TARTU ÜLIKOOL

LOODUS- JA TEHNOLOOGIATEADUSKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

Heleri Kirsip

Kas taimeviirused võivad olla uute valgudomeenide allikaks hulkraksetele loomadele?

Bakalaureusetöö

Juhendaja vanemateadur Aare Abroi, PhD

Kaasjuhendaja doktorant Tõnu Margus, MSc

TARTU 2013

SISUKORD

KASUTATUD LÜHENDID

BLAST – üldine lokaalse joonduse leidmise tööriist (*Basic Local Alignment Search Tool*)

CAV – mändvetika viirus (*Chara australis* viirus)

cf – SCOP-i klassifitseerimise pakkimise ehk voltumise tase (*common fold*)

cl – SCOP-i klassifitseerimise klassi tase (*class*)

CMV – kurgi mosaiikviirus (*Cucumber mosaic virus*)

EBLN – endogeenne bornaviiruste sarnane N element (*endogenous bornavirus-like N element*)

EDI – EVE-de abil loodud immuunsus (*EVE-derived immunity*)

EVE – endogeenne viiruse element (*endogenous viral element*)

fa – SCOP-i klassifitseerimise perekonna tase (*family*)

fg – femtogramm

fM – femtomolaarne

GRD – geminiviirustega seotud DNA (*geminivirus-related DNA*)

GTA – geenide ülekandja (*gene transfer agent*)

H2V – ülekanne peremehelt viirusele (*host-to-virus transfer*)

HCRSV – *Hibiscus chlorotic ringspot* viirus

HMM – peidetud Markovi mudel *(Hidden Markov model*)

HT-NGS – kõrge läbilaskega teise põlvkonna sekveneerimine (*High-throughput next generation sequencing*)

LCMV – Lümfotsütaarne koriomeningiitviirus (*Lympholytic Choriomeningitis virus*)

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

MAT – miljonit aastat tagasi (*million years ago, MYA*)

MSA – mitme järjestuse joondus (*Multiple Sequence Alignment*)

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

RDA – võrdlev geenieksрessioonianalüüs (*representational difference analysis*)

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

sf – SCOP-i klassifitseerimise superperekonna tase (*superfamily*)

SSH – supressioonil põhinev subtraktiivne hübridisatsioon (*supression subtactive hybridization*)

TMV-CP – tubaka mosaiigi viiruse sarnased kattevalgud

UP – UniProt andmebaas

V2H – ülekanne viiruselt peremehele *(virus-to-host transfer*)

VLP – viiruslaadne partikkel (*virus-like particle*)

SISSEJUHATUS

Viirused on väga suure arvukusega laialt levinud bioloogilised objektid. Viirused ei jäta üldjuhul endast fossiile ning nende kiire mutatsioonikiirus ei võimalda viiruse evolutsioneerumise kohta adekvaatseid hinnanguid teha. Pikka aega on teada retroviiruste järjestuste esinemisest eukarüootsete organismide genoomides. Uudseks on aga mitte-retroviiruslike elementide esinemine. Viiruslik järjestus, integreerudes ja kinnistudes peremeesorganismi genoomi, allub peremeesorganismi aeglasemale mutatsioonikiirusele ning seega peaks olema järjestuselt sarnasem integreerunud viirusele.

Selle bakalaureusetöö teoreetiline osa annab lühiülevaate integreerunud viiruslikest järjestustest ning nende avastamise võimalustest. Töös keskendutakse peamiselt mitte-retroviiruslikele järjestustele. Lisaks tutvustatakse valkude klassifitseerimise meetodit SCOP, mis on antud töös uuritavatele järjestustele aluseks.

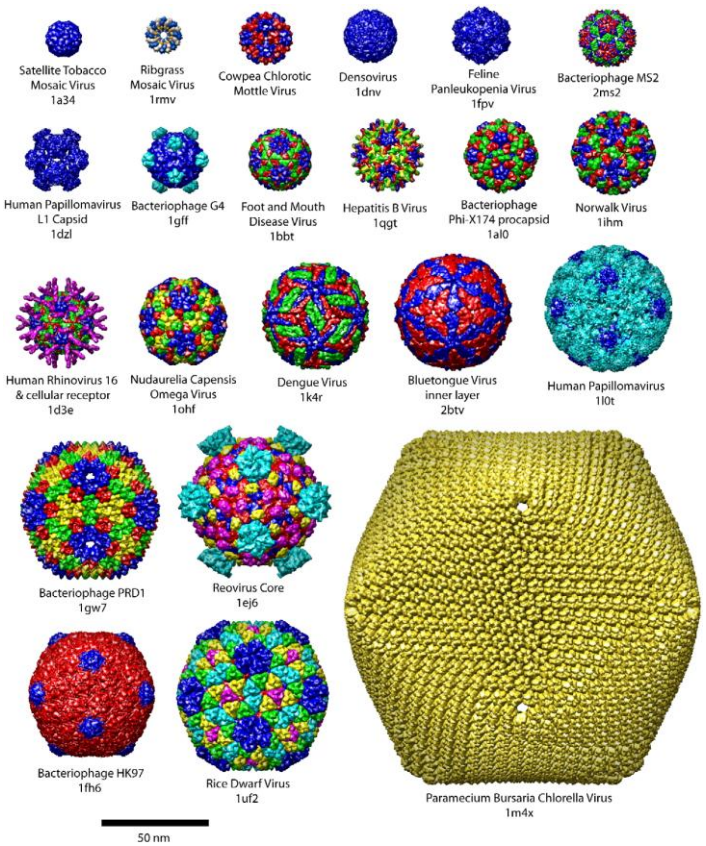
Töö eksperimentaalses osas tõendatakse, et kärbsed ja taimeviirused jagavad domeeni. Uuritakse toimunud ülekande suunda. See aitab aru saada viiruste, vektorite ja peremeesorganismide kompleksetest suhetest.

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

1. KIRJANDUSE ÜLEVAADE

1.1. Viirused erinevates biotoopides ja keskkondades

Viiruste hulk Maa erinevates elukeskkondades on võrreldes organismidega tunduvalt kõrgem. Näiteks inimese 10^{13} raku kohta esineb inimkehas 10x rohkem baktereid ning 100x rohkem viiruseid (Mokili *et al*., 2012). Viirused on oma mõõtudelt väikesed (~100 nm, sisaldades 0,2 fg süsinikku), moodustades ookeanis biomassilt suure grupi, jäädes alla vaid prokarüootide biomassile (Suttle, 2005). Suurimaks viiruseks peetakse *Megavirus chilensis*-t, kapsiidi diameetriga 520 nm (koos fiibritega on partikli diameeter 680 nm) ja genoomiga 1 259 197 aluspaari (Arslan *et al*., 2011). Väikseimaks viiruseks arvatakse olevat *Porcine circovirus* tüüp 1, partikliga 16-18 nm ja genoomiga 1 759 nukleotiidi (Finsterbusch ja Mankertz, 2009). Enamik kirjeldatud viiruste genoome jäävad siiski alla 200 000 bp.



Joonis 1. Viiruste suurused. Joonisel on kujutatud erinevate viiruste suurused ning nende all on toodud identifitseerivad nimetused, kasutades andmebaasi PDB (*Protein Databank*, http://www.rcsb.org/pdb/home/home.do). Joonis on võetud Goddard *et al*. (2005) artiklist „Software Extensions to UCSF Chimera for Interactive Visualization of Large Molecular Assemblies". Töös uuritav viiruse kapsiid on välja toodud esimeses reas vasakult teine – Ribgrass Mosaic Virus (1rmv).

Arvatakse et viiruseid esineb ookeanis 10^{29} (Wilhelm and Suttle, 1999) kuni 10^{30} viirust (Breitbart, 2012). Kui joondada kõik ookeanis leiduvad viirused, arvestades ühe viiruse diameetriks 50 nm, oleks viirustekett 400 000 valgusaastat pikk. Võrdluseks – Linnutee on vaid 25 000 valgusaastat pikk. (Weinbauer ja Rassoulzadegan, 2004)

Bergh *et al.* (1989) on leidnud, et ühes liitris merevees leidub ca 10^8 viiruse-sarnast partiklit (VLP), mis on umbes kümme korda rohkem kui prokarüootide kontsentratsioon samas keskkonnas. Wommack ja Colwell (2000) on leidnud, et merevee rannikuäärsetes pinnakihtides leidub ca 10^7 VLP-d milliliitri vee kohta. See teeb viiruste üldiseks kontsentratsiooniks ookeanides 17 femtomolaari (fM), samal ajal kui inimeste (6,9 miljardit) kontsentratsioon ookeanis oleks vaid 10^{-20} fM (Breitbart, 2012).

