiivid esinevad **ennustatult** kõigis uuritavates organismides.



Joonis 17. ConSurf 1VTM tulemus (valgu struktuur roheline, RNA peaahel koos kõrvalahelatega punane) ning sellele sobitatud kärbseliste (*Acalyptratae* ja *Calyptratae* koos) ennustatud struktuur LOMETS´a programmi abil (valgu struktuur sinine). Joonised on koostatud Swiss-PdbViewer programmiga (versioon 4.1.0; Guex ja Peitsch, 1997). RMS 1.31Å.

INTERAKTSIOONIDEKS OLULISTE AMINOHAPETE VÕRDLUSED KIRJANDUSE BAASIL

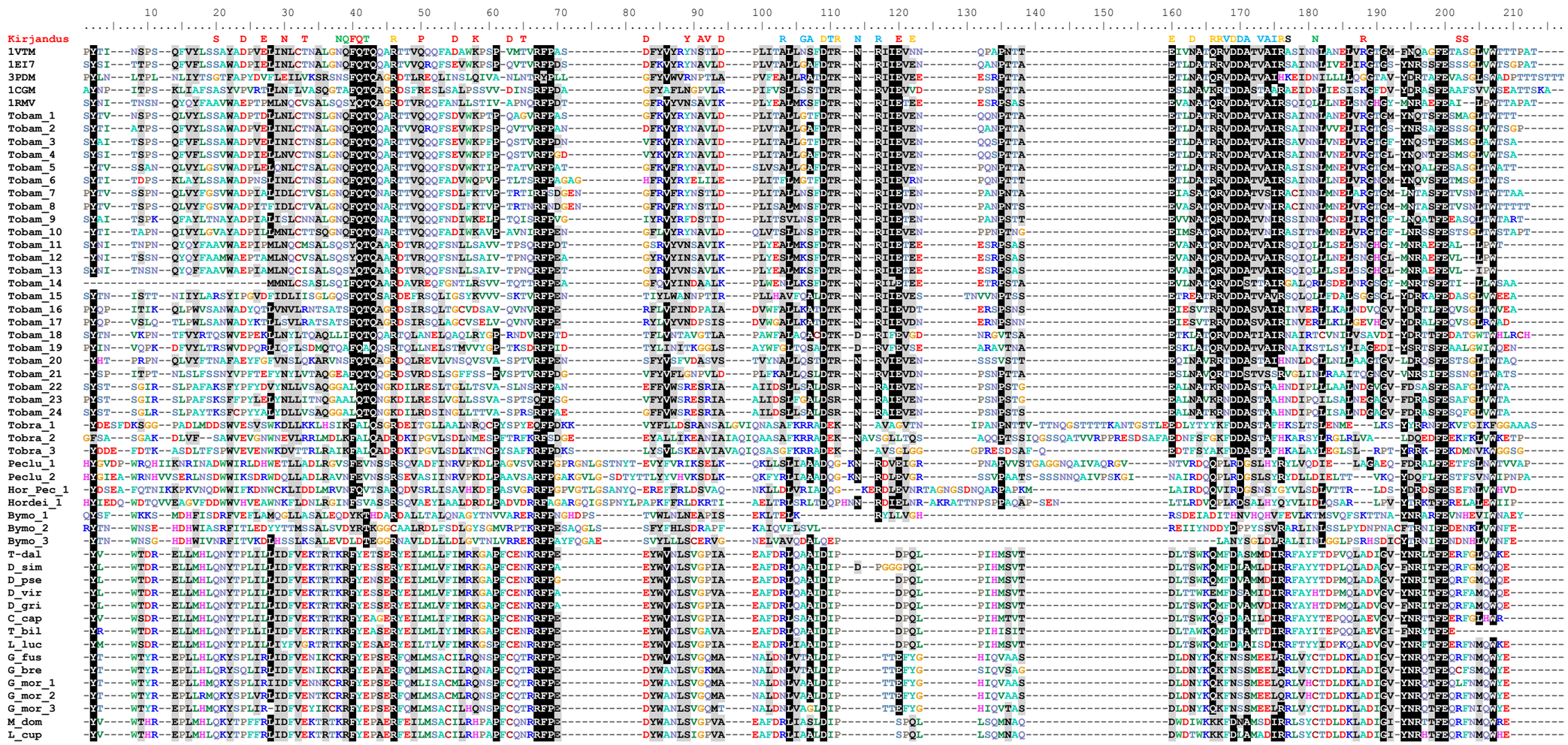
Võrreldes viiruste valk-valk või RNA-valk interaktsioonideks olulisi aminohappeid kärbseliste TMV-CP_{fly}, on võimalik vaadata, kas nendeks komplekside tekkeks olulised aminohapped on hoitud konserveerununa valiku abil. See võiks viidata valgu viiruslike omaduste alleshoidmisele. Kui interaktsioonideks olulised aminohapped on muutunud, nõrgestades või lõhkudes interaktsioone, võiks eeldada, et kärbselistes ei ole antud funktsiooni vaja ning seda ei ole alles hoitud.

Kirjanduse baasil otsiti välja TMV-CP struktuuris olulised aminohapped ning neid võrreldi tobamoviiruste TMV-CP_{viral} struktuuride, teiste TMV-CP_{viral} sisaldavate viiruste ning kärbseliste TMV-CP_{fly} järjestustega (joonis 18; TMV olulised aminohapped on välja toodud lisas 12).

Vaadates olulisi konserveerunud motiive (musta ja hallika taustaga esile toodud aminohapped joonisel 18), just täpsemalt tobamoviiruste baasil leituid, siis on näha üldist efekti – tobamoviiruste sees on leitud interaktsioonideks olulised motiivid olemas, tihti esineb neid ka tobraviirustes. Peclu-, hordei- ning bümoviirustel üldiselt puuduvad tobamoviirustel esinevad

konserveerunud motiivid. Kärbseliste puhul on näha *Acalyptratae* ja *Calyptratae* spetsiifilisi motiive. *Calpytatae* kärbselistest *Musca domestica´*l ning *Lucilia cuprina´*l esinevad motiivid on üldiselt segu *Calyptratae* ja *Acalyptratae* spetsiifilistest motiividest.

Kuigi hordei-, peklu- ning bümoviirustel puuduvad tubaka mosaiigiviiruse kirjeldatud RNA sidumiseks olulised aminohapped, ei tähenda see seda, et neil see omadus puudub. Arvatavasti kasutavad nad selleks teisi aminohappeid, mida täpsemalt ei ole teada, kuna nende struktuure ei ole nii põhjalikult uuritud. Kuna kärbselistes on tobamoviiruste viiruslike omaduste jaoks olulised aminohapped ning nende motiivid tunduvamalt erinevad, kuid kärbseliste enda sees väga konserveerunud, võiks eeldada, et viiruslikud omadused võiksid organismides puududa.



Joonis 18. TMV-CP fülogeneesipuu (struktuuri kaasarvamisel) koostatud MSA. Organismi nimed on lühendatud. MSA´le on lisatud kirjanduse baasil leitud tubaka mosaiigiviiruse kattevalgu interaktsioonideks (hüdrofiilse keskkonna loomiseks, RNA-VALK ning VALK-VALK) olulised aminohapped. Interaktsioonitüüpidele iseloomulikud aminohapped on märgitud erinevate värvidega: hüdrofiilse keskkonna tagamiseks oluliste interaktsioonide aminohapped – roheline; VALK-VALK interaktsioonideks – punane; RNA-VALK interaktsioonideks – sinine, RNA-VALK ja VALK-VALK interaktsionideks vajalikud aminohapped oranžiga ning RNA-VALK ja hüdrofiilse keskkonna loomiseks vajalikud aminohapped mustaga. Mitmese järjestuse joondus editeeriti BioEdit programmi abil (Hall, 1999). Joonise üleval on toodud üldise MSA positsioonide arv, interaktsioonid on arvestatud 1VTM järjestuse baasil ilma tühikuteta. Kirjanduses olulised aminohapped on välja toodud lisas 12.

2.4. ARUTELU

Käesoleva magistritöö eesmärgiks oli lähemalt uurida viiruste ja kärbseliste vahelist geneetilise materjali ülekannet. Antud töös keskendusi tubaka mosaiigiviiruse kattevalgu sarnase järjestuse detekteerimisele kärbselistes. Töö koosneb kolmest alateemast: toimunud integratsioonisündmuse täpsustamine (sündmuste arv ja aeg), alternatiivsete EVE´de tuvastamismeetodite efektiivsuste uurimine ning kärbselistes esineva TMV-CP_{fly} funktsiooni määramine.

TMV-CP_{fly} ESINEMINE ORGANISMIDES NING INTEGRATSIOONISÜNDMUSE TÄPSUSTAMINE

Antud tööga seotud bakalaureusetöös (Kirsip, 2013) detekteeriti TMV-CP_{fly} geeni 12 täielikult sekveneeritud *Drosophila* liigis SUPERFAMILY andmebaasi (Gough *et al*., 2001) abil. Kasutades juba tuvastatud TMV-CP_{fly} valgujärjestusi BLAST otsingus, määrati uuritav geen kolmes teises kärbselises, kelle genoome oli antud hetkel põhjalikumalt uuritud: *Musca domestica*, *Glossina morsitans morsitans* ning *Ceratitita capitata* (Kirsip, 2013).

Teostades kaks aastat hiljem analoogse BLAST analüüsi (kärbseliste valgujärjestuse baasil), detekteeriti TMV-CP_{fly} homolooge mitmetest uutest organismidest. See analüüs laiendas *Schizophora* tasemel peremeesorgansimide ringi, detekteerides TMV-CP_{fly} elementi ning *Drosophila melanogaster´*i sekveneeritud ning assambleeritud genoomi baasil määratud ümbritsevaid geene mitmetest kärbseliste perekondadest: *Lipara*, *Teleopsis*, *Drosphila*, *Themira*, *Ceratitis*, *Bactrocera*, *Rhagoletis*, *Glossina*, *Musca* ning *Lucilia*. Nende organismide esindajate puhul on peamiselt TMV-CP_{fly} geene detekteeritud terve genoomi või geenide sekveneerimisel. Lisaks on mõnedel tõendeid ka mRNA tasemel, transkriptoomi uurimise abil. Antud uuringus ei detekteeritud uuritavat TMV-CP_{fly} geeni perekondade *Delia*, *Sarcophaga* ning *Triarthia* esindajatel, küll aga esineb mõni ümrbitseva geen. Nende kolme perekonna esindaja puhul on ainult olemas transkriptoomi andmeid, seega see võib seletada miks teisi geene (kaasa arvatud TMV-CP_{fly}) ei deteketeeritud.

Nagu näha on, määrab meetodi edukust suurel määral olemasolevate andmete hulk ja kvaliteet, eriti just kvaliteetsed täisgenoomi sekveneerimised ja annoteerimised. Vaadates NCBI andmebaaside järgi bioprojektide arve kärbseliste sugukondade kohta, siis on näha, et enamus uuritavatest kärbselistest on kaasatud käesolevasse töösse (tabel 1, lisa 3). Lisaks kirjeldavad tabelis 1 (lisa 3) toodud mRNA´d ning valkude arvud organismide iseloomustatuse taset. See

kinnitab, et käesolev töö on olemasolevate andmete piires suutnud tuvastada uuritavaid geene efektiivselt.

Uurides lähemalt enne *Schizophora´*sid lahknenud kärbselisi, on võimalik täpsustada toimunud integratsiooni aega. Töö käigus ei tuvastatud uuritavat, TMV-CP_{fly} geeni, *Schizophora* kärbselistest varem lahknenud sirelastest (*Megaselia scalaris*) ega *Bombyliidae* kärbselistest (*Bombylius major*) uuritavat TMV-CP_{fly} geeni.

Antud töö käigus prooviti määrata ka katseliselt PCR analüüside abil uuritava TMV-CP_{fly} geeni olemasolu kärbselistes (*Drosophila melanogaster* ning *Syrphus*, *Helophilus* ja *Eristalis* perekondade sirelased). Katsete käigus suudeti PCR-i abil mõnel korral tuvastada uuritavat TMV-CP_{fly} geeni *Drosophila* kärbselistes (rakuliinist ja ka kärbsest), kuid antud metoodikat ei õnnestunud reprodutseeritaval moel tööle saada. Väitmaks, et TMV-CP_{fly} geeni sirelastel ei ole, peaks aga selle geeni detekteerimine *Schisophora* kärbselistel töötama usaldusväärselt ja robustselt. PCR-i ebaõnnestumise põhjuseks ei olnud ilmselt DNA puhtus või kvaliteet, sest mitokondriaalse geeni (tsütokroom c oksüdaasi I subühik) kui ka ühekoopialise tuumageeni (CAD´i geeni CPS domeen – karbomoüülfosfataasi süntetaas) detekteerimine õnnestus. Tõenäoliselt oli kitsaskohaks universaalsete praimerite disainimine TMV-CP_{fly} geenile konserveerunud valgudomeeni lühiduse tõttu (~ 600 nt ja vähemalt 80 miljonit aastat evolutsiooni). Tulenevalt eeltoodust, ei saadud katseliselt kindlaid tulemusi TMV-CP_{fly} eksisteerimise või mitteeksisteerimise kohta sirelastes.

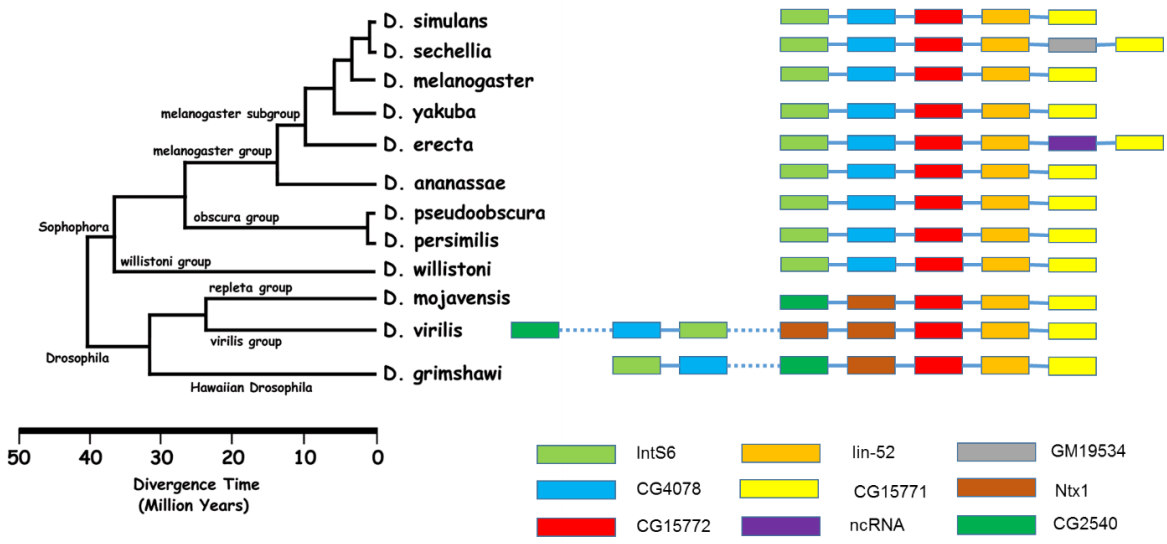
Praeguste andmete baasil võiks väita, et kirjeldatud integratsioon on toimunud kärbselistesse pärast sirelaste lahknemist *Schizophora´*dest, 80-120 MAT.