Pinnases on VLP kontsentratsioon ca 10^8 VLP-d ühe grammi pinnase kohta (Williamson *et al.*, 2003). Kolme erineva pinnatüübi (elamurajooni, metsa ja tööstuskompleksi) kohal atmosfääris mõõdetud viiruste hulgaks saadi $1{,}7x10^6$ kuni $4{,}0x10^7$ viirust m^{-3} (Whon *et al.*, 2012). Keskmine inimene hingab 24 tunni jooksul sisse ja välja ca 11 000 liitrit õhku, mis teeb 11 m^3 õhku ööpäevas ("How much oxygen does a person consume in a day?" 01.aprill 2000, http://health.howstuffworks.com/human-body/systems/respiratory/question98.htm, 18. mai 2013).

$$11m^3 * 4.4 * 10^7 \text{ viirust } m^{-3} = 4.84 * 10^8 \text{ viirust ööpäevas}$$

Seega 24 tunni jooksul peaks teoreetiliselt inimesest läbi käima ligikaudu $4{,}4*10^8$ viirust. Ja see on vaid üks kokkupuute viis viirustega. Inimese kokkupuude viirustega igal ajahetkel on kordades suurem. See on hea näide viiruste massiivsest arvukusest.

Tänaseks on täielikult sekveneeritud 3 252 viiruseliigi genoomi (arvestamata isolaate ja subtüüpe) (http://www.ebi.ac.uk/genomes/virus.html, 27.05.2013 seisuga). Viiruslike metagenoomide uurimistest on leitud, et 60-99% viiruste metagenoomsetest järjestustest ei oma homoloogsust ühegi teadaoleva järjestusega (Breitbart *et al.*, 2002), seega on viiruste diversiteet palju suurem, kui arvata võiks. Viirustel ei esine kindlat universaalset konserveerunud valku, valgu domeeni või muud järjestust, mis esineks kõikidel viirustel. Siiski on mõned järjestused, mis on iseloomulikud teatud viiruste gruppidele – nendest lähtutakse viiruste sugukondadesse ja kõrgematesse taksonitesse grupeerimisel (Suttle, 2005). Tähendusväärne on fakt, et on leitud geograafiliselt väga erinevatest piirkondadest (Lõunameri, Mehhiko laht ja Arktika sulavee järv) viiruste järjestusi, mis on peaaegu identsed nukleotiidsel tasemel (Short ja Suttle, 2005). Kui kohalik viiruste mitmekesisus

on väga suur, võib selle taustal globaalne diversiteet tunduda suhteliselt madal (Breitbart ja Rohwer, 2005). Ehk ühest keskkonnast võetud metagenoomi uuring võib näidata, kui palju erinevaid viiruseid võib tegelikult eksisteerida. Samas aga uurides üle maailma erinevaid keskkondi, avastame sarnaseid viirusjärjestusi kõikjalt, mis viitab sellele, et leitakse samu identifitseerimata ja identifitseeritud viiruseid üle maailma.

Viiruste kontsentratsioon keskkonnas sõltub tugevalt erinevatest teguritest: temperatuur, viiruse-bakteri suhe keskkonnas, organismi lüüsumise määr (Weinbauer *et al*., 1995).

Wommack ja Colwell (2000) leidsid seose aastaaegade vaheldumise ja viiruse arvukuse, viiruse-bakteri suhte ning lüsogeensuse hulga vahel. Üldiselt on viiruse arvukus kõrgem suviti ja sügiseti ning madalam talviti. Näiteks määrati talvel Norra rannikuvetes 10^4 viiruse partiklit ml^{-1} vees, kevadest sügiseni suurenes arvukus 100 korda.

Viiruste hulk on üldjuhul tugevas sõltuvuses bakterite arvuga keskkonnas – 3-10 VLP-d ühe bakteriraku kohta (Wommack ja Colwell, 2000). Bakteriofaagid lüüsivad keskmiselt 4-50% uutest paljunenud bakteritest (Heldal ja Bratbak, 1991), ülejäänud on protistidele toiduks (Fuhrman ja Noble, 1995).

Suttle (2007) hindas, et iga päev toimub ookeanides 10^{23} viiruseinfektsiooni, mille tagajärjel vabaneb keskkonda kuni 10^9 tonni süsinikku bakterite rakkudest. Tänu sellele on viirused olulised nii toitainete-, kui ka energiaringes. Viirused stimuleerivad organismide kasvu, vabastades bioloogilistest rakkudest orgaanilist substraati, mida kasutavad teised organismid kasvamiseks ja paljunemiseks (Middelboe ja Lyck, 2002). Lotka-Volterra mudel „tapa võitja“ (*kill-the-winner*) selgitab viiruste osalust bakteriliikide vohamises. Kui üks bakteriliik saavutab dominantsuse keskkonnas (nišis), siis bakteriliigile omane viirus lüüsib selle grupi, võimaldades teistel, nõrgematel, bakteriliikidel saavutada ülekaal (Thingstad ja Lignell, 1997).

Viirused mõjutavad lisaks ka peremeesorganismide kohasust, aidates bakteritel saavutada nõnda ülekaal teiste üle. Brüssow ja Hendrix (2002) leidsid, et bakterigenoom sisaldab keskmiselt 3-10% profaagi DNA-d. Lawrence *et al*. (2002) leidsid, et üks bakteri genoom sisaldab keskmiselt 2,6 profaagi genoomijärjestust ehk ~50% sekveneeritud bakteriliikidest esines vähemalt üks ennustatav profaag. Profaagide identifitseerimisel bakteriaalsetest genoomidest võib aga tekkida probleeme – nimelt on profaagide genoomne diversiteet niivõrd suur, et üle 50% ORF-idest ei leia andmebaasidest vasteid (Fouts, 2006). Viirused

kannavad osa oma geneetilisest materjalist transduktsiooni kaudu peremehele, mis võib viimastele anda eelise teiste suhtes (Anderson *et al*., 2011).

Peremehe kohasust ei mõjutata ainult geeniülekande kaudu. Paljudel organismidel on avastatud sümbiootiline suhe neid nakatavate viirustega (Roossinck, 2011).

Peedil, kurgil, paprikal, arbuusil ja tomatitaimel on avastatud parem vastupidavus põuale, kui taimedel esineb infektsioon kurgi mosaiikviirusega (CMV, *Cucumber mosaic virus*). Olenevalt taimest esinesid veepuudusele iseloomulikud tunnused 2-5 päeva hiljem, kui viiruse infektsioonita kontroll-taimedel. Põhjuseks võib olla osmoprotektantide (suhkrud, proliin, antioksüdandid) kuhjumine taimes, mis aitab lühiajalist veepuudust üle elada. (Xu *et al*., 2008)

Avastatud on viiruste tähtis osakaal erinevate organismide evolutsioonilisel kujunemisel. Näiteks pakkus Harris (1991) välja hüpoteesi trofoplasti tekkele: arenevad embrüod nakatusid emakas retroviirustega, tekitades rakkude proliferatsiooni ning trofoplasti teket, mis on oluline embrüo implantatsiooniks emakasse. See protsess pidi toimuma enne platsentaalsete imetajate lahknemist ning lisaks pidi integratsioon toimuma idurakkudes, et uus omadus päranduks järglastele. Teooria tõestuseks on süntsüütsiotrofoblastile sarnased morfoloogilised tunnused retroviiruse infektsioonil rakukultuurides – hulktuumse hiiglasliku raku teke.

Seega on viirused elukeskkonna ja peremeesorganismi varieeruvuse tagamisel oluliseks komponendiks.

1.2. Viiruste identifitseerimiseks kasutatavad metoodikad

Viiruse kultiveerimine, seroloogia, elektronmikroskoopia ja PCR on traditsioonilisemad meetodid, mida kasutatakse uute viiruste identifitseerimiseks (Tang ja Chiu, 2010), kuid neil igaühel esinevad omad piirangud viiruste süstemaatiliseks avastamiseks. Näiteks elektronmikroskoopia abil saab viiruseid eristada vaid morfoloogiliste erinevuste alusel (Roingeard, 2008). Meetodeid on võimalik kasutada kombineeritult koos, kuid protsess on kallis, ebaefektiivne ning aeganõudev (Dong *et al*., 2008).

Kliinilises ja keskkonnamikrobioloogias kasutatakse viiruste avastamiseks peamiselt kahte üldistatud meetodit: järjestusest sõltuvad ja järjestusest mittesõltuvad meetodid.