2012 aastal loodi i5K projekt, mille eesmärgiks on sekveneerida 5000 olulise lülijalgse organismi genoomi. Sekveneeritavate liikide valimisel keskenduti kindlatele kriteeriumitele: ökoloogilised rollid keskkonnas, inimese mõju, alalhoiu nõudmised, huvitav bioloogia ning lülijalgsete fülogeneesi ärakatvus (i5K consortium, 2013). Seega selle projekti abil võib saada veel usaldusväärsemaid andmeid kärbseliste lahknemise, TMV-CP_{fly} esinemise ning inttegratsiooniaja suhtes, seda just *Syrphidae* ja *Pipunculidae* sirelaste sekveneerimistega.

TMV-CP_{fly} geenipiirkonna uurimise eesmärgiks oli võrrelda geenide paiknevust kvaliteetselt sekveneeritud ning assambleeritud genoomiga kärbselistes. See võiks anda viiteid toimunud integratsioonisündmus kohta – kas on toimunud üks või mitu integratsiooni ning kirjeldada lookuse aktiivsust. Võrdlused on teostatud *Drosophila melanogaster* organismi baasil (joonis 19). Üldjuhul on ümrbitsev geeniala suhteliselt sarnane ortoloogsete geenide järjekorra alusel. *Drosophila* alamperekonnal (*D. grimshawi*, *D. virilis* ning *D. mojavensis*) erinevad teistest

Drosophila kärbselistest *Sophophora* alamperekonna piires kõige enam, omades Ntx1 (CG12752) ja CG2540 (*Drosophila melanogaster* geeninimed FlyBase FB2015_02 versiooni järgi) geene TMV-CP_{fly} vahetus läheduses. Ntx1 on ekspordivalk kärbselistes ning CG2540 geeni funktsiooni ei ole teada, kuid mõlemad on laialt levinud geenid. Üldjuhul lähedal asuvad geenid IntS6 ning CG4078 asuvad *D. virilis* ja *D. grimshawi* puhul kaugemal ning *D. mojavensis´*e puhul teisel kontiigil. Teiseks erinevuseks uuritava geenipiirkonna puhul on *D. erecta* inserteerunud mitte-kodeeriv RNA ning *D. sechellia* puhul tekkinud uus *de novo* geen, millel puuduvad teadaolevad ortoloogid.

Nende tulemuste põhjal võiks öelda, et kas *Drosophila* või *Sophophora* alamperekonnal on antud geenipiirkonnas toimunud muutused (arvatavasti translokatsioon), kuid selle sündmuse tagajärgi ei ole veel võimalik hinnata, kuna teistel kärbselistel puuduvad piisava ulatusega assambleeritud andmed. Samas geenipiirkonnadade üldine sarnane ülesehitus kinnitab vähemalt ***Drosophila* perekonna piires ühest integratsiooni aktiivsemasse lookusesse**.



Joonis 19. *Drosophila* kärbeste baasil välja toodud uuritava TMV-CP_{fly} ning ümbritsevate geenide järjekord suhtelise suunaga. Järjestikku asuvad geenid on ühendatud siniste joonstega ning kaugemal asuvad geenid sinise punktiirjoonega. All vasakul nurgas on välja toodud geenide nimed *Drosophila melanogaster* organismi baasil, välja arvatud hall GM19534, kuna antud geen on *D. sechellia* spetsiifiline ning puudub ortoloog teistes organismides. Joonis baseerub FlyBase andmebaasis asuval fülogeneesil ning järjestused on võetud samast andmebaasist (versioon FB2015_02; dos Santos *et al*., 2015).

ALTERNATIIVSED MEETODID EVE´DE DETEKTEERIMISEKS

Üldjuhul ei ole BLAST otsing heaks EVE esialgse otsimise alustamiseks, kuna antud meetodiga on võimalik uurida vaid teatud ajalises sügavuses toimunud protsesse. Lisaks homoloogseid valke kodeerivad viiruslikud järjestused, eriti nukleiinhappe tasemel, on võrreldes eukarüootsete organismidega niivõrd erinevad (Abroi ja Gough, 2011). Nukleiinhappe baasil detekteeritakse kõige vähem EVE´sid. Kasutades BLAST otsinguks viiruslikku valgujärjestust, on võimalik EVE´sid avastada juba suuremal määral. Kui arvestada, et valkude struktuur on ajas tunduvamalt konserveerunum kui järjestus, siis selle kaasamine otsingutesse peaks aitama uuritavaid EVE´sid paremini detekteerida.

EELASJÄRJESTUSE KAASAMINE

Eelasjärjestus konstrueeritakse teadaolevate järjestuste baasil nende ühise eellasseisundi jaoks, mis kaasates sarnaste järjestuste otsingule, võib aidata uute, integratsioonisündmusele lähedasemate järjestuste identifitseerimist. Eelasjärjestust võib kasutada kas viiruste või eukarüootsete organismide baasil. TMV-CP_{fly} uurimise puhul loodi nii viiruslikele järjestustele (*Virgaviridae* perekonna) kui ka kärbselistele (kõigile teadaolevatele kärbselistele ning *Acalyptratae* ja *Calyptratae* kärbselistele eraldi) kõige tõenäolisemad eelasjärjestused, teostades nende alusel BLAST otsinguid. Kahjuks käesolevas töös kirjeldatud meetod ei olnud efektiivne, tuvastades ainult varem detekteeritud järjestusi.

HMM ANALÜÜSID

Lisaks prooviti profiilidel baseeruvaid otsinguid, mis tunduvad olevat parem EVE´de detekteerimiseks, kui tavaline BLAST otsing, kuna kirjeldatud meetod võimaldab tuvastada ajaliselt sügavamaid suhteid. Antud võrdlusel tuli välja, et profiil-HMM meetod produtseerib tunduvamalt usaldusväärsemaid tulemusi kui BLAST analüüsid ning seega kirjeldatud meetodi juurutamine EVE´de uurimisse tundub olevat märksa efektiivsem.

Kindlasti esineb nii BLAST kui ka HMMER analüüsides piiripealseid tulemusi, mille kindlaksmääramine on raske tegevus. Antud juhul tulid välja *Apis dorsalis* ning *Apis mellifera* valgud (esinesid ka BLAST analüüsides, kuid tunduvamalt suuremate E-väärtusega ning need jäeti analüüsidest kõrvale). Valkudele teostati struktuuriennustused, mis piiripealselt määrasid valgustruktuurid TMV-CP_{viral} alla. Samas ei saa kindlalt väita, et antud valgudomeenid ei kuulu mõne teise viiruste sugukonna kattevalgu koosseisu, mille struktuuri ei ole hetkel määratud.

Antud töö puhul on kirjeldatud järjestused kõrvale jäetud, kuna kasutatud meetodite alusel ei ole võimalik usaldusväärselt ennustada nende kuulumist (ega ka mittekuulumist) TMV-CP superperekonda.

STRUKTUURI KAASAMINE FÜLOGENEESIUURINGUTESSE

Varasemates töödes on leitud, et tihti esineb fülogeneesiuuringutes ebakindlust, eriti just kaugelt lahknenud liikide puhul. Challis ja Schmidler (2012) pakkusid välja, et teadaolevate struktuuride kaasamine võiks antud probleeme lahendada.

Käesoleva töö puhul prooviti kirjeldatud meetodit, parandamaks kärbseliste ning viiruste kaugemaid lahknemisi. Käsitletava EVE puhul on teadaolevaid struktuure ainult *Virgaviridae* tobamoviiruste esindajatel. Kaasates need fülogeneesi konstrueerimisse, võiks loota kaugemate lahknemiste bootstrap väärtuste tõusu, samas kuna esindatud on ainult üks viiruse perekond, siis kindlasti esineb kallutatus ning üldine efekti suurus väheneb, mida on ka antud juhul näha (võrdluseks struktuuri arvestamisel loodud fülogeneesipuu joonisel 13 ning ilma struktuurida konstrueeritud fülogenees joonisel 20, lisas 11). Kuigi viiruslikud perekonnad ning kärbseliste alamgrupid on monofüleetilised ning lahknemised on toetatud kõrgete bootstrap väärtustega, esineb ka struktuuri arvestamisel siiski suur ebakindlus sügavamate lahknemiste puhul. Kaasates teiste viirusperekondade esindajate kattevalkude ja TMV-CP_{fly} struktuure, võiks loota probleemi paranemisele.

Kokkuvõtvalt võiks öelda, et uute meetodite, eriti profiil-HMM kaasamine aitab EVE´sid tuvastada tunduvamalt usaldusväärsemalt, kui kahe järjestuse võrdlused, ning see meetod tuleks kaasata kaugete homoloogide otsingutesse. Samas on antud meetodi puhul suureks probleemiks andmete vähesus, näiteks uuritavate valkude struktuuride puudumine, mis piirab kirjeldatud meetodite efektiivsust.

KÄRBSELISE TMV-CP_{fly} FUNKTSIOONI MÄÄRAMINE KÄRBSELISTES

Viirusliku järjestuse avastamise järel eukarüootsete organismide genoomidest, on välja pakutud, et kirjeldatud EVE´d võiksid toimida viiruseinfektsiooni kaitses. Antud sündmust on uuritud vaid ühe EVE baasil – EBLN elemendid oravates, kus järjeldati ei viiruslik kaitse võiks olemas olla, kuid antud fenomeni peaks veel lähemalt uurima.

Käsitletava EVE, TMV-CP_{fly}, viiruslikke omadusi uuriti bioinformaatiliste meetoditega, võrreldes kirjanduse baasil saadud VALK-RNA ning VALK-VALK interaktsioonideks olulisi aminohappeid kärbseliste omadega konserveeruvuse alusel. TMV-CP_{fly} geeni puhul oli toimunud mitmeid muutusi valgujärjestustes, samas kui struktuur on jäänud põhistruktuurilt samaks. See viitab, et uuritava geeni valgustruktuur võiks kas nõrgalt või üldse mitte siduda RNA´d ning arvatavasti ei moodusta suuremat filamentset viiruse kattevalgule sarnast struktuuri.

Nähes toimunud suuri muutusi, võib järeldada, et geen on omandanud uue funktsiooni ning sel juhul tuleks hakata uurima organismide ekspressiooni ning interaktsioonide andmeid, keskendudes funktsiooni määramisele.

TMV-CP_{fly} EKSPRESSEERUMISE SPETSIIFIKA BLAST JA MIKROKIIBI ANALÜÜSIDE BAASIL

Ekspressiooni- ning BLAST transkriptsioonianalüüsid kinnitavad TMV-CP_{fly} ekspresseerumist *Drosophila melanogaster* organismi piires nii emas- kui ka isasorgansimides. Kõige paremini on antud geeni ekspressiooni uuritud täiskasvanud organismis, kuid tõendeid esineb ka teiste elustaadiumite (embrüo, larva, nukk) ajal toimuvast ekspressioonist. Kõrget ekspressioonitaset esineb larva staadiumi rasvkehas ning nuku staadiumi anatoomiliste struktuuride piires. Kas mujal kudedes antud staadiumite puhul ekspressioon toimub, ei ole praeguse hetke andmete abil võimalik määrata. Täiskasvanud organismi piires esineb kõrge ekspressioonitase peas ja ajus ning testistes. Keskmise tasemega ekspressiooni on leida juba rohkemates kudedes, seega ei saa kindlalt väita, kas ekspresseerumine toimub vaid üksikute kudede piires või on vajalik elementaarsete ülesannete jaoks ning on tänu sellele levinud üle terve organismi ning kuna **valgu** ekspressiooni on leitud ka ajust, siis vähemalt selle koe piireks võiks eeldada TMV-CP_{fly} vajalikkust (ehk ei ole ainult transkribeeritud geen).

TMV-CP_{fly} EKSPRESSEERUMISE SPETSIIFIKA INTERAKTSIOONIDE BAASIL

Uurides erinevaid interaktsioone valkude, geenide, mikroRNA´de ning transkriptsioonifaktorite vahel, saab informatsiooni TMV-CP_{fly} ekspressioonitingimuste kohta, mis omakorda võiks kirjeldada kaudselt TMV-CP_{fly} ülesannet organismis.

TMV-CP_{fly} valk ekspresseerub proteoomika andmete alusel peas ning seda on detekteeritud nii tsütoplasma kui membraani fraktsioonides. Uuritava *Drosophila melanogaster CG15772*

valguga interakteeruvad tugeva skoori alusel kolm valku (FlyBase andmebaasi FB2015_02 versiooni järgi): letaalne (3)03670 (ID CG1715, 3R kromosoomis), DNA polümeraas δ (ID CG5949, 3L kromosoomis) ning CG1316 (3L kromosoomis).

l(3)03670 valk. Antud geeni on identifitseeritud neuroblastide eneseuuenduste protsessi kandidaat-geenina. Homosügootsed mutandid surid vastse teises staadiumis ning rakutsükkel oli mõjutatud, vähendades organismis esinevate rakkude arvu (Zhou, 2013). FlyAtlas´e (Chintapalli *et al*., 2007) ning modENCODE projekti (Contrino *et al*., 2012) järgi on antud geen ekspresseerunud vastse, nuku ning täiskasvanud staadiumis soo-spetsiifilisuseta. Yang *et al*. (2000) leidis, et antud geen ekspresseerub täiskasvanud organismis, kõrgenenud ekspressiooniga peas. Lisaks on väga kõrget ekspressiooni täheldatud ka vastse staadiumi tagasooles ning kõrgeid ekspressioone mitmetes kudedes: pea, silm, aju, tagasool, rasvkehad, süljenäärmed, süda ning emasorganismi *spermatheca* (FlyAtlas; Chintapalli *et al*., 2007). Lisaks on mass-spektroskoopia detekteerinud antud geeni valku 7-49 päevastes täiskasvanud kärbseliste südametes (Cammarato *et al*., 2011).

DNA polümeraas δ. Antud geeni ekspressiooni on detekteeritud embrüonaalsetes staadiumites (Weiszmann *et al*., 2009), kudedest kõrge ekspressiooniga esinevad *eye disc*, täiskasvanud organismi alakõht, munasari ning tiib (Genevisible; Hruz *et al*., 2008). Mass-spektroskoopia on detekteerunud antud valku 7-49 päevastes täiskasvanud kärbseliste südametes (Cammarato *et al*., 2011).

CG1316. Antud valgul on ennustatud mRNA ning nukleiinhappe sidumise võime, ning sellel on detekteeritud embrüonaalne ekspressioon (Weiszmann *et al*., 2009). Väga kõrget ekspressiooni esineb larva staadiumi *imaginal disc* ning kesknärvisüsteemis, nuku staadiumi rasvkehas ning täiskasvanud organismi peas, munasarjas, testistes ja *accessory gland* (modENCODE; Contrino *et al*., 2012).

Kirjeldatud interakteeruvate geenide ekspressioon on ülekattuv uuritava TMV-CP_{fly} omaga nii arengustaadiumite kui ka kudede piires. Seega tõenäosus, et antud valgud võiksid interakteeruda kindlates tingimustes, on tõenäoline, kuid vajavad katselist kinnitust. Seda just DNA polümeraas δ ja CG1316 puhul, kus mõlemal esineb nukleiinhapet siduv omadus ning kaksik-hübriidanalüüsides võib seetõttu välja selekteerida ka VALK1-RNA-VALK2 kompleksi, see tähendab nukleiinhappe vahendatud interaktsiooni.