1.2.1. Järjestusest sõltuvad meetodid

Nende alla kuuluvad PCR, kasutades konsensuspraimereid, ja hübridisatsioonimeetodid – mikrokiibiga analüüsimine. Meetodite eelduseks on uute viiruste nukleiinhapete järjestuste teadmine. Meetodites kasutatakse teadaolevate viiruste konserveerunud järjestusi, et identifitseerida uute viiruste genoome. Tuvastatavad viirused peaksid olema suguluses teadaolevate viirustega. (Mokili *et al*., 2012)

1.2.2. Järjestusest mittesõltuvad meetodid

Järjestusest mittesõltuvad tehnoloogiad ei vaja eelteadmisi proovis leiduvate viiruste kohta. Neid kasutatakse viiruste jaoks, mida leidub proovis väheses koguses. Protsess on töörohke ja suhteliselt keeruline ning esineb madal tundlikkus (välja arvatud HT-NGS). (Tang ja Chiu, 2010)

Sellisteks meetoditeks on:

- Superssioonil põhinev subtraktiivne hübridisatsioon (SSH) (Ambrose ja Clewley, 2006)

SSH-d kasutatakse üldiselt erinevate ekspressioonimustrite abil geenide vaheliste erinevuste leidmiseks (Stein ja Liang, 2002). Protsess põhineb pööratud kordusjärjestustel (ITR). Esmalt toimub mittevajaliku DNA eemaldamine, kuni jääb alles soovitud üheahelaline DNA, seejärel seda kloonitakse ja sekveneeritakse (Tang ja Chiu, 2010).

- Võrdlev geeniekspressioonianalüüs (RDA) (Stein ja Liang, 2002)

RDA on sarnane protsess SSH-ga, kuid ahelate eraldamine on ühendatud PCR amplifikatsiooni protsessiga, saades suure hulga soovitud puhastatud DNA järjestusi. Protsessi kasutatakse genoomi ja cDNA proovides erinevuste leidmiseks. (Delwart, 2007)

1.2.2.1. Viiruslik metagenoomika

Järjestustest mittesõltuvatest meetoditest on viiruste avastamises levinud kombineeritud meetod – viiruslik metagenoomika, mis on vähem kallutatud kui eelnevalt mainitud meetodid. Teoreetiliselt on võimalik avastada kõiki viiruseid: kultiveeritavaid ja

mittekultiveeritavaid, tuntud, kui ka võõraid viiruseid (Mokili *et al*., 2012). Koosneb kolmest etapist:

1. Proovi ettevalmistus (Mokili *et al*., 2012)

Võimalik on analüüsida kõikjalt saadud proove: merevesi, veri, väljaheited, meresetted ja kuumaveeallikad (Mokili *et al*., 2012). Esimese protseduurina mitte-viiruslike elementide eemaldamisel proov homogeniseeritakse, filtreeritakse ja ultratsentrifuugitakse (Thurber *et al*., 2009). Võõr-DNA eemaldatakse kloroformi ja DNaasiga töötlemisel, võõr-RNA RNaasiga töötlemisel (Lee ja Hallam, 2009). Kloroformiga töötlemisel tuleb olla ettevaatlik, kuna see võib põhjustada membraaniga viirustel lipiidkihi kaotuse, seega muuta viiruse DNA DNaasile ligipääsetavaks (Breitbart ja Rohwer, 2005).

2. Kõrge läbilaskvusega sekveneerimine (HT-NGS, *High-throughput next generation sequencing*, (Tang ja Chiu, 2010))

Sangeri ensümaatilist sekveneerimise meetodit on kasutatud peamise sekveneerimise meetodina alates 1977. aastast, millal see loodi (Sanger *et al*., 1977). Teise generatsiooni sekveneerimise meetodid pakkusid kiirust ja automatiseeritust, tõstes sekveneerimise efektiivsust (Mokili *et al*., 2012).

HT-NGS kolmanda põlvkonna sekveneerimistehnoloogia on arendamisjärgus. Selle eesmärgiks on ühe-molekuli sekveneerimine ehk proovitakse saavutada sekveneerimine ilma järjestuse amplifikatsioonita, mis vähendab tulemuste kallutatust (Schadt *et al*., 2010).

3. Bioinformaatiline analüüs

Metagenoomse meetodi kõige vaevanõudvam etapp on andmete analüüs. Tuleb kindlaks määrata proovis esinevad järjestused ning need assambleerida kontiigideks (Schatz *et al*., 2010). Järgnevalt otsitakse sekveneeritud järjestustele homolooge erinevatest andmebaasidest, kasutades otsingutööriistu (Mokili *et al*., 2012), näiteks BLAST (*Basic Local Alignment Search Tool*) perekonna programme (Altschul *et al*., 1990). Kasutatavad andmebaasid peaksid sisaldama ka peremeesorganismide ja bakteriaalseid järjestusi, et eemaldada taustmüra – võõr-DNA-d, mida ei suudetud eemaldada (Schmieder ja Edwards, 2011).

1.2.3. Klassikalise viroloogia probleemid „high-throughput“ ajastul

Arvestades viiruste hulka maailmas, tuleb tõdeda, et kõik viirused ei ole patogeensed, vaid võivad esineda ka organismi normaalse floora osana, nagu on näidatud heade viiruste näidetega (vaata lk 10). See on põhjuseks mille pärast peab enne ravi kinnitama parasiidi olemasolu kui haiguse tekitajat (Mokili *et al*., 2012).

Läbi ajaloo on haigustekitaja identifitseerimisel juhindutud Koch-i postulaatidest (Falkow, 2004):

1. Haigust tekitav parasiit või patogeen peab esinema igas vastavat haigust põdevas organismis.
2. Parasiit ei tohi esineda teistes haigustes juhuslikult ega kui mitte-patogeenne parasiit.
3. Pärast haigest organismist isoleerimist ja puhaskultuuris kasvatamist, peab haigustekitaja olema võimeline uueks eelnevaga identseks infektsiooniks.

Juba Koch-i eluajal tõdeti, et kõik haigustekitajad ei täida nõutud kriteeriume. Näiteks Koch-i kolmas postulaat nõuab, et patogeeni peab olema võimalik pärast eraldamist uuesti kasvatada puhaskultuuris ning järgnevalt peab olema võimalik indutseerida uus esialgse infektsiooniga identne infektsioon (Rivers, 1937). Kahjuks pole võimalik kõiki viiruseid kultuurides kasvatada (Mokili *et al*., 2012). Seega peaks esialgseid nõudeid kohandama viiruste jaoks. Seda tegi Falkow (2004), kohandades Koch-i postulaate, et need kehtiksid molekulaarsete meetodite kasutamisel: tuleks näidata tugevat seost haiguse fenotüübi ja põhjustava patogeeni perekonna või liigi esindajate vahel geeni tasemel. Ehk haigust tekitav geen peaks esinema kõigil patogeensetel tüvedel, kuid peaks puuduma mittepatogeensetel tüvedel.

Neid juhtnööre kasutades peaks olema võimalik identifitseerida ühe või mitme viirusliku elemendi põhjustatud haiguse tekitaja.

Viimasel ajal on aga avastatud bakteriofaagidele sarnaste elementide esinemine keskkonnas – GTA-d (*gene transfer agent*). GTA-d sarnanevad bakteriofaagi morfoloogilisele välimusele (pea ja saba struktuur) ja suurusele (pea diameeter 30-45 nm; saba pikkus kuni 190 nm), sisaldades juppi organismi genoomist, mis on väiksem kui partikli enda kodeerimiseks vajalik järjestus ehk GTA ei sisalda (seniste teadmiste kohaselt) täielikku komplekti kapsiidi kodeerivaid geene (Lang *et al*., 2012). Seega tekib küsimus: kuidas eristada viiruseid ja GTA-sid, kui nad on oma olemuselt üksteisele nii

sarnased? See on tõsiseks probleemiks viiruste identifitseerimisel ja keskkonna-viroloogial.

Viirusteks peetakse üldiselt nukleiinhapet sisaldavaid bioloogilisi objekte, kes suudavad end taastoota. GTA-d aga pole selleks võimelised. Kasutades viiruste identifitseerimiseks NGS või elektronmikroskoopiat, pole võimalik teada saada, kas avastatud objekt on viiruslikku päritolu või hoopis näiteks GTA. Seega tuleks viiruste identifitseerimisel rakendada modifitseeritud Koch-i postulaate, mis seda võimaldavad.