Nagu näha on, omavad kirjeldatud protsessid sarnaseid ekspressiooniaegu sarnaste kudede piires. Nende andmete järgi võiks välja pakkuda hüpoteesi, et uuritav valk on oluline kärbseliste embrüogeneesis erinevate kudede arenemisel ning hilisemalt täiskasvanud organismis võib

uuritav valk toimida kindlates kudedes (ajus, neuronites) spetsiifilise ülesandega. Seda kinnitavad ka seondunud mikroRNA´d, mis on olulised närvirakkude arengus, ning transkriptsioonifaktorid, mis on seotud baaskudede tekke ja arenguga.

KOKKUVÕTE

Järjest rohkem on teadlasi hakanud huvitama uute tekkinud geenide päritolu ning ekspresseerumise erinevused võrreldes eelasgeeniga. Näiteks kas horisontaalsel geeniülekandel omandatud uus geen on integreerunud organismis pseudogeenistunud, säilitanud algse või tekitanud uue funktsiooni. Horisotaalsel geeniülekandel *de novo* tekkinud geenide näiteks võivad olla järjest rohkem tuvastatud endogeensed viiruslikud elemendid (EVE, ing k *endogenous viral elements*), mida on deteketeeritud erinevate taimede, loomade ja seente genoomidest.

Käesolevas magistritööga seotud bakalaureusetöös uuriti lähemalt tubaka mosaiigiviiruste sarnase kattevalgu (TMV-CP) esinemist kärbseliste genoomides. Antud töös määrati geneetilise materjali ülekandesuunaks viirustelt eukarüootsetele organismidele 60-250 milionit aastat tagasi (Kirsip, 2013).

Antud töö eesmärgiks oli uurida lähemalt TMV-CP võimalikku funktsiooni uutes peremeesorganismides, püüdes määrata seda bioinformaatiliste meetodite abil analüüsides teadaolevaid andmeid. Algselt võrreldi viiruste ning kärbselistesse integreerunud valkude järjestusi ning struktuure. Kärbseliste ennustatud struktuuridel oli näha piisavalt suuri muutusi, eriti just interaktsioonides RNA´ga, mis viitab antud interaktsiooni omaduse kaotusele või olulisele nõrgenemisele. Samuti olid toimunud muutused valk-valk interaktsioonideks olulistes aminohapetes, tänu millele saaks viirustes tekkida filamentne struktuur. Mõistes, et kärbselistes eksisteeriv TMV-CP_{fly} valgul ei pruugi olla viirusliku kattevalgu omadusi ja seega ka mitte funktsiooni, keskenduti järgnevalt uue funktsiooni ennustamisele. Selleks kasutati mikrokiibi ning sekveneeritud transkripte iseloomustavaid andmeid. Lisaks võeti arvesse ka valk-valk ning geen-mikroRNA/transkriptsioonifaktori interaktsioone, püüdes leida viiteid uuritava geeni ekspressioonitingimustele. Kogu saadud informatsiooni arvestades, võiks hinnata, et uuritav TMV-CP_{fly} on oluline kärbseliste embrüonaalses arengus, täpsemalt erinevate kudede morfogeneesis. Lisaks leidub viiteid täiskasvanud organismi närvisüsteemispetsiiflisele ekspressioonile, mida tuleks järgnevalt katseliselt kinnitada või ümber lükata.

Siiani on EVE´sid peamiselt detekteeritud kahe järjestuse võrdluse alusel, kasutadeks selleks BLAST perekonna programme. Töö teiseks ülesandeks oli hinnata alternatiivsete EVE´de deketeerimise meetodite efektiivsust TMV-CP_{fly} geeni baasil. Kõige efektiivsemaks neist osutus peidetud Markovi mudelite (HMM) kasutamine otsingu teostamiseks. Antud juhtudel on erinevate struktuurseid valgudomeene arvestavad andmebaasid loonud ühisesse superperekonda kuuluvate valgujärjestusi iseloomustavaid profiile. Kuna need arvestavad ka

struktuuri, mis on ajas tunduvamalt konserveerunud kui nukleiinhape või aminohappeline järjestus, siis on see meetod suuteline detekteerima ühist päritolu järjestusi väga väikese valgujärjestuse identsuse alusel. Teostades EVE´de deketeerimiseks süstemaatilisi otsinguid viiruslike järjestuste baasil, kasutades selleks eelnevalt kirjeldatud andmebaase, peaks saama võimalikult laia peremeesteringi uuritavatele valkudele, seal hulgas ka eukarüootsed organismid, mida on hiljem võimalik laiendada tavaliste BLAST otsingutega.

Kolmandaks ülesandeks oli toimunud integratsiooni sündmuse toimumise aja täpsustamine. Selleks uuriti *Schizophora* kärbselistest eelnevalt lahknenud, kuid sekveneeritud, organisme. Käesoleva töö käigus leitud, et varasemalt lahknenud sirelastel (*Aschiza*) uuritavat TMV-CP_{fly} geeni ei ole tuvastatud, sama leiti ka veel varem lahknenud *Bombyliidae* kärbselistel. See viitab, et integratsioon peaks olema toimunud enne 80 MAT ja pärast 120 MAT (ajad võetud Wiegmann *et al*., 2011). Arvestades, et tobamoviiruste vanuseks on määratud 120 MA ning kuna uuritavat geeni eksisteerib ka teistes *Virgaviridae* sugukondades, siis see kinnitab veelkord toimunud integratsioonisuunda viirustelt peremeesorganismidesse.

Käesolevas töös uuriti põhjalikumalt kärbselistes tuvastatud tubaka mosaiigiviiruste kattevalgu sarnaseid järjestusi, täpsustades varem saadud integratsiooniaega ning määrates uuritava geeni funktsiooni organismides. Lisaks pakuti välja efektiivsem meetod EVE´de detekteerimiseks, kui seda on praegu laialdaselt kasutusel olev BLAST analüüs. Implementeerites käesolevas töös teostatud analüüse teiste EVE´de tuvastamisteks ning analüüsimisteks, on võimalik teha edusamme antud uurimisvalgkonnas.

Investigation of the gene, *TMV-CP*, transferred from viruses to flies, with focus on time and function

Heleri Kirsip

SUMMARY

New genes can emerge by different mechanisms, but this Master´s thesis is concentrated on horizontal gene transfer (HGT). HGT is well known process in bacteria, but is a rare event in higher eukaryotes with the exception of plants. One example of it could be endogenous viral elements (EVE´s) that have been discovered in different plants´ and animals´ genomes. This process is thoroughly researched regarding retroviral integrations, but new non-retroviral integration examples are emerging.

This Master´s thesis is a continuation of undergraduate work (Kirsip, 2013), where tobacco mosaic virus´s coat protein´s (TMV-CP) protein domain was discovered in different *Drosophila* fly genomes. In this previous work, it was determined, that the integration was from viruses to fly genomes and the event took place about 60-250 MYA (million years ago).

In the literature part, a review was given about *de novo* gene emergence, focusing on HTG, and previously detected EVE´s with described common general detection methodology. Since protein structure is much more conserved in time than sequence, protein structure implementation to detection methodology was also reviewed.

In the experimental part of the work, there were three main objectives: further research of the integration event with main focus on time, determination of the effectiveness of new alternative methods, especially compared to BLAST analysis, and evaluation of the function of TMV-CP_{fly} in flies.

For the first part of the work, BLAST analysis was used to identify organisms that contained TMV-CP_{fly} gene or transcript. From the work, it was deduced that *Schizophora* flies contained aforementioned gene, but previously diverged groups, hoverflies (*Aschiza*) and bee flies (*Bombyliidae*) do not contain TMV-CP_{fly}. This indicates that viral TMV-CP_{viral} integrated into the flies genomes about 80-120 MYA.

Previous works of EVE identification uses mainly BLAST analysis. For this part, alternative methods like ancestral sequence, protein structures and profile-HMM (hidden Markov models) were tested to measure their effectiveness, especially against commonly used BLAST analysis.

It was deduced that profile-HMM was the most effective alternative method, giving more reliable results than BLAST, because it uses protein domains that have been classified to same structural superfamilies and creates profiles based on their sequences which can be searched against. This new method should be used, especially in distant homologue detection.

The third aim of the work was to identify TMV-CP_{fly}´s possible function in flies with bioinformatical methods. Firstly the viral and fly TMV-CPs were compared, focusing on necessary amino acids for RNA-PROTEIN and PROTEIN-PROTEIN interactions. Form the many significant changes in the amino acid sequences, it was deduced that TMV-CP_{fly} may not have the same function as viral TMV-CP. Next step was to analyse available expression data. Taking all the information into account, it could be assessed that TMV-CP_{fly} is essential in embryonic development, especially in morphogenesis of different tissues. In addition there are indications that it also might have an essential role in adult fly nervous system.

In this Master´s thesis the integration event between viruses and eukaryotic organisms was thoroughly investigated. From this work, it was deduced that this event took place about 80-120 MYA from *Virgaviridae* viruses to *Schizophora* flies and the integrated protein is necessary for the development of flies in the early embryonic stages. In addition to that, alternative method for EVE detection was also recommended, especially when identifying distant homologues, because this method´s results are more reliable than generally used BLAST analysis.

KASUTATUD KIRJANDUSE LOETELU

- **ARTIKLID**

Abroi, A., Gough, J. (2011). Are viruses a source of new protein folds for organisms? – Virosphere structure space and evolution. Bioessays. 33(8): 626-635.

Aiewsakun, P., Katzourakis, A. (2015). Endogenous viruses: connecting recent and ancient viral evolution. Virology.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acid Res. 25: 3389-3402.

Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, S. E., Chothia, C., Murzin, A. G. (2008). Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 36(Database issue): D419-425.

Andrews, J., Bouffard, G G.,, Cheadle, C., Lu, J., Becker, K. G., Oliver, B. (2000). Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis. Genome Res. 10: 2030-2043.

Arriagada, G., Gifford, R. J. (2014). Parvovirus-derived endogenous viral elements in two South American rodent genomes. J Virol. 88(20): 12158-12162.

Ashby, M. K., Warry, A., Bejarano, E. R., Khashoggi, A., Burrell, M., Lichtenstein, C. P. (1997). Analysis of multiple copies of geminiviral DNA in the genome of four closely related Nicotiana species suggest a unique integration event. Plant Mol Biol. 35: 313–321.

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., Pupko, T. (2012). FastML: a web server for peobabilistic reconstruction of ansectral sequences. Nucleic Acids Res. 40 (Web Server issue): W580-584.

Aswad, A., Katzourakis, A. (2014). The first endogenous herpesvirus, identified in the tarsier genome, and novel sequences from primate rhabdoviruses and lymphocryptoviruses. PLoS Geneti. 10(6).

Bai. Y., Casola, C., Feschotte, C., Betran, E. (2007). Comparative genomics reveal a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. Genome Biol. 8 (1): 1-9.

Ballinger, M. J., Bruenn, J. A., Taylor, D. J. (2012). Phylogent, integration and expression of sigma virus-like genes in Drosophila. Mol Phylogenet Evol. 65(1): 251-258.

Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V., Robinson-Rechavi, M. (2008)
Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. *in* DILS: Data Integration in Life Sciences. Lecture Notes in Computer Science. 5109:124-131.

Begun, D. J., Lindfors, H. A., Kern, A. D., Jones, C. D. (2007). Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. Genetics. 176: 1131.

Bejarno, E. R., Khashoggi, A., Witty, M., Lichtenstein, C. (1996). Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. Proc Natl Acad Sci USA. 93(2): 759-764.

Belyi, V. A., Levine, A. J., Skalka, A. M. (2010 A). Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old. J Virol. 84(23): 12458-12462.

Belyi, V. A., Levine, A. J., Skalka, A. M. (2010 B). Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. PLoS Pathoq. 6(7).

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Research. 28: 235-242.

Bertone, M.A., Courtney, G.W., Wiegmann, B.W., (2008). Phylogenetics and temporal diversification of the earliest true flies (Insecta: Diptera) based on multiple nuclear genes. Systematic Entomology. 33: 668-687.

Betran, E., Long, M. (2003). *Dntf-2r*, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. Genetics. 164: 977-988.

Betran, E., Thornton, K., Long, M. (2002). Retroposed new genes out of the X in Drosophila. Genome Res. 12: 1854–1859.

Bhyravbhatla, B., Watowich, S. J., Caspar, D. L. (1998). Refined atomic model of the four-layer aggregate of the tobacco mosaic virus coat protein at 2.4 A resulotion. Biophys J. 74(1): 604-615.

Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. A., Boudreau, M. E., Nesbo, C. L., Case, R. J., Doolittle, W. F. (2003). Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet. 37: 283–328.

Brammer, C. A., von Dohlen, C. D. (2007). The evolutionary history of Stratiomyidae (Insecta: Diptera): the molecular phylogeny of a diverse family of flies. Mol Phyl Evol. 43: 660-673.

Brenner, S. E., Chothia, C., Hubbard, T. (1998). Assessment of sequence comparison methods with reliable structurally-identifed distant evolutionary relationships. Proc Natl Acad Sci USA. 95: 6073-6078.

Bruenn, J. A., Warner, B. E., Yerramsetty, P. (2015). Widespread mitovirus sequences in plant genomes. 3.

Brunner, E., Ahrens, C. H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., Deutsch, E. W., Panse, C., de Lichtenberg, U., Rinner, O., Lee, H., Pedrioli, P. G., Malmstrom, J., Koehler, K., Schrimpf, S., Krijgsveld, J., Kregenow, F., Heck, A. J., Hafen, E., Schlapbach, R., Aebersold, R. (2007). A high-quality catalog of the Drosophila melanogaster proteome. Nat Biotechnol. 25(5): 576-583.

Caetano-Anolles, G., Nasir, A. (2012). Benefits of using molecular structure and abundance in phylogenimic analysis. Front Genet. 3(172).

Calderone, A., Castagnoli, L., Cesareni, G. (2013). mentha: a resource for browsing integrated protein-interaction networks. Nat Methods. 10(8): 690-691.

Cammarato, A., Ahrens, C. H., Alayari, N. N., Qeli, E:, Rucker, J., Reedy, M. C., Zmasek, C. M., Gucek, M., Cole, R. N., Van Eyk, J. E., Bodmer, R., O´Rourke, B., Bernstein, S. I., Foster, D. B. (2011). A mighty small heart: the cardiac proteome of adult *Drosophila melanogaster*.

Challis, C. J., Schmidler, S. C. (2012). A stochastic evolutionary model for protein structure alignment and phylogeny. Mol Evol Biol. 29(11): 3575-3587.

Chintapalli, V. R., Wang, J., Dow, J. A. (2007). Using FlyAtlas to identify better Drosophila melanogaster models for human disease. Nat Genet. 39(6): 715-720.

Chothia, C., Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. EMBO J. 5(4): 823-826.

Chu, H., Jo, Y., Cho, W. K. (2014). Evolution of endogenous non-retroviral genes integrated into plant genomes. Curr Plant Biol. 1: 55-59.

Contrino, S., Smith, R. N., Butano, D., Carr, A., Hu, F., Lyne, R., Rutherford, K., Kalderimis, A., Sullivan, J., Carbon, S., Kephart, E. T., Lloyd, P., Stinson, E. O., Washington, N. L., Perry, M. D., Ruzanov, P., Zha, Z., Lewis, S. E., Stein, L. D., Micklem, G. (2012). modMine: flexible access to modENCODE data. Nucleic Acids Res. 40(Database issue): D1082-1088.

Cornelis, G., Heidmann, O., Bernard-Stoecklin, S., Reynaud, K., Veron, G., Mulot, B., Dupressoir, A., Heidmann, T. (2012). Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. Proc Natl Acad Sci USA. 109(7): E432-441.

Cornelis, G., Heidmann, O., Degrelle, S. A., Vernochet, C., Lavialle, C., Letzelter, C., Bernard-Stoecklin, S., Hassanin, A., Mulot, B., Guillomot, M., Hue, I., Heidmann, T., Dupressoir, A. (2013). Captured retroviral envelope syncytin gene associated with the unique placental structure of higher ruminants. Proc Natl Acad Sci USA. 110(9): E828-837.

Cui, J., Holmes, E. C. (2012 A). Endogenous RNA viruses of plants in insect genomes. Virology. 427(2): 77-79.

Cui, J., Holmes, E. C. (2012 B). Evidence for an endogenous papillomavirus-like element in the platypus genome. J Gen Virol. 93(Pt 6): 1362-1366.

Cui, J., Zhao, W., Huang, Z., Jarvis, E. D., Gilbert, M. T., Walker, P. J., Holmes, E. C., Zhang, G. (2014). Low frequency of paleoviral infiltration across the avian phylogeny. Genome Biol. 15(12): 539.

Darriba, D., Taboada, G. L., Doallo, R., Prosada, D. (2011). ProtTest 3: fast selection of best-fit modelid of protein evolution. Bioinformatics. 27: 1164-1165.

de Souza, W., Motta, M. C. (1999). Endosymbiosis in protozoa of the Trypanosomatidae family. FEMS Microbiol Lett. 173(1): 1-8.

Del-Toro, N., Dumousseau, M., Orchard, S., Jimenez, R. C., Galeota, E., Launay, G., Goll, J., Breuer, K., Ono, K., Salwinski, L., Hermjakob, H. (2013). A new reference implementation of the PSICQUIC web service. Nucleic Acids Res. 41(Web server issue): W601-606.

Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., Aebersold, R. (2006). The PeptideAtlas project. Nucleic Acids Res. 34(Database issue): D655-D658.

dos Santos, G., Schroeder, A. J., Goodman, J. L., Strelets, V. B., Crosby, M. A., Thurmond, J., Emmert, D. B., Gelbart, W. M., FlyBase Consortium. (2015). FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res. 43(Database issue): D609-D697.

Dupressoir, A., Vernochet, C., Bawa, O., Harper, F., Pierron, G., Opolon, P., Heidmann, T. (2009). Syncytin-A knockout mice demonstrate the criticalrole in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. Proc Natl Acad Sci. 106: 12127–12132.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5): 1792-1797.

Emerson, J. J., Kaessmann, H., Batran, E., Long, M. (2004). Extensive gene traffic on the mammalian X chromosome. Science. 303(5657): 537-540.

Evanoff, E., McIntosh, W. C., Murphey, P. C. (2001). Stratigraphic summary and Ar/Ar Geochronology of the Florissant Formation, Colorado. Proceeding of the Denver Museum of Nature and Science. 4(1).

Fablet, M., Bueno, M., Potrzebowski, L., Kaessmann, H. (2009). Evolutionary origin and functions of retrogene introns. Mol Biol Evol. 26: 2147-2156.

Fan, G., Li, J. (2011). Regions identity between the genome of vertebrates and non-retroviral families of insect viruses. Virol J. 8(511**).**

Finn, R. D., Clements, J., Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39 (Wed Server issue): W29-37.

Fort, P., Albertini, A., Van-Hua, A., Berthomieo, A., Roche, S., Delsuc, F., Pasteur, N., Capy, P., Gaudin, Y., Weill, M. (2012). Fossil rhabdoviral sequences integrated into arthropod genomes: ontogeny, evolution, and potential functionality. Mol Biol Evol. 29(1): 381-390.

Frank, A. C., Wolfe, K. H. (2009). Evolutionary capture of viral and plasmid DNA by yeast nuclear chromosomes. Eukaryot Cell. 8(10): 1521-1531.

Fujino, K., Horie, M., Honda, T., Merriman, D. K., Tomonaga, K. (2014). Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. PNAS. 111(36): 13175-13180.

Gaunt, N. W., Miles, M. A. (2002). An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. Mol Biol Evol. 19(5): 748-761.

Geering, A. D., Maumus, F., Copetti, D., Choisne, N., Zwickl, D. J., Zytnicki, M., McTaggart, A. R., Scalabrin, S., Vezzulli, S., Wing, R. A., Quesneville, H., Teycheney, P. Y. (2014). Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. Nat Commun. 5(5269).

Gilbert, C., Meik, J. M., Dashevsky, D., Card, D. C., Castoe, T. A., Schaack, S. (2014). Endogenous hepadnaviruses, bornaviruses and circoviruses in snakes. Proc Biol Sci. 281(1791).

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L. Jr,, White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., Rothberg, J. M. (2003). A protein interaction map of Drosophila melanogaster. Science. 302 (5651): 1727-1736.

Gladyshev, E. A., Meselson, M., Arkhipova, I. R. (2008). Massive horizontal gene transfer in bdelloid rotifers. Science. 320: 1210–1213.

Gough J, Chothia C (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Research. 30 (1): 268-272.

Gough, J. (2002). The SUPERFAMILY database in structural genomics. Acta Crystallogr D Biol Crystallogr. 58 (Pt 11): 1897-1900.

Gough, J., Karplus, K., Hughey, R., Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. J Mol Biol. 313: 903-919.

Graham, J., Butler, P. J. G. (1979). Binding of oligonucleotides to teh disc of Tobacco-Mosaic-Virus protein. Eur J Biochem. 93: 333-337.

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B.J., Chiang, H.R., King, N., Degnan, B.M., Rokhsar, D.S., and Bartel, D.P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature. 455: 1193–1197.

Guex, N., Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis. 18: 2714-2723.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser. 41: 95-98.

Haspel, M. V., Knight, P. R., Daff, R. G., Rapp, F. (1973). Activation of a latent measles virus infection in hamster cells. J Virol. 12: 690-695.

Hayward, J. A., Tachedjian, M., Cui, J., Field, H., Holmes, E. C., Wang, L. F., Tachedjian, G. (2013). Identification of diverse full-length endogenous betaretroviruses in megabats and microbats. Retrovirology. 10(35).

Heidmann, O., Vernochet, C., Dupressoir, A., Heidmann, T. (2009). Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: A new ‘‘syncytin’’ in a thirdorder of mammals. Retrovirology. 6 : 107.

Herman, J. L., Challis, C. J., Novak, A., Hein, J., Schmidler, S. C. (2014). Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. Mol Biol Evol. 31(9): 2251-2266.

Hoffmeister, M., Martin, W. (2003). Interspecific evolution: microbial symbiosis, endosymbiosis and gene transfer. Environ Microbiol. 5(8): 641-649.

Holmes, E. C. (2009). The evolutionary genetics of emerging viruses. Annu Rev Ecol Evol Syst. 40: 353-372.

Holmes, K. C. (1979). Protein-RNA interactions druing TMV assembly. J Supramol Struct. 12: 305-320.

Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., Ikuta, K., Jern, P., Gojobori, T., Coffin, J. M., Tomonaga, K. (2010). Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature. 463(7277): 84-87.

Horie, M., Kobayashi, Y., Suzuki, Y., Tomonaga, K. (2013). Comprehensive analysis of endogenous bornavirus-like elements in eukaryote genomes. Philos Trans R Soc Long B Biol Sci. 368(1626).

Horie, M., Tomonaga, K. (2011). Non-retroviral fossils in vertebrate genomes. Viruses. 3: 1836-1848.

Hotopp, J.C., Clark, M.E., Oliveira, D.C., Foster, J.M., Fischer, P., Torres, M.C., Giebel, J.D., Kumar, N., Ishmael, N., Wang, S., Ingram, J., Nene, R. V., Shepard, J., Tomkins, J., Richards, S., Spiro, D. J., Ghedin, E., Slatko, B. E., Tettelin, H., Werren, J. H. (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* 317: 1753–1756.

Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W. Zimmermann, P. (2008). Genevestigator V3: a reference expression database for the meta-analysis of transcriptomes. Advances in Bioinformatics. 420747.

Illegard, K., Ardell, D. H., Elofsson, A. (2009). Structure is three to ten times more conserved than sequence – a study of structural response in protein cores. Proteins. 77(3): 499-508.

Jacob, F. (1977) Evolution and tinkering. Science 196: 1161–1166.

Jaroszewski, L., Rychlewski, L., Li, Z., Li, W., Godzik, A. (2005). FFAS03: a server for profile-profile sequence alignments. Nucleic acids research. 33(Web Server issue): W284-288.

Johnson, M. E., Viggiano, I., Bailey, J. A., Abdul-Rauf, M., Goodwin, G., Rocchi, M., Eichler, E. E. (2001). Positive selection of a gene family during the emergence of humans and African apes. Nature. 413: 514-519.

Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. Genome Research 20: 1313–1326.

Kaessmann, H., Vinckenbosch, N., Long, M. (2009). RNA-based geen duplication: mechanistic and evolutionary insights. Nat Rev Genet. 10 (1): 19-31.

Karplus, K., Barrett, C., Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. Bioinformatics. 14(10): 846-856.

Katzourakis, A., Gifford, R. J. (2010). Endogenous viral elements in animal genomes. PLoS Genet. 6(11).

Keeling, P. J., Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. 9: 605–618.

Kenton, A., Khashoggi, A., Parokonny, A., Bennett, M. D., Lichtenstein, C. (1995). Chromosomal location of endogenous geminivirus-related DNA sequences in Nicotiana tabacum L. Chromosome Res. 3(6): 346-350.

Kimmins, S., Sassone-Corsi, P. (2005). Chromatin remodelling and epigenetic features of germ cells. Nature. 434(7033): 583-589.

Kirsip, H. (2013). Kas taimeviirused võivad olla uute valgudomeenide allikaks hulkraksetele loomadele? Bakalaureusetöö. Tartu Ülikool.

Kleene, K. C. (2001). A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mech Dev. 106(1-2): 3-23.

Kleene, K. C. (2005). Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. Dev Biol. 277(1): 16-26.

Kondo, H., Chiba, S., Suzuki, N. (2015). Detection and analysis of non-retroviral RNA virus-like elements in plant, fungal, and insect genomes. Methods Mol Biol. 1236: 73-88.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., Haussler, D. (1994). Hidden Markov models in computational biology: applications to protein modeling. J Mol Biol. 235: 1501-1531.

Kutty, S. N., Pape, T., Wiegmann, B. M., Meier, R. (2010). Molecular phylogeny of the Calyptratae (Diptera: Cyclorrhapha) with an emphasis on the superfamily Oestroidea and the position of Mystacinobiidae and McAlpine´s fly. Systematic Entomology. 35(4): 614-635.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., Ben-Tal, N. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucl Acids Res. 33:W299-W302.

Lavialle, C., Cornelis, G., Dupressoir, A., Esnault, C., Heidmann, O., Vernochet, C., Heidmann, T. (2013). Paleovirology of ´syncytins´, retroviral env genes exapted for a role in placentation. Philos Trans R Soc Lond B Biol Sci. 368(1626).

Levine, M.T., Jones, C.D., Kern, A.D., Lindfors, H.A., and Begun, D.J. (2006). Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proc. Nat. Acad. Sci. USA 103, 9935–9939.

Li, C. X., Shi, M., Tian, J. H., Lin, X. D., Kang, Y. J., Chen, L. J., Qin, X. C., Xu, J., Holmes, E. C., Zhang, Y. Z. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. Elife. 4.

Li, N., Li, Q. (2015). Identification and characterization of endogenous viral elements for the three key schistosomes of humans. Pak J Pharm Sci. 28(1): 375-382.

Li, X. Y., Song, B. A., Hu, D. Y., Chen, X., Wang, Z. C., Zeng, M. J., Yu, D. D., Chen, Z., Jin, L. H., Yang, S. (2013). The Conformations and Interactions of the Four-Layer Aggregate Revealed by X-ray Crystallography Diffraction Implied the Importance of Peptides at Opposite Ends in Their Assemblies. PDB ID: 4GQH. Unpublished data.

Li, X., Song, B., Chen, X., Wang, Z., Xeng, M., Yu, D., Hu, D., Chen, Z., Jin, L., Yang, S., Yang, C., Chen, B. (2013). Crystal structure of a four-layer aggregate of engineered TMV CP implies the importance of terminal residues for oligomer assembly. PLoS ONE. 8(11).

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., Castagnoli, L., Cesareni, G. (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 40(Database issue): D857-861.

Lieberman, B. S. (2003). Taking the pulse of the cambrian radiation. Integr Comp Biol. 43(1): 229-237.

Lim, K. Y., Matyasek, R., Lichtenstein, C. P., Leitch, A. R. (2000). Molecular cytogenetic analyses and phylogenetic studies in the Nicotiana section Tomentosae. Chromosoma. 109: 245–258.

Liu, H., Fu, Y., Jiang, D:, Li, G., Xie, J., Cheng, J., Peng, Y., Ghabrial, S. A., Yi, X. (2010). Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. J Virol. 84(22): 11876-11887.

Liu, H., Fu, Y., Xie, J., Cheng, J., Ghabrial, S. A., Li, G., Peng, Y., Yi, X., Jiang, D. (2011). Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. J Virol. 85(19): 9863-9876.

Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C., Murzin, A. G. (2002). SCOP database 2002: refinements accommodate structural genomics. Nucleic Acids Res. 30(1): 264-267.

Lobley, A., Sadowski, M. I., Jones, D. T. (2009). pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics. 25(14): 1761-1767.

Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., Janssens, H., Ji, W., Mclaren, P., North, P., Rana, D., Riley, T., Sullivan, J., Watkins, X., Woodbridge, M., Lilley, K., Russell, S., Ashburner, M., Mizuguchi, K., Micklem, G. (2007). FlyMine: an integrated database for Drosophila and Anopheles genomemics. Genome Biol. 8(7): R129.

Madera, M. (2008). Profile Comparer: a program for scoring and aligning profile hidden Markov models. Bioinformatics. 24(22): 2630-2631.

Maori, E., Tanne, E., Sela, I. (2007). Reciprocal sequence exchange between non-retro viruses and hosts leading to the appearance of new host phenotypes. Virology. 362(2): 342-349.

Marques, A., Dupanloup, I., Vinckenbosch, N., Reymond, A., Kaessmann, H. (2005). Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 3.

Mi, S., Lee, X., Li, X., Veldman, G. M., Finnerty, H., Racie, L., LaVallie, E., Tang, X. Y., Edouard, P., Howes, S., Keith, J. C. Jr., McCoy, J. M. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature. 403: 785–789.

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T., Beutel, R. G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A. J., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T. R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L. S., Kawahara, A. Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D. D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J. L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von Reumont, B. M., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N. U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M. G., Wiegmann, B. M., Wilbrandt, J., Wipfler, B., Wong, T. K., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D. K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K. M., Zhou, X. (2014). Phylogenomics resolve the timing and pattern of insect evolution. Science. 364(6210): 763-767.

Morrison D. A. (2009). Why would phylogeneticists ignore computerized sequence alignment? Syst. Biol. 58: 150–158.

Mostovski, M. B. (2000). Contributions to the study of fossil snipe-flies (Diptera: Rhagionidae), the genus Palaeobolbomyia. *Paleontological Journal.* 34(Suppl 3): S360-S366.

Murad, L., Bielawski, J. P., Matyasek, R., Kovarik, A., Nichols, R. A., Leitch, A. R., Lichtenstein, C. P. (2004). The origin and evolution of geminivirus-related DNA sequences in Nicotiana. Heredity (Edinb). 92(4): 352-358.

Murad, L., Lim, K. Y., Christopodulou, V., Matyasek, R., Lichtenstein, C. P., Kovarik, A. (2002). The origin of tobacco's T genome is traced to a particular lineage within Nicotiana tomentosiformis (Solanaceae). Am J Bot. 89: 921–928.