Mokili *et al*. (2012) töötasid välja metagenoomsed Koch-i postulaadid, mis põhinevad metagenoomsete omaduste ehk haiguste molekulaarsete markerite identifitseerimisel haigusega organismides:

1. Haigusega metagenoom peaks erinema oluliselt kontroll-metagenoomist.
2. Tervet organismi inokuleerides haigustekitajaga peab järgnema haiguslik seisund.
3. Eelnevalt selektiivselt inokuleeritud materjaliga nakatades uut tervet organismi, peab kaasnema samade sümptomitega haigus.

Uued avastused hajutavad ajalooliselt kindlaks määratud selgeid piire. Seega tuleks defineeritud mõisted üle vaadata ning kohandada tänapäevaste teadmistele sobivateks, nagu seda on tehtud näiteks Koch-i postulaatidega.

1.3. EVE-d

EVE-d ehk endogeensed viiruse elemendid (*endogenous viral elements*) on viiruse geenid või genoom, mis on integreerunud organismi genoomi ning fikseerunud põlvkondade jooksul (Bejarno *et al*., 1996). Kõige tihedamini integreeruvad retroviirused pöördtranskriptaasi abil (Lander *et al*., 2001), kuid on leitud tõendeid ka teiste replikatsioonitüüpidega viiruste integratsioonist peremehe genoomi. Integratsioon toimub kas mitte-homoloogse rekombinatsiooni (Katzourakis ja Gifford, 2010) või peremeesorganismi retroelementide abil (Geuking *et al*., 2009).

On teada, et eukarüootide genoomid sisaldavad erineval hulgal retroviiruslikke järjestusi sellepärast, et retroviiruste elutsükli üheks osaks on peremeesgenoomi integreerumine. Üllatuseks on mitte-retroviiruslike genoomijärjestuste leidmine, kuna enamik replitseeruvad tsütoplasmas ning neil puuduvad ensüümid dsDNA sünteesiks. (Crochu *et al.*, 2004)

1.3.1. EVE-de avastamine

Esimene kinnitatud tõend RNA viiruse genoomi DNA järjestuse leidmisest elusorganismist presenteeris Klenerman *et al.* (1997), tõestades LCMV (*Lympholytic Choriomeningitis virus*) DNA järjestuste olemasolu hiirte maksarakkudes 200 päeva pärast infektsiooni. Võimalust, et viiruse genoomi DNA järjestus on integreerunud peremeheorganismi genoomi, ei uuritud.

Tänaseks on EVE-d leitud paljudest erinevatest organismidest:

- loomade genoomidest (Holmes, 2011)
- mitmetest seente genoomidest (Taylor ja Bruenn, 2009)
- bakterite genoomidest (Salanoubat *et al*., 2002)

Bejarno *et al.* tõestasid esimest korda viiruse DNA integratsioonist peremeesraku genoomi taimedes. Tubakataime (*Nicotiana tabacum*) genoomist avastati geminiviiruste-sarnane DNA (GRD ehk *geminivirus-related* DNA). GRD-d leiti ka teistest alltoodud tubakataime liikidest:

- *Nicotiana tomentosa*
- *Nicotiana tomentosiformis*

Geminiviiruse DNA järjestus peab olema integreerunud meristeemkoe genoomi tubakataimede evolutsiooni algetappidel, kui tekkisid taime erinevad liigid. Arvatavasti kinnistas selle neutraalne geneetiline triiv, lähedal asuva geeni ekspressiooni muutmine või pakkus järjestuse integratsioon geminiviiruste vastu resistentsust. (Bejarno *et al*., 1996)

Kindla tõendi viiruse geneetilise materjali sisestusest seente genoomi esitasid Taylor ja Bruenn (2009), kes leidsid integreerunud totiviiruse sarnaseid järjestusi kolmest seeneliigist:

- *Candida parapsilosis*
- *Penicillium marneffei*
- *Uromyces appendiculatus*

Fülogeneetilised tõendid viitavad, et geneetiline materjal kandus totiviirustelt seentele. Leiti ka tõendeid horisontaalsest geeniülekandest seente vahel. (Taylor ja Bruenn, 2009)

Bornaviiruste sarnaseid N elemente (EBLN) on leitud mitmetest selgroogsete liikide esindajate genoomidest: primaadid, närilised, kiskjad, nahkhiired, putuksööjad,

kukkurloomad ja kalad. EBLN-e on leitud ka inimese genoomist **(**Horie ja Tomonaga, 2011**).** Horie *et al*. (2010) leidsid, et osad primaatide EBLN-id sisaldavad ka tervet ORF-i ja neid ekspresseeritakse mRNA-dena.

Crochu *et al*. (2004) avastasid 2/3 erinevate flaviviiruste genoomsetest järjestusest sääskedest *Aedes albopictus* ja *Aedes aegypti*. Uuriti nii laboratoorseid kui ka looduses leiduvaid sääski, kindlustamaks laboritingimustes tekkinud artefakti puudumist. Samuti detekteeriti flaviviiruste järjestuse mRNA-d organismis, viidates järjestuse olulisusele peremehele. Tähendusväärne on fakt, et moskiitod on vastavalt Aasia ja Aafrika päritolu ning nende lahknemine (34-42 MAT) on toimunud kordades varem kui uuringus käsitletud flaviviiruste lahknemine (3 500 – 350 000 aastat tagasi), lisaks olid järjestused piisavalt erinevad, et leiud võiksid viidata kahele iseseisvalt toimunud integratsioonile. See aga annab tõestust, et viiruse järjestuste integratsioon peremeesorganismi ei ole nii erandlik sündmus kui algselt arvati.

1.3.2. EVE-de bioloogiline tähtsus

Üldjuhul on EVE-d passiivsed reliikviad, akumuleerides raaminihkeid ja stopp-koodoneid, mida põhjustab geneetilise triivi jõud (Holmes, 2011).

Kuid esineb ka seda, et organism saab integreerunud viirusest kasu – EDI-d (*EVE-derived immunity*). EDI-d käituvad kui viiruse infektsiooni inhibiitorid: blokeerivad viiruste rakku sisenemise, inhibeerivad replikatsiooni või omandavad peremeesorganismi genoomis olles hoopis uusi funktsioone. Seega on EVE-d üheks viiruste ja nende peremeesorganismide võidurelvastumise osaks. (Aswad ja Katzourakis, 2012)

Paleoviroloogia kasutab EVE-dest saadud informatsiooni viiruste evolutsiooni määramiseks (Aswad ja Katzourakis, 2012; **Joonis 2**, vaata lk 18) – otsitakse ortoloogseid EVE-sid fülogeneetiliselt mitte väga lähedastelt peremeesliikide genoomidest. Seejärel vaadatakse organismide evolutsioonilist põlvnemist ning proovitakse dateerida insertsiooni ligikaudne aeg. See võib näidata, et viiruse sugukond on miljoneid aastaid vanem, kui on saadud eksogeensete sugulaste fülogeneetiliste analüüside tulemusena (Holmes, 2011).

Näiteks *Circoviridae* perekond ei ole mitte <500 aastat vana (Firth *et al*., 2009), vaid EVE-de dateerimisega on saadud >40 miljonit aastat (Belyi *et al*., 2010).

Teiseks näiteks on filoviirused, arvatava vanusega 10 000 aastat (Suzuki ja Gojobori, 1997). Taylor *et al*. (2010) näitasid, uurides erinevatest organismidest leitud filoviiruste järjestusi (nahkhiired, närilised, kukurloomad), et filoviirused on kümneid miljoneid aastaid vanemad, kui algselt arvati.