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247: 536-540.

Mushegian, A. R., Elena, S. F. (2015). Evolution of plant virus movement proteins from the 30K superfamily and of their homologs integrated in plant genomes. Virology. 476: 304-315.

Namba, K., Pattanayek, R., Stubbs, G. (1989). Visualization of protein-nucleic acid interactions in a virus. J Mol Biol. 208: 307-325.

Neuwald, A., Liu, J., Lipman, D., Lawrence, C. (1997). Extracting protein alignment models from the sequence database. Nucl Acids Res. 25: 1665-1667.

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., Sninsky, J. J., Adams, M. D., Carqill, M. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3(6).

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G., Hermjakob, H. (2014). The MintAct project – IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 42(Database issue): D358-363.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., Chothia, C. (1998). Sequence comparisons using multiple sequences detect twice as many remote homologues as pairwise methods. J Mol Biol. 284: 1201 -1210.

Park, J., Teichmann, S. A., Hubbard, T., Chothia, C. (1997). Intermediate sequences increase the detection of homology between sequences. J Mo. Biol. 273: 249-254.

Pattanayek, R., Stubbs, G. (1992). Structure of the U2 strain of tobacco mosaic virus refined at 3.5 A resolution using X-ray fiber diffraction. J Mol Biol. 228(2): 516-528.

Paulding, C. A., Ruvolo, M., Haber, D. A. (2003). The Tre2 (USP6) oncogene is a hominoid-specific gene. Proc Natl Acad Sci. 100: 2507–2511.

Pei, J., Kim, B. H., Grishin, N. V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 36(7): 2295-2300.

Rice, D. W., Alverson, A. J., Richardson, A. O., Young, G. J., Sanchez-Puerta, M. V., Munzinger, J., Barry, K., Boore, J. L., Zhang, Y., de Pamphilis, C. W., Knox, E. B., Palmer, J. D. (2013). Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm Amborella. Science. 342(6165): 1468-1473.

Sassone-Corsi, P. (2002). Unique chromatin remodeling and transcriptional regulation in spermatogenesis. Science. 296(5576): 2176-2178.

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E., Ye, J. (2009). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 37(Database issue): D5-15.

She, X., Horvath, J. E., Jiang, Z., Liu, G., Furey, T. S., Christ, L., Clark, R., Graves, T., Gulden, C. L., Alkan, C., Bailey, J. A., Sahinalp, C., Rocchi, M., Haussler, D., Wilson, R. K., Miller, W., Schwartz, S., Eichler, E. E. (2004). The structure and evolution of centromeric transition regions within the human genome. Nature. 430: 857–864.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics. 21: 951-960.

Song, D., Cho, W. K., Park, S. H., Jo, Y., Kim, K. H. (2013). Evolution of and horizontal gene transfer in the Endornavirus genus. PLoS One. 8(5).

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34(Database issue): D535-539.

Stenglein, M. D., Leavitt, E. B., Abramovitch, M. A., McGuire, J. A., DeRisi, J. L. (2014). Genome sequence of a bornavirus recovered from an African Garter Snake (Elapsoidea loveridgei). Genome Announc. 2(5).

Stobbe, A. H., Melcher, U., Palmer, M. W., Roossinck, A. J., Shen, G. (2012). Co-divergence and host-switching in the evolution of tobamoviruses. Journal of General Virology. 93: 408- 418.

Stubbs, G. (1999). Tobacco mosaic virus particle structure and the initiation of disassembly. Phil Trans R Soc Lond B. 354: 551-557.

Zaiss, D. M., Kloetzel, P. M. (1999). A second gene encoding the mouse proteasome activator PA28beta subunit is part of a LINE1 element and is driven by a LINE1 promoter. J Mol Biol. 287: 829-835.

Zhdanov, V. M., Bogomolova, N. N., Gavrilov, V. I., Andzhaparidze, O. G., Deryabin, P. G., Astakhova, A. N. (1974). Infectious DNA of tick-borne encephalitis virus. Archiv für die gesamte Virusforschung. 45: 215-224.

Zhdanov, V. M., Parfanovich, M. I. (1974). Integration of measles virus nucleic acid into the cell genome. Archiv für die gesamte Virusforschung. 45: 225-234.

Zhou, H., Zhou, Y. (2004). Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins. 55: 1005-1013.

Zhou, H., Zhou, Y. (2005). Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins. 58: 321-328.

Zhou, X. (2013). L (3) of 03670 genes in the study of molecular mechanism of asymmetric cell division in the Drosophila neural stem. Master´s thesis. Zhejiang University.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S. (2013). MEGA6: molecular evolutionary genetics analüsys version 6.0. Molecular Biology and Evolution. 30: 2725-2729.

Tanne, E., Sela, I. (2005). Occurrence of a DNA sequence of a non-retro RNA virus in a host plant genome and its expression: evidence for recombination between viral and host RNAs. Virology. 332(2): 614-622.

Taylor, W. R. (1986). Identifcation of protein sequence homology by consensus template alignment. J Mol Biol. 188: 233-258.

Tewary, S. K., Oda, T., Kendall, A., Bian, W., Stubbs, G., Wong, S. M., Swaminathan, K. (2011). Structure of hibiscus latent singapore virus by fiber diffraction: a nonconserved his122 contributes to coat protein stability. J Mol Biol. 406(3): 516-526.

Umber, M., Filloux, D., Muller, E., Laboureau, N., Galzi, S., Roumagnac, P., Iskra-Caruana, M. L., Pavis, C., Teycheney, P. Y., Seal, S. E. (2014). The genomes of African yam (Dioscorea cayenensis-rotundata complex) hosts endogenous sequences from four distinct Badnavirus species. Mol Plant Pathol. 15(8): 790-801.

Wang, H., Stubbs, G. (1993). Molecular dynamics in refinement against fiber diffraction data. Acta Crystallogr A. 49(3): 504-513.

Wang, H., Stubbs, G. (1994). Structure determination of cucumber green mottle mosaic virus by X-ray fiber diffraction. Significance for the evolution of tobamoviruses. J Mol Biol. 239(3): 371-384.

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., Barton, G. J. (2009). Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. Bioinformatics. 25(9): 1189-1191.

Weiszmann, R., Hammonds, A. S., Celniker, S. E. (2009). Determination of gene expression patters using high-throughput RNA in situ hybridization to whole-mount Drosophila embryos. Nat Protoc. 4(5): 605-618.

Vernochet, C., Heidmann, O., Dupressoir, A., Cornelis, G., Dessen, P., Catzeflis, F., Heidmann, T. (2011). A syncytin-like endogenous retrovirus envelope gene of the guinea pig specifically expressed in the placenta junctional zone and conserved in Caviomprpha. Placenta. 32(11): 885-892.

Wiegmann, B. M., Trautwein, M. D., Winkler, I. S., Barr, N. B., Kim, J. W., Lambkin, C., Bertone, M. A., Cassel, B. K., Bayless, K. M., Heimberg, A. M., Wheeler, B. M., Peterson, K. J., Pape, T., Sinclair, B. J., Skevington, J. H., Blagoderov, V., Caravas, J., Kutty, S. N., Schmidt-Ott, U., Kampmeier, G. E., Thompson, F. C., Grimaldi, D. A., Beckenbach, A.

T., Courtney, G. W., Friedrich, M., Meier, R., Yeates, D. K. (2011). Episodic radiations in the fly tree of life. Proc Natl Acad Sci USA. 108(14): 5690-5695.

Wiegmann, B. M., Yeates, D. K., Thorne, J. L., Kishino, H. (2003). Time flies, a new molecular time-scale for brachyceran fly evolution without a clock. Syst Biol. 52(6): 745-756.

Vinckenbosch, N., Dupanloup, I., Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci. 103: 3220–3225.

Winterton, S. L., Wiegmann, B. M., Schlinger, E. I. (2007). Phylogeny and Bayesian divergence time estimations of small-headed flies (Diptera: Acroceridae) using multiple molecular markers. Mol Phylogenet Evol. 43(3): 808-832.

Wu, S., Zhang, Y. (2007). LOMETS: A local meta-threading-server for protein structure prediction. Nucleic Acids Research. 35: 3375-3382.

Wu, S., Zhang, Y. (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins. 72: 547-556.

Xu, D., Jaroszewski, L., Li, Z., Godzik, A. (2014). FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. Bioinformatics. 30(5): 660-667.

Xu, Y. and Xu, D. (2000). Protein threading using PROSPECT: design and evaluation. Proteins. 40: 343-354.

Yan, R., Xu, D., Yang, J., Walker, S., Zhang, Y. (2013). A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Sci Rep. 3: 2619.

Yang, M. Y., Wang, Z., MacPherson, M., Dow, J. A., Kaiser, K. (2000). A novel Drosophila alkaline phosphatase specific to the ellipsoid body of the adult brain and the lower Malphiginal (renal) tubule. Genetics. 154(1): 285-297.

Yeates, D. K., Wiegmann, B. M. (1999). Congruence and controversy: toward a higher-level phylogent of Diptera. Annu Rev Entomol. 44: 397-428.

Yu, J., Finley, R. L. Jr. (2009). Combining multiple positive training sets to generate confidence scores for protein-protein interactions. Bioinformatics. 25(1): 105-111.

Yu, J., Pacifico, S., Liu, G., Finley, R. L. Jr. (2008). DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. BMC Genomics. 9:461.

- **RAAMATUD**

Bertone, M. A., Wiegmann, B. M. (2009). True flies (Diptera). In Hedges, S. B., Kumar, S., The timetree of life, Oxford University Press, Oxford, UK.

Evenhuis, N. L. (1994). Catalogue of the fossil flies of the world (Insecta: Diptera). Blackhuys Publishers Lieden. 438(1).

Gradstein, F. M., Ogg, J. G., Smith, A. G. (2005). A geologic time scale 2004. Cambridge University Press. Cambridge, UK.

Grimaldi, D. I., Cumming, J. M. (1999). Brachyceran Diptera in Cretaceous ambers and Mesozoic diversification of the Eremoneura. Bulletin of the American Museum of Natural History (American Museum of Natural History). 239: 1–124

Krzeminski, W., Evenhuis, N. L. (2000). Review of Diptera paleontological records. *Contributions to a manual of Palaearctic Diptera*. 1: 535-564.

Labandeira, C. C. (2005). Fossil history and evolutionary ecology of Diptera and their associations with plants. In Yeates, D. K., Wiegmann, B. M (ed.), The evolutionary biology of flies, Columbia University Press, New York.

Marshall, S. A. (2006). Insects: their natural history and diversity: with a photographic guide to insects of Eastern North America. Firefly Books, Buffalo, USA.

Pisani, D. (2009). Arthropods (Arthropoda). In Hedges, S. B., Kumar, S., The timetree of life, Oxford University Press, Oxford, UK.

Yeates, D. K., Wiegmann, B. M. (2005). Phylogeny and evolution of Diptera: recent insights and new perspectives. In Yeates, D. K., Wiegmann, B. M (ed.), The evolutionary biology of flies, Columbia University Press, New York.

KASUTATUD VEEBIAADRESSID

- http://bgee.unil.ch/bgee/bgee
- http://consurf.tau.ac.il/
- http://digitalinsectcollection.wikispaces.com/Common+Green+Bottle+Fly?responseToken=0710d33594337d0a1d97719aeb0d364ab
- http://fastml.tau.ac.il/
- http://flyatlas.org/
- http://flybase.org/
- http://genevisible.com/search
- http://hmmer.janelia.org/
- http://mentha.uniroma2.it/
- http://prodata.swmed.edu/promals3d/promals3d.php
- http://supfam.org/SUPERFAMILY/
- http://zhanglab.ccmb.med.umich.edu/LOMETS/
- http://thebiogrid.org/
- http://www.droidb.org/
- http://www.ebi.ac.uk/intact/
- http://www.ebi.ac.uk/Tools/webservices/psicquic/view/main.xhtml
- http://www.flymine.org/
- http://www.modencode.org/
- http://www.ncbi.nlm.nih.gov/
- http://www.peptideatlas.org/
- http://www.rcsb.org/pdb/home/home.do
- http://www.targetscan.org/
- https://genevestigator.com/gv/
- https://www.pymol.org/

LISAD

LISA 1. SCOP andmebaasi (versioon 1.75; Lo Conte *et al.*, 2002) valgudomeenide klassifitseerimise hierarhia ning lühitutvustused.

Valgud klassifitseeritakse hierarhiliselt mitmete tasemete abil:

- Klass (ing k *class)*. Valgud jagatakse sekundaarstruktuuri alusel klassidesse.
    - Ainult alpha struktuurid (ing k *all alpha proteins*). Valgudomeen koosneb põhiliselt vaid alpha-heeliksist. (Murzin *et al*., 1995)
    - Ainult beeta struktuurid (ing k *all beta proteins*). Valgudomeen koosneb põhiliselt vaid beeta-lehtedest. (Murzin *et al*., 1995)
    - Alpha ja beeta struktuurid (ing k *alpha and beta proteins; α/β*). Valgudomeen koosneb segamini alpha-heeliksitest ja beeta-lehtedest. (Murzin *et al*., 1995)
    - Alpha pluss beeta struktuurid (ing k *alpha plus beta proteins; α+β*). Valgudomeen koosneb eraldatud alpha-heelikstest ja beeta-lehtedest. (Murzin *et al*., 1995)
    - Multi-domeenide strktuurid (ing k *multi-domain proteins*). Valgudomeenid, millel pole homolooge teada (Murzin *et al*., 1995) ning mille domeenid kuuluvad mitme erineva klassi alla (http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.html, 14.04.2015).
    - Membraani- ja raku pinnastruktuurid (ing k *membrane and cell surface proteins and peptides*). Siia klassi ei kuulu immuunsüsteemi valgud (http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.html, 14.04.2015).
    - Väikesed valgud (ing k *small proteins*). Üldjuhul on domineeritud metalli-ligandide, heme ja/või disulfiidsildade poolt (http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.html, 14.04.2015).
- Pakkimise ehk voltumise tase (ing k *common fold*). Valgudomeenid, mis omavad sarnaseid suuremaid sekundaarstruktuuri motiive sarnases asetuses samade topoloogiliste ühendustega. Peamisteks erinevusteks on perifeerias asuvad elemendid ning pöörde-elementide suurused ja konformatsioonid. Struktuuride sarnasused on tekkinud arvatavasti valkude füüsikaliste ja keemiliste omaduste poolt, mis eelistavad kindlaid pakkimise motiive ning ahela topoloogiaid. (Murzin *et al*., 1995)
- Superperekond (ing k *superfamily*). Valgudomeenid, millel esineb väikene järjestuste sarnasus, kuid mille struktuurid (seal hulgas funktsionaalsus) viitab ühisele evolutsioonilisele päritolule (Murzin *et al*., 1995).

- Perekond (ing k *family*). Valgudomeenid, mille järjestuse identsus on vähemalt 30% ning mille funktsioon ja struktuur on väga sarnased (Murzin *et al*., 1995).
- Valk (ing k *protein*). Valgudomeenid, millel on järjestused ja funktsioon identsed, aga pärinevad bioloogiliselt erinevatelt organismidelt või esindavad ühe organismi erinevaid isovorme (Andreeva *et al*., 2008).
- Liik (ing k *species*). Eraldiseisvad valgujärjestused ning nende nii looduslikud kui ka tehislikud alternatiivid (Andreeva *et al*., 2008).

LISA 2. Tabel uuritava TMV-CP$_{fly}$ ning naabergeenide esimenise kohta kärbselistes.

Tabel 1. Uuritava geeni TMV-CP$_{fly}$ (märgitud *CG15772*) ning naabergeenide (*Drosophila melanogaster* geeninimed: *IntS6*, *CG4078*, *lin-52*, *CG15771*; dos Santos *et al*., 2015) esinemine kärbselistes ning nende detekteerimine (genoom – DNA või transkriptoom – RNA). Punasest joonest jäävad ülespoole *Schizophora* kärbselistest eelnevalt lahknenud organismid ning allapoole *Acalyptratae* kärbselised, sinisest joonest allapoole jäävad aga *Calyptratae* kärbselised. Tumedamas kirjas on välja toodud liigid, kelle genoom on täielikult sekveneeritud. Sulgudes on välja toodud bioprojektide arvud sugukonna, perekonna ning liigi tasemel NCBI bioprojektide andmebaasi järgi (Sayers *et al*., 2009). Lisaks on välja toodud mRNA ning valkude arv NCBI nukleotiidide ning valkude andmebaasi järgi (Sayers *et al*., 2009). *CG15772* puhul tumedamalt märgitud esinemised on tuvastatud konstrueeritud ansestraalsete järjestuste abil BLAST otsingu tulemusena eelnevate tulemustele lisaks ning rohelisena on märgitud uue BLAST´imise (12.05.2015 seisuga) detekteeritud uued genoomi või transkriptoomi andmed (Sayers *et al*., 2009).

Sugukond	**Perekond**	**Liik**	**IntS6**	**CG4078**	**CG15772 (TMV-CP$_{fly}$)**	**lin-52**	**CG15771**	**mRNA arv**	**Valkude arv**
Bombyliidae (4)	Bombylius (1)	Bombylius major (1)	RNA	RNA		RNA	RNA	2	27
Phoridae (5)	Megaselia (5)	**Megaselia scalaris (3)**	DNA	DNA		DNA		13	93
Chloropidae (1)	Lipara (1)	Lipara lucens (1)	RNA	RNA	RNA	RNA	RNA	0	0
Diopsidae (7)	Teleopsis (7)	Teleopsis dalmanni (5)	RNA	RNA	DNA/ RNA	RNA	RNA	9	215
		Teleopsis whitei (1)	RNA	RNA	DNA/ RNA	RNA	RNA	0	34
Drosophilidae (1536)	Drosophila (1534)	Drosophila albomicans (2)	DNA	DNA	DNA	DNA	DNA	1	73
		Drosophila mojavensis (10)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	14627	30501
		Drosophila virilis (18)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	14588	30058
		Drosophila grimshawi (3)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	15003	30200
		Drosophila ananassae (14)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	15106	31831
		Drosophila bipectinata (2)	DNA	DNA	DNA	DNA	DNA	3	190
		Drosophila malerkotliana malerkotliana (2)				RNA	RNA	9595	46
		Drosophila malerkotliana pallens (1)	RNA	RNA		RNA	RNA	19162	41
		Drosophila pseudoananassae nigrens (1)	RNA	RNA	DNA/ RNA	RNA	RNA	29642	22
		Drosophila pseudoananassae pseudoananassae (1)	RNA			RNA	RNA	32925	20
		Drosophila ercepeae (1)	RNA			DNA/RNA		15457	12
		Drosophila merina (1)	RNA	RNA		RNA		19354	11
		Drosophila elegans (2)	DNA	DNA	DNA	DNA	DNA	1	70

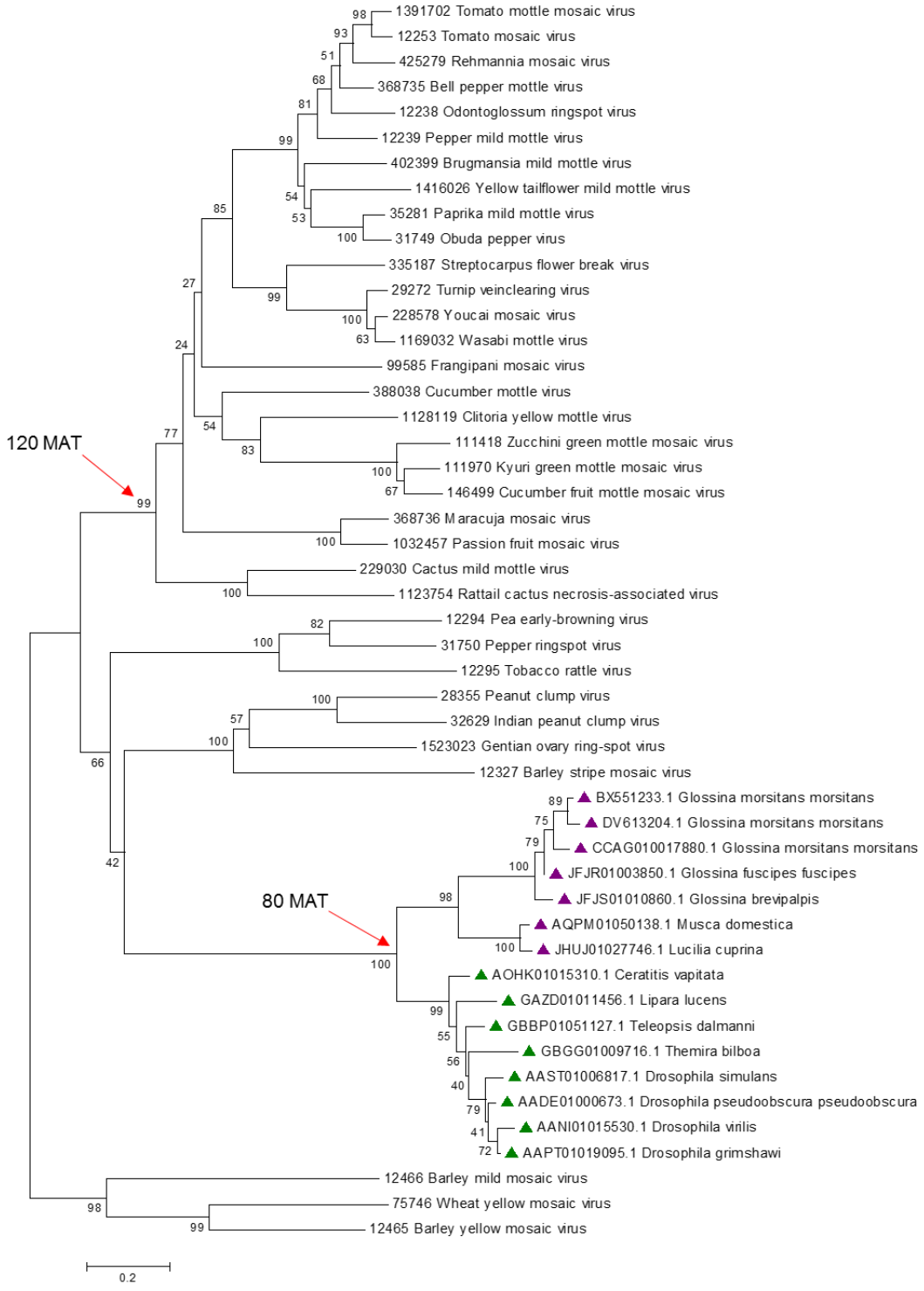
		Drosophila eugracilis (2)	DNA	DNA	DNA	DNA	DNA	5	55
		Drosophila ficusphila (2)	DNA	DNA	DNA	DNA	DNA	2	45
		Drosophila erecta (6)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	15088	30632
		Drosophila melanogaster (1361)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	80653	132379
		Drosophila sechellia (7)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	16488	33749
		Drosophila simulans (41)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	15748	43895
		Drosophila yakuba (23)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	16901	34972
		Drosophila auraria		DNA		RNA		3	66
		Drosophila kikkawai (2)	DNA	DNA	DNA	DNA	DNA	2	235
		Drosophila serrata (4)		RNA	DNA/ RNA	DNA/ RNA	RNA	9364	85
		Drosophila rhopaloa (2)	DNA	DNA	DNA	DNA	DNA	4	6
		Drosophila biarmipes (2)	DNA	DNA	DNA	DNA	DNA	0	45
		Drosophila suzukii (4)	DNA	DNA	DNA	DNA	DNA	1	122
		Drosophila takahashii (2)	DNA	DNA	DNA	DNA	DNA	19	84
		Drosophila teissieri		DNA				15	579
		Drosophila miranda (6)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	2	2671
		Drosophila persimilis (3)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA	DNA/ RNA	16913	34219
		Drosophila pseudoobscura (22)	RNA	RNA		RNA	RNA	121	5428
		Drosophila pseudoobscura pseudoobscura (3)	DNA/RNA	DNA/RNA	DNA/ **RNA**	DNA/RNA	DNA/RNA	16875	35227
		Drosophila willistoni (5)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	15540	31394
Sepsidae (3)	Themira (2)	Themira biloba (1)	RNA	RNA	DNA/ RNA	RNA	RNA	0	6
Tephritidae (55)	Ceratitis (15)	Ceratitis capitata (15)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	45009	45986
	Bactrocera (35)	Bactrocera dorsalis (21)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	18089	44702
		Bactrocera tryoni (4)	DNA	DNA	DNA	DNA	DNA	11	203
		Bactrocera oleae (4)	RNA	RNA		RNA	RNA	228	737
		Bactrocera minax (2)	RNA	RNA		RNA	RNA	11	456
		Bactrocera cucurbitae (3)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	20754	38857
	Rhagoletis (2)	Rhagoletis pomonella (2)				RNA	RNA	24376	138

Glossinidae (19)	Glossina (19)	Glossina brevipalpis (1)	DNA	DNA	DNA	DNA		0	6
		Glossina austeni (1)	DNA	DNA	DNA	DNA		2	9
		Glossina pallidipes (1)	DNA	DNA	DNA	DNA		5	29
		Glossina morsitans morsitans (8)	DNA	DNA/ RNA	DNA/ RNA	DNA	DNA	2763	2780
		Glossina fuscipes fuscipes (2)	DNA	DNA	DNA	DNA		2	88
		Glossina palpalis gambiensis (3)	DNA	DNA	DNA	DNA		0	55
Anthomyiidae (2)	Delia (1)	Delia antiqua (1)	RNA	RNA		RNA	RNA	44	282
Muscidae (25)	Musca (16)	Musca domestica (16)	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	DNA/ RNA	26300	26407
Calliphoridae (16)	Lucilia (5)	Lucilia cuprina (3)	DNA	DNA/ RNA	DNA/ **RNA**	DNA	DNA	64	547
	Cochliomyia (7)	Cochliomyia hominivorax (5)		RNA				35	354
Sarcophagidae (9)	Sarcophaga (5)	Sarcophaga crassipalpis (5)					RNA	21024	168
Tachinidae (4)	Triarthria (1)	Triarthria setipennis (1)	RNA	RNA		RNA		0	4

LISA 3. LOMETS´a metaprogrammide nimekiri ning nende lühitutvustus.

- **cdPPAS**, **Neff-PPAS** ja **wdPPAS** on järjestuse profiil-profiil võrdluse meetodid, mis võtavad skoori arvutamisel arvesse ka sekundaarstruktuuri (Yan *et al*., 2013).
- **FFAS03** on profiil-profiil võrdluse meetod, mis võtab arvesse ka valkude pakkimise viise (Jaroszewski *et al*., 2005). **FFAS-3D** on originaalse FFAS modifitseeritud versioon, mis võtab skoori arvutamisel arvesse üksikut järjestust (Xu *et al*., 2014).
- **MUSTER** profiil-profiil võrdluse meetod, mis võtab arvesse ka struktuurset informatsiooni, lahuse ligipääsetavust, selgroo nurkasid ning üldist hüdrofoobsust (Wu ja Zhang, 2008).
- **HHSEARCH** kasutab HMM-HMM võrdluse meetodit. Antud meetodi baail on loodud **HHSEARCH 1** ja **HHSEARCH 2**, mis on kirjeldatud meetodi modifitseeritud versioonid (Soding, 2005).
- **pGenTHREADER** on parameetriline profiil-profiil võrdluste meetod, mis võtab arvesse valkude pakkimisviisi (Lobley *et al*., 2009).
- **PRC** on profiil-HMM-profiil-HMM võrdluse meetod, mis arvestab nii otsitava kui tulemuste profile (Madera, 2008).
- **PROSPECT2** võtab skoori arvutamisel arvesse järjestuste mutatsioone, sekundaarstruktuuri, lahuse ligipääsetavust ning jääkide paariviisilisi kontakte (Xu ja Xu, 2000).
- **SPARKS-X** võtab skoori arvutamise arvesse nii järjestust kui ka sekundaarstruktuuri (Zhou ja Zhou, 2004).
- **SP3** kasutab profiil-profiil joondust, mis võtab arvesse nii järjestust kui ka struktuuri (Zhou ja Zhou, 2005).

LISA 4. Struktuuri arvestamata konstrueeritud fülogeneesipuu.



Joonis 20. TMV-CP viiruslike ning eukarüootsete järjestuste baasil konstrueeritud fülogeneetiline NJ puu (Poissoni korrektsiooni mudeli baasil paariviisiliste gappide deleteerimisega), mis on juuritud bümoviiruste baasil. All on märgitud aminohapete asenduste arvu positsiooni kohta (0.2). Antud analüüsis osales 49 järjestust 216 positsiooniga. Roheliste kolmnurkadega on märgitud *Acalyptratae* kärbselised ning lillade kolmnurkadega *Calyptratae* kärbselised).

LISA 5. Tabel uuritava TMV-CP$_{fly}$ geeniekspressiooni kohta kärbselistes.