1.3.3. EVE-de tuvastamise meetodid

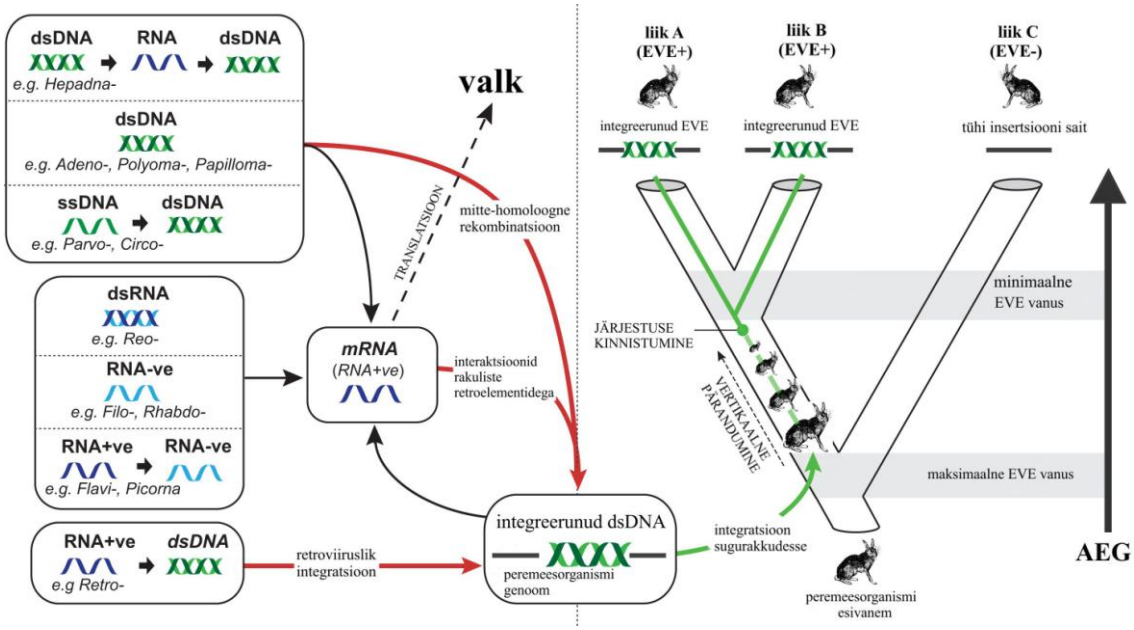
Sekveneerides organismi genoomi, tuleks kindlasti saadud järjestused kontrollida:

1. annoteerimise eesmärgil,
2. saastuse mittetekkimise tõestamise eesmärgil.

Sekveneeritud genoomi järjestusi tuleks vastandada erinevate andmebaaside vastu (Katzourakis ja Gifford, 2010). Selleks võib kasutada BLAST-i perekonna programme (Altschul *et al*., 1990).

Teiseks võimaluseks on süsteemne EVE-de skriining organismide genoomidest. Selleks teostatakse viiruse järjestuse homoloogide otsinguid erinevate andmebaaside vastu (Horie *et al*., 2010). Katzourakis ja Gifford (2010) lõid algoritmi *in silico* endogeensete mitte-retroviiruslike insertsioonide tuvastamiseks genoomidest. Nad lõid <100 Kb genoomiga viirustest valgujärjestuste andmebaasi, mida hakati võrdlema erinevate organismide sekveneeritud genoomijärjestustega. Leides järjestusi, mis sobisid nende seatud kriteeriumitega (E-väärtus <0.001), joondati organismi ja viiruse järjestused ning konstrueeriti fülogeneetilised puud. Üldiselt kasutatakse bootstrap-i väärtust puu siseste lahknemiste usaldusväärtuse hindamiseks (Horie *et al*., 2010).

Hinnates EVE-de insertsiooni toimumise ligikaudset aega, on võimalik järeldada ka viiruse vanust (**Joonis 2**). Selleks tuleks uurida viiruslike järjestuste olemasolu/puudumist EVE insertsiooniga organismi lähedastelt liikidelt. EVE lookuse esinemine lähedaste liikide samas lookuses viitab integratsioonile enne nende liikide divergeerumist. Sellega määratakse EVE integratsiooni minimaalne aeg. Tühja insertsioonisaidiga organism, kes on viirusliku järjestusega organismile kõige lähemalt suguluses, määrab ligikaudse maksimaalse EVE vanuse. (Feschotte ja Gilbert, 2012)



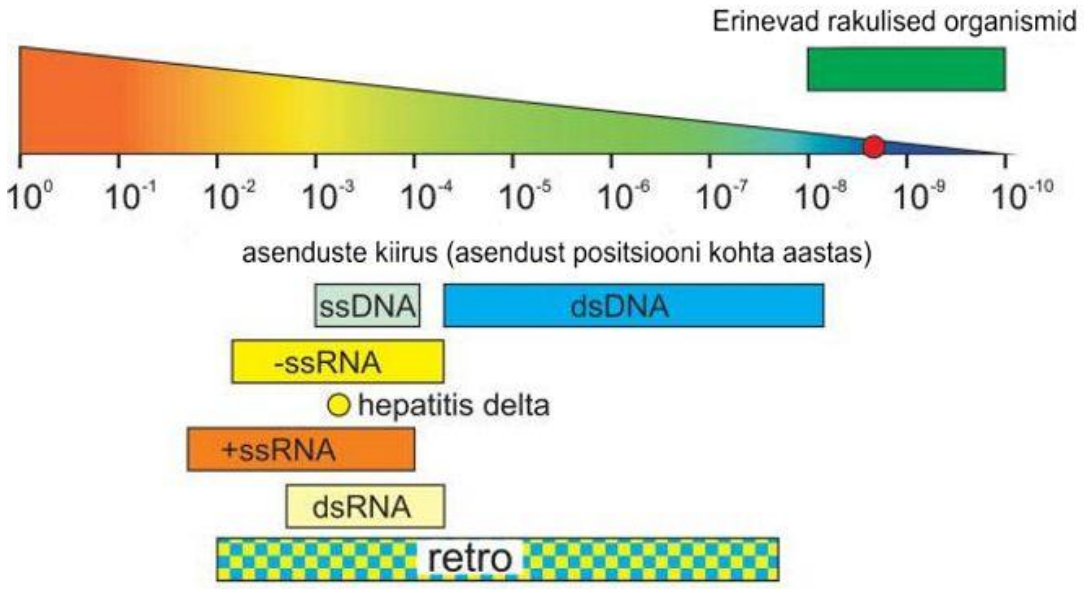
Joonis 2. EVE-de integratsioon ning selle abil viiruste vanuse määramise üldskeem. Joonise originaali koostasid Katzourakis ja Gifford (2010); eesti keelde tõlgiti töö autori poolt.

EVE insertsioonid lähedaste liikide erinevates lookustes või esinemine lähedastes perekondades, kuid mitte igas liigi esindajas, võib viidata kahele iseseisvale insertsioonile (Cui ja Holmes, 2012; Crochu *et al.*, 2004). Sel juhul tuleks uurida tõendeid, mis viitaksid kahele iseseisvale insertsioonile. Näiteks *Aedes* liikide erinev geograafiline päritolu (vaata lk 16 (Crochu *et al.*, 2004)). Samas EVE mitteleidmine lähedaselt liigilt võib viidata ka sellele, et mingil põhjusel on EVE selles organismis kaduma läinud ning viiruslik järjestus on tegelikult veelgi vanem.

Organismide vanuse määramisel eri meetoditega tekkivad suuremõõtmelised erinevused tulenevad sellest, et molekulaarse kella ja mutatsioonide tekke kiirus piiravad järjestuse vanuse määramist. Näiteks mutatsioonikiirusega 0,14% miljoni aasta kohta on võimalik dateerida aega kuni 90 miljonit aastat (Belyi *et al.*, 2010). Samas aga arvestades peremeesorganismiga ko-evolutsioneerumist, on võimalik objekti vanuse määramisel minna ajas tunduvalt kaugemale.

RNA viirustel on väga kõrge mutatsioonide tekke kiirus – 0.1 kuni 1.0 mutatsiooni genoomi replikatsiooni kohta (Duffy *et al.*, 2008). Üldiselt on RNA viirustel nukleotiidide asenduskiirus 10^{-2} – 10^{-7} asendust positsiooni kohta aastas (Hanada *et al.*, 2004; Jenkins *et al.*, 2002), samas kui eukarüootidel on see 10^{-9} asendust positsiooni kohta aastas (**Joonis 3**). See tuleneb viirustel tõenäoliselt RNA-sõltuva RNA polümeraasi (RdRp) abil

replitseerumisest, mis on suure vigade määraga. Samas ei tohiks unustada, et enamus mutatsioone on kahjulikud ning need kaotatakse puhastava valikuga (Holmes, 2009). Lisaks muteeruvad genoomi erinevad järjestused erinevate kiirustega (Holmes, 2003). See suurendab järjestuste stabiilsust ning peremeesorganismiga ko-evolutsioneerumine aitab vähendada järjestuse muutlikkust (Vandamme *et al*., 2000).



Joonis 3. Viiruste ja eukarüootsete organismide keskmine mutatsioonikiirus aastas toimuvate positsiooni asenduste kohta. Punase täpiga on märgitud imetajate tuuma kodeerivate järjestuste keskmine asenduste kiirus. Joonis on võetud Abroi ja Gough (2011) artiklist „Are viruses a source of new protein folds for organisms? – Virosphere structure and space evolution". Joonise tõlkis eesti keelde töö autor.