Tabel 2. Ekspressiooni andmebaaside vasted, mis on kategoriseeritud kudede ning kirjeldatud soo ja/või arengustaadiumi baasil. Välja on toodud *probe´*ide arv antud koe ja soo/staadiumi uuringus ning ekspressioonitase (H – kõrge, MH – kõrgem keskmine, M – keskmine, L – madal) vastavate andmebaaside kohta (Bgee, Genevestigator – GV, FlyMine – FM). Osad koed on suuremate kudede piires kokku võetud (täpsustused on *probeset´*i juures välja toodud). (Bastian *et al*., 2008; Hruz *et al*., 2008; Lyne *et al*., 2007)

Organ/stage	Male	Female	Sex not assigned	Embryo	Larva	Pupal	Adult
Capitellum[1]							3 probeset (Bgee) - H
Basiproboscis[2]							5 probeset (Bgee) - H
Antenna							4 probeset (Bgee) - H
Anatomical structure						3 probeset (P-stage) (Bgee) – H 7 probeset (prepupal) (Bgee) - H	1 probeset (A3 stage) (Bgee) - H
Abdomen		22 probeset (GV) – L-M					22 probeset (GV) – L-M
Brain		3 probeset (GV) - L	12 probeset (GV) – H/MH 12 probeset (GV) – M 3 probeset (GV) - L	3 probeset (GV) - L	12 probeset (GV) - M		15 probeset (GV) L-H 13 probeset (Bgee) – H 3 probeset (haltere pedicel[3]) (Bgee) - H
Central nervous system			4 probeset (GV) - M		4 probeset (GV) - M		
Crop[4]			4 probeset (GV) - M				
Ejaculatory duct	4 probeset (GV) - M						
Eye			7 probeset (GV) - M				

[1] Luude ühendus
[2] Londi osa
[3] Aju osa
[4] Pugu

Eyedisc			8 probeset (GV) - M				
Fat body		19 probeset (GV) - M	31 probeset (GV) – M-H 4 probeset (GV) - M		31 probeset (GV) – M-H		4 probeset (GV) – M 19 probeset (GV) - M
Head	24 probeset (GV) – H (20 day) (FM) – M (1day, 4 day) (FM) - MH	(1 day) (FM) – M (4day) (FM) - MH	76 probeset (GV) – H/MH				100 probeset (GV) – MH-H 123 probeset (Bgee) - H 2 probeset (P13 - pharate) (Bgee) – H 3 probeset (Bgee) - H
Head capsule							4 probeset (Bgee) - H
Heart			4 probeset (GV) - M				4 probeset (GV) - M
Hindgut			4 probeset (GV) – L-M 4 probeset (GV) – L-M		4 probeset (GV) – L-M		4 probeset (GV) – L-M
Lower reproductive tract		4 probeset (GV) - M					
Malpighian tubule[5]			8 probeset (GV) - M		8 probeset (GV) - M		
Midgut		18 probeset (GV) - M	4 probeset (GV) – M 4 probeset (GV) - M		4 probeset (GV) - M		22 probeset (GV) - M
Ovary		10 probeset (GV) – M-H					
Proboscis[6]		5 probeset (GV) - M					
Salivary gland			12 probeset (GV) – M 4 probeset (GV) – M 4 probeset (GV) - M		4 probeset (GV) - M	12 probeset (GV) - M	4 probeset (GV) - M

[5] Seedesüsteemi organid
[6] Lont

Tagma[7]							8 probeset (Bgee) – H 8 probeset (Bgee) - H
Testis	16 probeset (GV) – M 3 probeset (apical testis) (GV) – M 3 probeset (proximal testis) (GV) – M 3 probeset (distal testis) (GV) - M	12 probeset (spermathecum) (GV) - M					2 probeset (A1 stage) (Bgee) - H
Trachea			4 probeset (GV) – L-M		4 probeset (GV) – L-M		
Whole body	4 probeset (GV) – M 196 probeset (GV) – M-H 12 probeset (GV, third instar) - M	4 probeset (GV) – M 193 probeset (GV) – M-H 6 probeset (third instar) (GV) - M	75 probeset (GV) – M-H 21 probeset (third instar) (GV) – L-H 8 probeset (first instar) (GV) – M 4 probeset (early third instar) (GV) – L-M 27 probeset (second instar) (GV) - L	63 probeset (GV) – L-M	39 probeset (third instar) (GV) – L-H 8 probeset (first instar) (GV) – M 4 probeset (early third instar) (GV) – L-M 27 probeset (second instar) (GV) - L	8 probeset (prepupa) (GV) – M (P15, P9-10) (FM) - M	464 probeset (GV) – M-H
Wing imaginal disc[8]			36 probeset (GV) – L(M)				

[7] Mitme organi ühendnimetud, mis seob neid funktsionaalseks morfoloogiliseks ühikuks. Tihti putukatel näiteks pea, ringmik ning alakeha.
[8] Putukavastse osa, millest moodustuvad nuku transformatsiooni staadiumis täiskasvanud organismi välised osad.

LISA 6. Tabel BLAST EST andmebaasi tulemuste iseloomustus uuritava TMV-CP_{fly} kohta.

Tabel 3. EST andmebaasi vasted otsingule „*Drosohphila melanogaster CG15772* regioon“ (Sayers *et al*., 2009). Välja on toodud NCBI accession ID, nimi, BLAST (TBLASTN) otsingu bit-skoor ning E-väärtus. Lisaks bioproovide andmed vanuse, soo ja koe kohta. AP – andmed puuduvad.

NCBI ligipääsu ID	Nimi	Bit-skoor	E-väärtus	Vanus	Sugu	Kude
FG088519.1	*Ceratitis capitata*	257.684	1.43E-84	AP	AP	Pea
FG088519.1	*Ceratitis capitata*	257.684	5.43E-85	AP	AP	Pea
BF491991.2	*Drosophila melanogaster*	179.104	6.72E-55	AP	Isasorganism	Täiskasvanud organismi testis
BF491991.2	*Drosophila melanogaster*	179.104	2.56E-55	AP	Isasorganism	Täiskasvanud organismi testis
BF487486.2	*Drosophila melanogaster*	160.614	1.10E-47	AP	Isasorganism	Täiskasvanud organismi testis
BF487486.2	*Drosophila melanogaster*	160.614	4.18E-48	AP	Isasorganism	Täiskasvanud organismi testis
BI568953.1	*Drosophila melanogaster*	151.369	5.46E-44	AP	Isas- ja emasorganism	Pea
BI568953.1	*Drosophila melanogaster*	151.369	2.08E-44	AP	Isas- ja emasorganism	Pea
BI629955.1	*Drosophila melanogaster*	91.2781	2.25E-21	AP	Isas- ja emasorganism	Pea
BI629955.1	*Drosophila melanogaster*	91.2781	5.91E-21	AP	Isas- ja emasorganism	Pea
BI589607.1	*Drosophila melanogaster*	91.2781	2.25E-21	AP	Isas- ja emasorganism	Pea
BI589607.1	*Drosophila melanogaster*	91.2781	5.91E-21	AP	Isas- ja emasorganism	Pea
BI620843.1	*Drosophila melanogaster*	78.9518	5.94E-17	AP	Isas- ja emasorganism	Pea
BI620843.1	*Drosophila melanogaster*	78.9518	1.56E-16	AP	Isas- ja emasorganism	Pea
BI627728.1	*Drosophila melanogaster*	75.485	1.09E-15	AP	Isas- ja emasorganism	Pea
BI627728.1	*Drosophila melanogaster*	75.485	2.88E-15	AP	Isas- ja emasorganism	Pea

CK659670.1	*Drosophila melanogaster*	123.25	6.04E-33	AP	Isas- ja emasorganism	Terve organism
CK659670.1	*Drosophila melanogaster*	123.25	2.30E-33	AP	Isas- ja emasorganism	Terve organism
AI388163.1	*Drosophila melanogaster*	116.701	7.41E-31	AP	Isas- ja emasorganism	Pea
AI388163.1	*Drosophila melanogaster*	116.701	1.95E-30	AP	Isas- ja emasorganism	Pea
AW940721.1	*Drosophila melanogaster*	80.4925	8.20E-18	AP	Isas- ja emasorganism	Pea
AW940721.1	*Drosophila melanogaster*	80.4925	2.15E-17	AP	Isas- ja emasorganism	Pea
EC198829.1	*Drosophila melanogaster*	167.162	1.03E-50	AP	AP	AP
EC198829.1	*Drosophila melanogaster*	167.162	3.93E-51	AP	AP	AP
EB558208.1	*Drosophila virilis*	103.605	1.03E-25	AP	AP	Terve organismi embrüo ja täiskasvanud staadiumi RNA segu
EB558208.1	*Drosophila virilis*	103.605	2.69E-25	AP	AP	Terve organismi embrüo ja täiskasvanud staadiumi RNA segu
DV613204.1	*Glossina morsitans morsitans*	151.754	5.20E-43	AP	Isas- ja emasorganism	Immuun-stimuleeritud organsimi rasvkeha
DV613204.1	*Glossina morsitans morsitans*	151.754	1.98E-43	AP	Isas- ja emasorganism	Immuun-stimuleeritud organsimi rasvkeha
DV617147.1	*Glossina morsitans morsitans*	151.754	9.02E-43	AP	Isas- ja emasorganism	Immuun-stimuleeritud organsimi rasvkeha
DV617147.1	*Glossina morsitans morsitans*	151.754	3.44E-43	AP	Isas- ja emasorganism	Immuun-stimuleeritud organsimi rasvkeha
DV602461.1	*Glossina morsitans morsitans*	52.7582	5.59E-08	AP	Isas- ja emasorganism	Immuun-stimuleeritud organsimi rasvkeha
DV602461.1	*Glossina morsitans morsitans*	52.7582	1.47E-07	AP	Isas- ja emasorganism	Immuun-stimuleeritud organsimi rasvkeha
DV620083.1	*Glossina morsitans morsitans*	59.3066	6.90E-10	AP	Isas- ja emasorganism	Immuun-stimuleeritud organsimi rasvkeha
DV620083.1	*Glossina morsitans morsitans*	59.3066	1.81E-09	AP	Isas- ja emasorganism	Immuun-stimuleeritud organsimi rasvkeha

BX551233.1	*Glossina morsitans morsitans*	155.221	1.13E-45	AP	AP	Täiskasvanud organismi infekteeritud seedekulgla
BX551233.1	*Glossina morsitans morsitans*	155.221	4.31E-46	AP	AP	Täiskasvanud organismi infekteeritud seedekulgla
BX555471.1	*Glossina morsitans morsitans*	60.8474	8.41E-11	AP	AP	Täiskasvanud organismi infekteeritud seedekulgla
BX555471.1	*Glossina morsitans morsitans*	60.8474	2.21E-10	AP	AP	Täiskasvanud organismi infekteeritud seedekulgla
FM982036.1	*Glossina morsitans morsitans*	122.865	1.80E-32	AP	AP	T. brucei infekteeritud süljenäärmed
FM982036.1	*Glossina morsitans morsitans*	122.865	6.86E-33	AP	AP	T. brucei infekteeritud süljenäärmed
FM965076.1	*Glossina morsitans morsitans*	115.546	4.44E-30	AP	AP	T. brucei infekteeritud süljenäärmed
FM965076.1	*Glossina morsitans morsitans*	115.546	1.17E-29	AP	AP	T. brucei infekteeritud süljenäärmed

LISA 7. Tabel BLAST TSA andmebaasi tulemuste iseloomustus uuritava TMV-CP_{fly} kohta.

Tabel 4. TSA andmebaasi vasted otsingule „*Drosophila melanogaster CG15772* regioon“ (Sayers *et al., 2009*). Välja on toodud NCBI accession ID, nimi ning BLAST otsingu bit-skoor ja E-väärtus. Lisaks biorpoovide andmed: vanus, sugu ja kude. AP – andmed puuduvad.

NCBI ligipääsu ID	Nimi	Bit-koor	E-väärtus	Vanus	Sugu	Kude
GBXI01004483.1	*Bactrocera cucurbitae*	259.225	6.06E-81	segu	Isas- ja emasorganism	Terve organism
GBXI01004483.1	*Bactrocera cucurbitae*	259.225	5.93E-82	segu	Isas- ja emasorganism	Terve organism
GAKP01003220.1	*Bactrocera dorsalis*	259.996	1.89E-81	AP	AP	AP
GAKP01003220.1	*Bactrocera dorsalis*	259.996	1.85E-82	AP	AP	AP
GAMC01007375.1	*Ceratitis capitata*	259.61	1.26E-82	AP	AP	Terve organism
GAMC01007375.1	*Ceratitis capitata*	259.61	1.23E-83	AP	AP	Terve organism
GAZD01011456.1	*Lipara lucens*	250.366	1.29E-81	AP	AP	AP
GAZD01011456.1	*Lipara lucens*	250.366	1.27E-82	AP	AP	AP
GBBP01051127.1	*Teleopsis dalmanni*	270.011	1.04E-87	4 nädalat	Isasorganism	Pea
GBBP01051127.1	*Teleopsis dalmanni*	270.011	1.02E-88	4 nädalat	Isasorganism	Pea
GBBQ01050027.1	*Teleopsis whitei*	148.673	4.73E-43	4 nädalat	Isasorganism	Testis
GBBQ01050027.1	*Teleopsis whitei*	148.673	4.64E-44	4 nädalat	Isasorganism	Testis
GBBQ01049690.1	*Teleopsis whitei*	118.242	5.90E-32	4 nädalat	Isasorganism	Testis
GBBQ01049690.1	*Teleopsis whitei*	118.242	5.78E-33	4 nädalat	Isasorganism	Testis
GBGG01009716.1	*Themira biloba*	238.81	7.21E-78	72 tunnine nukk	AP	Terve organism
GBGG01009716.1	*Themira biloba*	238.81	7.06E-79	72 tunnine nukk	AP	Terve organism

LISA 8. Tabel TMV-CP$_{fly}$ valk-valk interaktsioonide kohta.

Tabel 5. *CG15772* valgu interaktsioonid teiste valkudega ning nendele vastavate skooride väärtused erinevate andmebaaside piires. Rohelisega on välja toodud esimese katse autori poolt välja toodud kõrged ning oranžiga madalad skoorid. (Giot *et al*., 2003; Orchard *et al*., 2014; Licata *et al*., 2012; Yu ja Finley, 2009)

Valk 1	**Valk 2**	**Esimene autor**	**Interaktsiooni detekteerimise meetod**	**Autori skoor (0 ... 1)**	**IntAct skoor (0 ... 1)**	**Mentha skoor (0 ... 1)**	**DroiD skoor (0 ... 1)**
CG15772	**CG1316**, Q9VZE4	Giot *et al*. (2003)	Finley labori kaksik-hübriid Curagen´i kaksik-hübriid	0.56	0.37	0.183	0.428
CG15772	**CG33129**, CG6089, Q9VKM7	Giot *et al*. (2003)	Curagen´i kaksik-hübriid	0.14	0.37	0.183	0.411
CG15772	**CG3776**, l(2)k03704, Jhebp29, Q9W0X3	Giot *et al*. (2003)	Curagen´i kaksik-hübriid	0.33	0.37	0.183	-
CG15772	**CG8199**, Q9VHB8	Giot *et al*. (2003)	Curagen´i kaksik-hübriid	0.33	0.37	0.183	0.383
CG15772	**comt**, CG1618, Q9VYF4	Giot *et al*. (2003)	Curagen´i kaksik-hübriid	0.42	0.37	0.21	0.324
CG15772	**DNApol-δ**, CG5949, Q9VUW8	Giot *et al*. (2003)	Curagen´i kaksik-hübriid	0.61	0.37	0.21	0.345
CG15772	**Fak**, CG10023, B7YZM0	Giot *et al*. (2003)	Curagen´i kaksik-hübriid	0.34	0.37	0.183	0.374
CG15772	**l(3)03670**, CG1715, Q9VA18	Giot *et al*. (2003)	Finley labori kaksik-hübriid Curagen´i kaksik-hübriid	0.98	0.37	0.183	0.418
CG15772	**Ripα**, CG18145	Giot *et al*. (2003)	Curagen´i kaksik-hübriid	0.42	0.37	0.183	0.406
CG15772	**sina**, CG9949, Q9VVB0	Giot *et al*. (2003)	Curagen´i kaksik-hübriid	0.22	0.37	0.21	0.403
CG15772	**Tim9a**, CG1660, Q1ECC2	Giot *et al*. (2003)	Curagen´i kaksik-hübriid	0.26	0.37	0.21	0.414

LISA 9. Valk-valk interaktsioonide andmebaaside skooride skaalad ning arvutuste lahtiseletused.

Esimese eksperimendi autori poolt määratud skoorid on välja toodud skaalas 0-1. Antud töös määrati olulisuse lävendiks 0.5 oluliste interaktsioonide jaoks (Giot *et al*., 2003).

IntAct andmebaasi poolt määratud skoorid on välja toodud skaalas 0-1, mis on määratud manuaalse annotatsiooni poolt, võttes arvesse kasutatud meetodit, interaktsiooni tüüpi, publikatsioonide arvu (maksimaalselt 8 tükki). Iga parameeter on normaliseeritud skaalale 0-1 ning ühe interaktsiooni kumulatiivne skoor on lisaks uuesti normaliseeritud skaalale 0-1. (http://www.ebi.ac.uk/intact/faq , 11.05.2015)

Mentha andmebaasi poolt määratud skoorid on välja toodud skaalas 0-1. Skoor võtab arvesse mitmeid andmeid: ekperimendi suurus ja tüüp ning publikatsioonide arv. Näiteks skoor 1 tähendab seda, et antud interaktsioon on kirjeldatud mitmete publikatsioonide ning eksperimenditüüpide poolt. Skoor 0.3 aga antud interaktsioon on tõendatud ühe kõrgekvaliteetse eksperimendi poolt. (Licata *et al*., 2012)

DroiD andmebaasi poolt määratud skoorid on välja toodud skaalas 0-1. Mida kõrgem skoor, seda tõenäolisemalt on oluline interaktsioon, võrreldes madalamate skooridega interaktsioonidega. (Yu ja Finley, 2009)

LISA 10. Tabel TMV-CP$_{fly}$ geeniga interakteeruvate mikroRNA´de kohta.

Tabel 6. *CG15772* geeni interaktsioonid mikroRNA´dega, koos ekspressiooni staadiumi, bioloogilistes protsessides osalemise ning fenotüübi kirjeldustega. Konserveeruvus *Dorosophila* perekonnas baseerub perekonna esindajate arvukusel, kellel antud seondumiskoht esineb. Tabel on koostatud DroID andmebaasi abil (Yu ja Finley, 2009). Interaktsioonide konserveeruvused on võetud TargetScan andmebaasist (versioon 6.2, juuni 2012; Grimson *et al*., 2008).

miRNA nimi	**Seondumisasukoht CG15772 geenil**	**Konserveeruvus *Drosophila* perekonnas**	**GO bioloogiline protsess**	**RNA ekspressiooni staadium**	**Fenotüüp**
mir-8	140-146	Nõrgalt	Geenivaigistamine, Wnt signalisatsioonirada, epidermaalsete kasvufaktorite retseptorite siganlisatsiooniraja negatiivne regulaator, gliiarakkude areng ning kasv, neuroblastide areng, neurogeneesi regatiivne regulaator, valmiku koorumise positiivne regulaator	Embrüonaalne ja täiskasvanud staadium	Väike keha, arengukiirus defektne, vähenenud rakkude arvukus, osaliselt ja täielikult letaalne, *locomotor* käituvus defektne, neurofüsioloogia defektne
mir-124	237-243	Tugevalt	Geenivaigistamine, sünaptiliste transmissioonide negatiivne regulatsioon, dendriitide morfogenees, neuroblastide proliferatsioon	Embrüonaalne staadium, larva 1L staadium	Surmav enne nuku staadiumi lõppu, neuroanatoomia defektne, *locomotor* käitumine defektiivne, paaritusmiskäitumine defektne
mir-277	129-135	Nõrgalt	Neuroni surma positiivne regulaator, geenivaigistamine	-	Osaliselt surmav nuku staadiumis, kuuma-stressi vastus puudulik, neuroanatoomia puudulik.
mir-4	356-362	Tugevalt	-	Embrüonaalne staadium, larva 1L staadium	Nähtav
mir-932	3-10	Nõrgalt	*Smoothened* signalisatsiooniraja regulatsioon	-	Surmav, kõik surevad enne larva/nuku staadiumi lõppemist.
mir-983	141-147	Nõrgalt			

LISA 11. TMV-CP_{fly} geeniga interakteeruvad transkriptsioonifaktorid ning nende osalused bioloogilistes protsessides koos GO koodidega DroID andmete järgi (Yu ja Finley, 2009).

Cad (FBgn0000251; CG1759). *Biological GO process: positive regulation of cell proliferation* (0008284), *anterior/posterior axis specification* (0009948), *analia development* (0007487), *positive regulation of transcription from RNA polymerase II promoter* (0045944), *germ-band extension* (0007377), *positive regulation of antimicrobial peptide biosynthetic process* (0002807), *Malpighian tubule morphogenesis* (0007443), *hindgut morphogenesis* (0007442), *blastoderm segmentation* (0007350), *gastrulation involving germ band extension* (0010004), *genital disc anterior/posterior pattern formation* (0035224), *segment specification* (0007379), *genital disc development* (0035215).

Kr (FBgn0001325; CG3340). *Biological GO process: Malpighian tubule bud morphogenesis (0061332), neuroblast fate determination* (0007400), *regulation of development* (0040034), *heterochronic, regulation of hemocyte proliferation* (0035206), *wing disc development* (0035220), *compound eye development* (0048749), *zygotic determination of anterior/posterior axis* (0007354), *embryo, negative regulation of transcription from RNA polymerase II promoter* (0000122), *ventral cord development* (0007419), *ganglion mother cell fate determination* (0007402), *negative regulation of transcription* (0045892), *DNA-templated, muscle organ development* (0007517), *Malpighian tubule morphogenesis* (0007443), *chromatin silencing* (0006342), *axon guidance* (0007411), *trunk segmentation* (0035290), *negative regulation of gene expression* (0010629).

LISA 12. Kirjanduse andmete järgi TMV-CP struktuuris olulised aminohapped.

Sinisega on märgitud hüdrofiilse keskkonna loomiseks ning RNA-valk interaktsioonides osalevad ühised aminohapped; rohelisega märgitud need aminohapped mis osalevad nii RNA-valk kui ka valk-valk interaktsioonides:

- Hüdrofiilse keskkonna loomiseks:
    - Asn33, Gln34, Thr37, **Ser123**, Asn127 (Namba *et al*., 1989).
- RNA-valk interaktsioonid:
    - Stabiilsuse loomine:
        - **Arg41**, **Arg 90**, Arg92 (Namba *et al*., 1989; Tewary *et al*., 2011; Wang ja Stubbs, 1994; Li *et al*., 2013; Subbs, 1999; Pattanayek ja Stubbs, 1992).
    - Nukleiinhappe aluse äratundmise spetsiifilisuse loomine:
        - **Asp115** (Namba *et al*., 1989; Holmes, 1979).
    - Arg82, Gly85, Ala86, **Asp88**, Thr89, Asn91, **Glu97**, **Glu106**, **Asp109**, **Arg112**, **Arg113**, Val114, Asp116, Ala117, Val119, Ala120, Ile121, **Arg122**, **Ser123** (Namba *et al*., 1989; Holmes, 1979; Wang ja Stubbs, 1994; Stubbs, 1999; Graham ja Butler, 1979; Bhyravbhatla *et al*., 1998; Pattanayek ja Stubbs, 1992)
- Valk-valk interaktsioonid.
    - Virioni struktuuri kokkupanemiseks ning lahtiharutamiseks:
        - Glu50-Asp77/Arg134 (Pattanayek ja Stubbs, 1992; Stubbs, 1999; Tewary*et al*., 2011; Li *et al*., 2013; Wang ja Stubbs, 1994).
        - **Glu106**-Glu95/**Glu97**/**Asp109** (Pattanayek ja Stubbs, 1992; Bhyravbhatla *et al*., 1998; Stubbs, 1999; Tewary *et al*., 2011).
    - Glu95-**Arg112** (Wang ja Stubbs, 1994)
    - Ser15-Asn25 (Li *et al*., 2013).
    - Thr28-Tyr72 (Li *et al*., 2013).
    - **Asp88**-Phe35/Gln36/**Arg112**/**Arg122** (Namba *et al*., 1989; Wang ja Stubbs, 1994; Li *et al*., 2013).
    - Arg134-Asp19/Asp57/Asp66 (Wang ja Stubbs, 1994; Li *et al*., 2013).
    - Glu22-Lys53 (Li *et al*., 2013).
    - **Arg113**-**Arg90**/**Arg115** (Namba *et al*., 1989; Wang ja Stubbs, 1994; Bhyravbhatla *et al*., 1998).
    - **Asp115**- Gln36 (Bhyravbhatla *et al*., 1998).
    - **Asp41**-**Arg122** (Holmes, 1979; Wang ja Stubbs, 1994).

- Pro45, Thr59, Ala74, Val75, Ser147, Ser148 (Holmes, 1979; Namba *et al*., 1989).

LISA 13. Tabel töös kasutatud järjestuste kohta. Välja on toodud NCBI andmebaasi järgi (Sayers *et al.*, 2009) taksonoomia ID, järjestuse ID ning nimi. Lisaks on märgitud järjestusele, mida kasutati fülogeneeside koostamisel ning mida eelasjärjestuste kosntrueerimisel.

Taksonoomia ID NCBI	**Järjestuse ID NCBI**	**Nimi/organism**	**Fülogenees**	**Joonise lühendid**	**Eelasjärjestus**
		1CGM.pdbchainEs001	+	1CGM	
		1EI7.pdb chainA s002	+	1EI7	
		1RMV.pdbchainAs003	+	1RMV	
		1VTM.pdbchainPs004	+	1VTM	
		3PDM.pdbchainPs006	+	3PDM	
7213	AOHK01015310.1	Ceratitis capitata	+	C_cap	
7213	NW_004523913.1	Ceratitis capitata			+
7213	FG088519.1	Ceratitis_capitata			+
7291	ACVV01094903.1	Drosophila albomicans			+
7217	NW_001939297.1	Drosophila ananassae			+
125945	AFFD02001848.1	Drosophila biarmipes			+
42026	AFFE02006047.1	Drosophila bipectinata			+
30023	AFFF02008357.1	Drosophila elegans			+

7220	NW_001956551.1	Drosophila erecta			+
29029	AFPQ02004970.1	Drosophila eugracilis			+
30025	AFFG02007800.1	Drosophila ficusphila			+
7222	AAPT01019095.1	Drosophila grimshawi	+	D_gri	
7222	NW_001961677.1	Drosophila grimshawi			+
30033	AFFH02007845.1	Drosophila kikkawai			+
7227	NC_004354.4	Drosophila melanogaster			+
7229	GALP01000229.1	Drosophila miranda			+
7230	NW_001979118.1	Drosophila mojavensis			+
7234	NW_001985964.1	Drosophila persimilis			+
46245	AADE01000673.1	Drosophila pseudoobscura pseudoobscura	+	D_pse	
46245	NW_001589959.2	Drosophila pseudoobscura pseudoobscura			+
1041015	AFPP02029731.1	Drosophila rhopaloa			+
7238	NW_001999693.1	Drosophila sechellia			+
7274	GAHN01006963.1	Drosophila serrata			+
7240	AAST01006817.1	Drosophila simulans	+	D_sim	+
7240	AASR01044027.1	Drosophila simulans			+

7240	NC_011089.1	Drosophila simulans			+
28584	CAKG01011790.1	Drosophila suzukii			+
29030	AFFI02008573.1	Drosophila takahashii			+
7260	NW_002032860.1	Drosophila willistoni			+
7244	AANI01015530.1	Drosophila virilis	+	D_vir	
7244	NW_002014440.1	Drosophila virilis			+
7245	NC_011091.1	Drosophila yakuba			+
7395	JMRR01004949.1	Glossina austeni			+
7395	JFJS01010860.1	Glossina brevipalpis	+	G_bre	+
201502	JFJR01003850.1	Glossina fuscipes fuscipes	+	G_fus	+
37546	DV613204.1	Glossina morsitans morsitans	+	G_mor_1	+
37546	BX551233.1	Glossina morsitans morsitans	+	G_mor_2	+
37546	CCAG010017880.1	Glossina morsitans morsitans	+	G_mor_3	+
37546	FM982036.1	Glossina morsitans morsitans			+
37546	DV617147.1	Glossina morsitans morsitans			+
37546	EZ422866.1	Glossina morsitans morsitans			+
7398	JMRQ01000691.1	Glossina pallidipes			+

67801	JXJN01029566.1	Glossina palpalis gambiensis			+
1323540	GAZD01011456.1	Lipara lucens	+	L_luc	+
7375	JHUJ01027746.1	Lucilia cuprina	+	L_cup	+
7370	AQPM01050138.1	Musca domestica	+	M_dom	
7370	XM_005176796.1	Musca domestica			+
139649	GBBP01051127.1	Teleopsis dalmanni	+	T_dal	+
292399	GBGG01009716.1	Themira biloba	+	T_bil	+
12466	NP_604490.1	Barley mild mosaic virus	+	Bymo_1	
12327	NP_604486.1	Barley stripe mosaic virus	+	Hordei_1	+
12465	NP_149000.1	Barley yellow mosaic virus	+	Bymo_3	
368735	YP_001333653.1	Bell pepper mottle virus	+	Toba_1	+
402399	YP_001974326.1	Brugmansia mild mottle virus	+	Toba_10	+
229030	YP_002455907.1	Cactus mild mottle virus	+	Toba_18	+
1128119	YP_004956730.1	Clitoria yellow mottle virus	+	Toba_20	+
146499	NP_072164.1	Cucumber fruit mottle mosaic virus	+	Toba_24	+
12235	NP_044580.1	Cucumber green mottle mosaic virus			+
388038	YP_908763.1	Cucumber mottle virus	+	Toba_21	+

99585	YP_003915156.1	Frangipani mosaic virus	+	Toba_15	+
1523023	YP_009047253.1	Gentian ovary ring-spot virus	+	Hor_Pec_1	
233051	NC_025381.1	Hibiscus latent Fort pierce virus			+
185955	YP_720000.1	Hibiscus latent Singapore virus			+
32629	NP_835263.1	Indian peanut clump virus	+	Peclu_2	+
111970	NP_619687.1	Kyuri green mottle mosaic virus	+	Toba_22	+
368736	YP_950424.1	Maracuja mosaic virus	+	Toba_16	+
31749	NP_620844.1	Obuda pepper virus	+	Toba_8	+
12238	NP_056812.1	Odontoglossum ringspot virus	+	Toba_6	+
35281	NP_671721.1	Paprika mild mottle virus	+	Toba_7	
1032457	YP_004465361.1	Passion fruit mosaic virus	+	Toba_17	+
12294	NP_040351.1	Pea early-browning virus	+	Tobra_2	+
28355	NP_620028.1	Peanut clump virus	+	Peclu_1	+
12239	NP_619743.1	Pepper mild mottle virus	+	Toba_5	+
31750	NP_620037.1	Pepper ringspot virus	+	Tobra_3	+
1123754	YP_004936169.1	Rattail cactus necrosis-associated virus	+	Toba_19	+
425279	YP_001041892.1	Rehmannia mosaic virus	+	Toba_2	+

51680	YP_005476603.1	Ribgrass mosaic virus			+
335187	YP_762620.1	Streptocarpus flower break virus	+	Toba_14	+
111418	NP_624339.1	Zucchini green mottle mosaic virus	+	Toba_23	+
12241	NP_062916.1	Tobacco mild green mosaic virus			+
12242	NP_597750.1	Tobacco mosaic virus			+
12295	NP_620682.1	Tobacco rattle virus	+	Tobra_1	+
12253	NP_078449.1	Tomato mosaic virus	+	Toba_4	+
1391702	YP_008492931.1	Tomato mottle mosaic virus	+	Toba_3	+
29272	NP_046154.1	Turnip veinclearing virus	+	Toba_11	+
1169032	NP_543052.1	Wasabi mottle virus	+	Toba_13	+
75746	NP_059448.1	Wheat yellow mosaic virus	+	Bymo_2	+
1416026	YP_008802587.1	Yellow tailflower mild mottle virus	+	Toba_9	+
228578	NP_740759.1	Youcai mosaic virus	+	Toba_12	+

LIHTLITSENT

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Heleri Kirsip (sünnikuupäev: 5. aprill 1991)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Viirustelt kärbselistele ülekandunud geeni, *TMV-CP*, integratsiooniaja ning funktsiooni uurimine“, mille juhendajad on Aare Abroi

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

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu alates **1.06.2017** kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

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

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

Tartus, 25.mai 2015