Fülogeneetilised puud konstrueeritakse järjestuste joondumise alusel. Challis ja Schmidler (2012) näitasid globiinide baasil, et programmid, mis arvestavad nii valgujärjestusi, kui ka struktuuri, annavad tõepärasemaid tulemusi, kui need programmid, mis arvestavad ainult valgujärjestusi (Katoh *et al*., 2002). Järjestuse baasil loodud fülogenees ei klasterdanud kõiki selgroogseid ühte rühma – valgu tertsiaarstruktuur on rohkem konserveerunud kui järjestus sellepärast, et selektiivne jõud mõjub valgule funktsioneerimise tasemel. MAFFT programm (mitme järjestuse joondus) andis esimese konstrueeritud puu puhul tõepärasele väga ligilähedase puu, kuid eemaldades lähedased homoloogsed järjestused, jättes alles vaid kauged homoloogid, langes MAFFT-i konstrueeritud tõepärased tulemused, andes 10-st variandist vaid ühe tegelikkusele lähedase topoloogia. Järjestuse-struktuuri liitprogramm andis aga 6 tõest topoloogiat. Seega fülogeneetilist ajalugu rekonstrueerivad programmid,

mis põhinevad mitmejärjestuse joonduse informatsioonile, ei võimalda kaugesse minevikku ulatuvad (fülogeneetilist) ajalugu usaldusväärselt rekonstrueerida, sest need positsioonid, mis toetaksid sügavaid harusid on ajas küllastunud. Seega tuleks valida tunnused, mis muutuvad ajad aeglasemalt, näiteks järjestuse ruumiline struktuur.

1.4. SCOP – valkude struktuurse klassifikatsiooni andmebaas

Üldiselt koosnevad kõik valgud ühest või mitmest domeenist. Domeen on kolmedimensionaalse struktuuri ühik (Murzin *et al*., 1995; Orengo *et al*., 1997), samas peetakse seda ka valkude evolutsiooniühikuks (Riley ja Labedan, 1997). Neid domeene saab valkude moodustamiseks kombineerida erinevatel viisidel (Gordana *et al*., 2001).

Enamus valke omavad struktuurset sarnasust teiste valkudega ja seega võivad omada ka ühist evolutsioonilist päritolu. Bioloogiliste makromolekulide struktuuri andmebaasis PDB-s (*Protein Databank*) on kirjeldatud väga palju erinevaid valke. Valkude struktuuride evolutsioneerumise mõistmiseks loodi andmebaas, mis klassifitseerib valke kolme-dimensionaalse struktuuri sarnasuse ja evolutsioonilise põlvnemise baasil – SCOP-i andmebaas. Klassifikatsiooni ühikuks peetakse ühte valgu domeeni. (Murzin *et al*., 1995)

Hierarhiliselt jagatakse valgud neljal tasemel:

- Klass (cl; *class*)
- Pakkimise ehk voltumise tase (cf; *common fold*)
- Superperekond (sf; *superfamily*)
- Perekond (fa; *family*)

Vastavalt valgustruktuurile jagatakse valgud kaheksaks klassiks (17.04.2013, http://scop.berkeley.edu/):

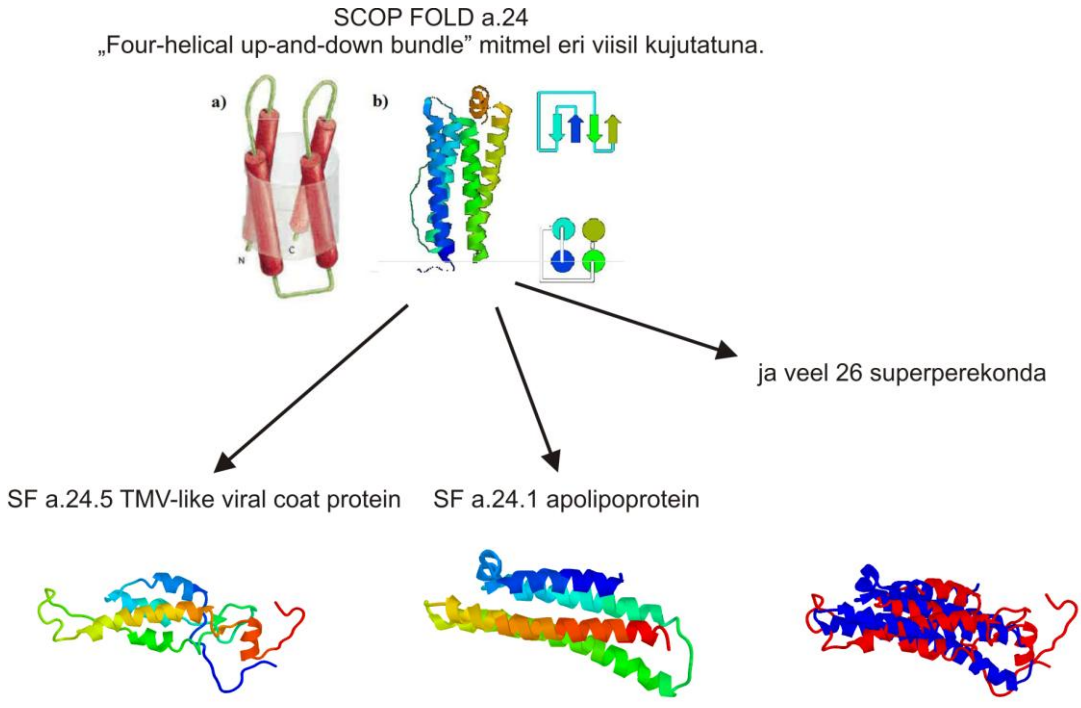
1. alfa-heeliksitest koosnevad valgud (*all alpha proteins*)
2. beeta-lehtedest koosnevad valgud (*all beta proteins*)
3. alfa- ja beetastruktuurid intervallidena jaotunud valgud (*alpha and beta proteins*) (α/β)
4. alfa pluss beetastruktuurid segregeerunult valgud (*alpha and beta proteins*) (α+β)
5. multidomeensed valgustruktuurid (*multi-domain proteins*)

6. membraani- ja raku pinnavalgud (*membrane and cell surface proteins and peptides*)
7. väikesed valgud (*small proteins*)
8. keerdunud valgud (*coiled coil proteins*)

Voltumise tasemel (*common fold*) klassifitseeritakse ühte gruppi need valgud, mis omavad ühist voltumisviisi, omades samu peamisi sekundaarstruktuuri elemente samade topoloogiliste ühendustega (**Joonis 4**). See tuleneb tõenäoliselt sarnasest valgu pakkimise viisist (Murzin *et al*., 1995).

Valkude perekonnad, kelle struktuur on sarnane, kuid järjestuste identsus on väike, viitavad ühisele evolutsioonilisele päritavusele ning need klassifitseeritakse ühte superperekonda (Murzin *et al*., 1995). Samasse sf-i klassifitseerimine eeldab valkude südamiku sarnast pakkimisviisi.

Valgud, mille järjestus, struktuur ja funktsioon on väga sarnased, viitavad ühisele evolutsioonilisele päritolule ning klasterdatakse ühte perekonda (Murzin *et al*., 1995).



Joonis 4. Näide valkude jaotumisest SCOP-i andmebaasis. Alpha heeliksitest koosnevad valgud, mis on üles-alla kipudeks pakitud, jagunevad 28 superperekonnaks. Esile on toodud TMV-sarnaste viiruste kattevalkude (TMV-CP) superperekond valgu domeeni d1cgme (vasakul) näitel ning apolipoproteiinide superperekond d1le2a (keskel) näitel. Parempoolseim valkude joonis näitab d1cgme ja d1le2a struktuurse joonduse kattuvusi ning erinevusi, mille alusel on valgudomeenid erinevatesse superperekondadesse jaotatud. Joonise koostas Aare Abroi.

1.5. SUPFAM andmebaas

SUPFAM (SUPERFAMILY andmebaas) on valkude ja genoomide struktuurse ja funktsionaalse annotatsiooni andmebaas, mille ülesandeks on avastada ning klassifitseerida valgujärjestusi, millel on teada struktuuri esindaja. See protsess toimub SCOP-i superperekonna (sf) tasemel ning toimub peidetud Markovi mudelite (HMM) baasil. Analüüsis kasutatakse täielikult sekveneeritud genoome. (Gough, 2002)

HMM-id on üks tundlikumaid järjestuste võrdlemise meetodeid (Park *et al*., 1998). SUPFAM-is määrati genoomide järjestustele SCOP-i domeenid, kasutades HMM-e, ning selle baasil loodi HMM-ide raamatukogu, millel baseerub SUPFAM-i andmebaas (Gordana *et al*., 2001). Iga loodud HMM mudel vastandatakse kõikidele olemasolevatele määratud järjestustele, kontrollimaks mudelite kvaliteeti. Kui mudelisse sobitub teise superperekonna valgustruktuur, siis uuritakse mudelit, identifitseerides ja lahendades tekkinud probleemid (Gough, 2002). Seega halbu mudeleid ei eemaldata, vaid parandatakse ning mida rohkem on valkude struktuure teada, seda vähem esineb valepositiivseid.

Andmebaasi kasutatakse genoomi annotatsioonide, struktuurse genoomika, geeni ennustamise ja domeeni-põhisteks genoomi uuringuteks (Gough ja Chothia, 2002).

Sekveneeritud genoomide järjestusi saab võrrelda SUPFAM-i mudelite raamatukogu järjestustega ning niimoodi annoteerida järjestusi, leides otsitavatele kaugeid homolooge tuntud valkude seast (Gough *et al*., 2001).

1.6. Kirjanduse kokkuvõte

Viiruseid esineb maailmas tohutus koguses. Igal ajahetkel puutub inimene kokku sadade, kui mitte rohkemate viirustega. Sellise tiheda kokkupuutega mõjutab nii viirus peremeesorganismi kui ka vastupidi.

Viimastel aastatel on järjest rohkem avastatud ja kirjeldatud viiruslike järjestuste avastamist eukarüootsetest organismidest. Leitud ei ole mitte ainult retroviiruslikke elemente, vaid ka positiivse ja negatiivse polaarsusega RNA, dsRNA ning ssDNA viiruseid. Seni leitud järjestused on tõendite baasil määratud viiruslikeks, ülekande suunaga viiruselt peremehele (V2H).

Integratsiooni toimunud aja määramisel on kasutatud valkude järjestust. Antud töös proovitakse määrata toimunut kasutades valgu järjestuste joondusele lisaks valkude struktuure, mis peaks parandama fülogeneetiliste puude usaldusväärsust ning sellekaudu andma täpsemaid tulemusi. Lisaks peaks kombineeritud meetod võimaldama seletada kaugemas ajalises sügavuses toimunut, kui seda võimaldavad ainult valkude järjestuste joondustele ülesehitatud meetodid.

2. EKSPERIMENTAALNE OSA

2.1. TÖÖ EESMÄRGID

Antud bakalaureusetöö eesmärgiks on uurida viiruse ja eukarüoodi vahelist geneetilise materjali ülekannet:

- Kinnitada ülekande toimumine
- Leida argumente ülekande suuna määramiseks
- Uurida, kas on toimunud üks ülekanne, mis on levinud evolutsioneerumise käigus või on toimunud mitmed sõltumatud ülekanded
- Insertsiooni toimumise aja hindamine

Töö kaugemaks eesmärgiks on kogu protsessi automatiseerimine, luues programm, mis suudab viiruste ja organismide valkude järjestuse ja struktuuri abil hinnata geneetilise materjali ülekannet.

2.2. MATERJALID JA METOODIKA

Töö metoodika on toodud skeemina (**Joonis 5**, vaata lk 26).

SUPFAM. Uuringus kasutati SUPFAM (versioon 1.75) andmebaasi. Uuritav objekt (SUPFAM-i kood 47195, SCOP ID a.24.5) ostutus valituks lähtudes kriteeriumitest:

- Andmete üldine maht on analüüsitaval hulgal
- Viiruslikke järjestusi on leitud vähemalt 5 eukarüootsest organismist ja enam kui 10-st viirusest

Lisaks kasutati SUPFAM-i integreeritud NCBI viiruslikke (vl) järjestusi (versioon 2013-2) ja UNIPROT-i (UP) järjestusi (UP väljalase 2013-3). UP andmebaas on suhteliselt kõdunud. Igast liigist on esindatud mitmed tüved. NCBI andmebaas on kureeritud, mis tõstab kvaliteeti, kuna iga lisatud järjestus läbib kontroll-etapi. Lisaks on NCBI andmebaasis iga liik esindatud ühekordselt – isolaadid ja erinevad tüved on välja jäetud.

ANNOTSIATSIOONI KONTROLL. Järjestuste annotatsiooni kontroll teostati kasutades SUPFAM-is igale järjestusele välja toodud linki vastava assambleeritud ja annoteeritud genoomi kodulehele. Genoomide resekveneerimiste või assableerimiste ajal võib avastada

eelnevalt tekkinud viiruslikku saastet. Kontiigi pikkus ei tohiks olla RNA viiruse suurusjärgus ning külgnevad alad ei tohiks sarnaned viiruslike järjestustega.

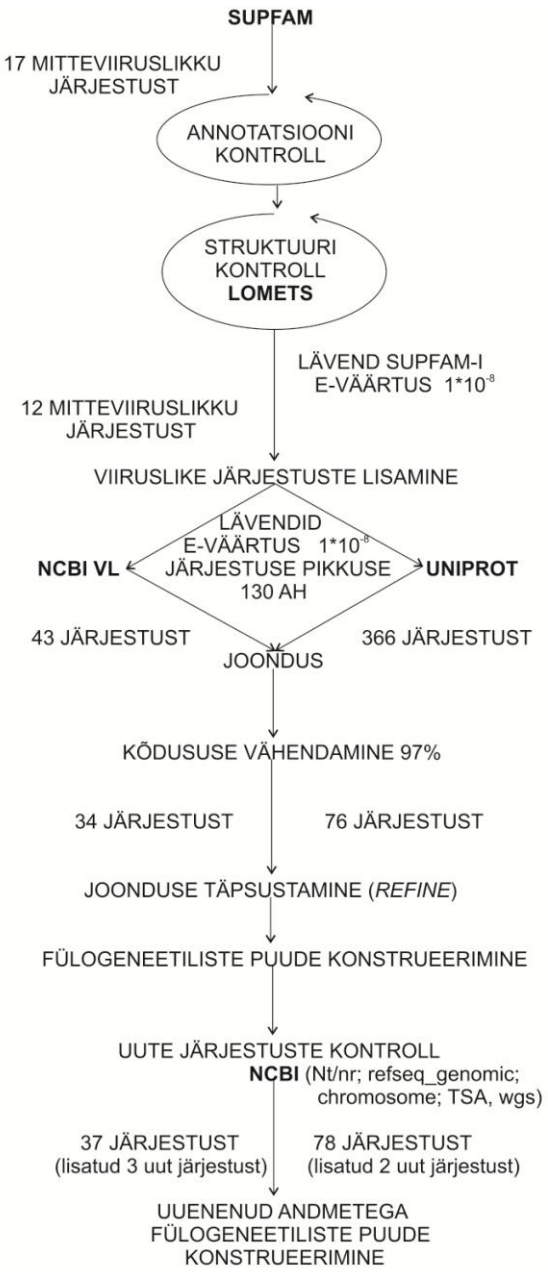
LOMETS. SUPFAM-i ennustuse õigsuse hindamiseks teostati kontroll antud järjestuste valkude struktuuride ennustamise programmipaketiga LOMETS (versioon 2.0, viimane uuendus 4.04.2010). LOMETS genereerib kolmedimensionaalseid valkude struktuuri ennustusi, kasutades metaservereid. Programmis on kombineeritud 9 erinevat algoritmide *threading* programmid: HHsearch, MUSTER, PPA, PROSPECT2, SAM-T02, SP3, FFAS ja PRC (Wu ja Zhang, 2007).

MSA. Järjestustele määrati lävendiks SUPFAM-i E-väärtus $1*10^{-8}$ ja järjestuse regiooni pikkus 130 aminohapet. Järjestused joondati programmiga Muscle (Edgar, 2004), kasutades programmipaketti Jalview (versioon 2.8, viimane uuendus 12.11.2012; Waterhouse *et al*., 2009). Järjestustest eemaldati kõduvus, määraga 97%, et vähendada fülogeneetiliste puude konstrueerimisel suuremahuliste andmete tekitatud müra. Joondatud järjestus täpsustati (*-refine*) tühimike (*gap*) piirkonnas kasutades Muscle-t (versioon 3.8.31).

FÜLOGENEETILISTE PUUDE KONSTRUEERIMINE. Fülogeneetilised puud konstrueeriti kasutades programmipaketti MEGA (versioon 5.1; Tamura *et al*., 2011). Puud konstrueeriti distantsimeetodil (NEIGHBOR) parameetritega: boostrap meetod 500, aminohappeline järjestus, Jones-Taylor-Thorton-i mudelit kasutades, tühikute paariviisilise kustutamisega.

BLAST. Järjestuste uuenenud otsing teostati kasutades NCBI koduleheküljel asuvat BLAST perekonna programmi (BLAST+ 2.2.28; uuendatud 2013-3), tblastn, erinevate andmebaaside vastu. Kasutatud andmebaasid:

- *Nucleotide collection* (Nt/nr)
- *Reference genomic sequences* (refseq_genomic)
- *NCBI chromosome sequences* (chromosome)
- *Transcriptome Shotgun Assembly sequences* (TSA)
- *Whole-genome shotgun contigs* (wgs)



Joonis 5. Töö metoodika üldskeem.

2.3. TULEMUSED

2.3.1. TMV-CP potensiaalsed struktuursed homoloogid rakkudes

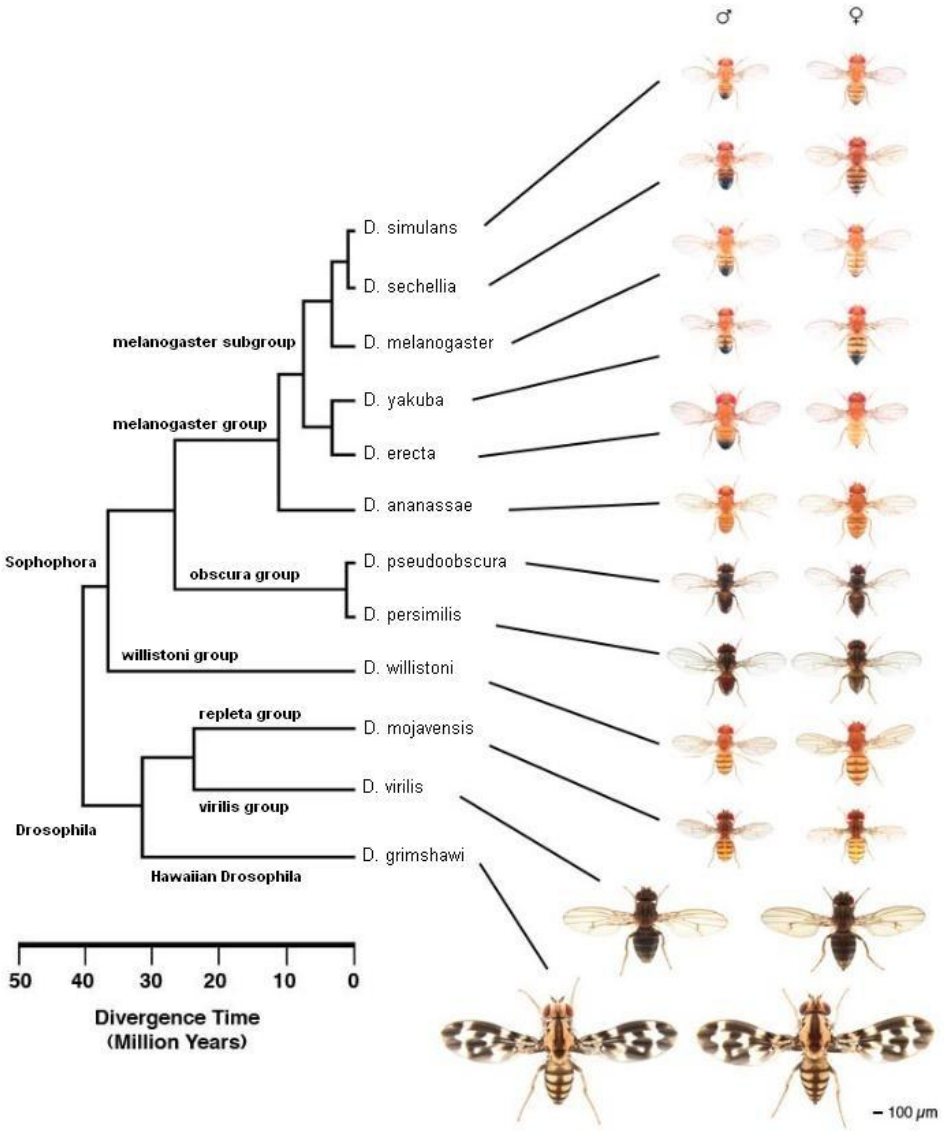
SUPFAM-is leiti TMV-CP superperekonna HMM-dele sobivad vasted 18 valgus, mis pärinesid 17 organismist:

12 *Drosophila* liiki (**Joonis 6**, vaata lk 28):

- *Drosophila grimshawi*
- *Drosophila willistoni*
- *Drosophila pseudoobscura*
- *Drosophila persimilis*
- *Drosophila yakuba*
- *Drosophila simulans*
- *Drosophila sechellia*
- *Drosophila erecta*
- *Drosophila melanogaster* (andmebaasidest Ensemble ja FlyBase)
- *Drosophila ananassae*
- *Drosophila virilis*
- *Drosophila mojavensis*

Nematood *Pristinochus pacificus*, kaks seeneliiki *Mucor circinelloides* ja *Phycomyces blakesleeanus* ning kõrstaim *Panicum virgatum*.

SUPFAM-is esines äädikakärbse (*D. melanogaster*) kolm variant, mis olid lisatud erinevatest andmebaasidest (kaks Ensemble-st ja üks Flybase-ist). Järjestused olid identsed ning arvestasin neid kui ühte järjestust.



Joonis 6. Töös kasutavate *Drosophila*-de evolutsiooniline lahknemine. Joonis on võetud FlyBase-i koduleheküljelt, mai 2013.

2.3.2. Järjestuste annotatsiooni kontroll

Enne uuringu alustamist peaks kontrollima, kas uuritavad järjestused on endiselt andmebaasis ning ega sekveneeritav materjal ei ole olnud saastunud viirusega. Antud järjestused esinesid kõikide organismide genoomide kõige uuemates versioonides. Kontiigid ja kromosoomid jäid üldiselt suuruselt 1.7 – 22.4 miljoni aluspaari vahemikku (**Lisa 3)**. Teise põlvkonna sekveneerimisel peaks viiruslik järjestus jääma eraldi kontiigina ning need ei tohiks olla assambleeritud suurteks kontiigideks. Eranditeks olid *P.*

blakesleaanus-e kontiig 40, mis oli 370 815 nukleotiidi pikk ja *P. virgatum*-i kontiig 50 355 pikkusega 5 693 nukleotiidi.

Järjestust ümbritsevad alad ei sarnanenud viirusliku päritoluga. Järjestuse ümbritsev ala on *Drosophila* erinevatel liikidel kõrge ortoloogsusega, erinedes üksikute ümberkorraldustega geenide järjestuses. Lisaks esinesid uuritavad järjestused kõigis organismides (välja arvatud *P. pacificus* (3 eksonit)) ühe eksonina. Pikkade kontiigide olemasolu tõstab, et tegemist on tõepoolest genoomse järjestusega.

2.3.3. SUPFAM-i ennustuse kontroll teiste meetoditega

SUPFAM-is juhindutakse struktuuri mustrite äratundmisel HMM-st. Mudelite täpsuse ja kvaliteedi määramiseks võib valkude struktuure ennustada alternatiivsete meetoditega, mis arvestavad teisi parameetreid.

Kärbeste ja teiste uuritavate organismide valguregioonide struktuuride ennustamine andis kõrge usaldatavuse skooriga valgustruktuurid 1vtm_P, 1cgm_E, 1rmv_A ja 1ei7_A, mis vastavad SCOP-i andmebaasi (1.75B, uuendatud jaanuar 2013) järgi TMV-sarnaste viiruste kattevalkude gruppi (a.24.5.1; näited **Lisa 6**). Samad neli valgustruktuuri on SUPFAM-is TMV-CP neljale HMM mudelile aluseks. Lisaks andis LOMETS usaldusväärseks vasteks valgustruktuuri 3pdm, mida ei ole SCOP-i andmebaasis praeguseks hetkeks (mai, 2013) klassifitseeritud, kuid PDB andmebaas annab struktuurse sarnasuse valguga 1ei7. 3pdm struktuur on pärit *Hibiscus latent Singapore* viirusest, mis kuulub tobamoviiruste hulka ning on suure tõenäosusega TMV-CP superperekonda kuuluv valk.

Seega võib usaldada SUPFAM-i TMV-CP HMM mudeleid ning ennustatavad valgud kuuluvad õigesse valkude superperekonda.

2.3.4. Fülogeneetilisteks uuringuteks vajaliku andmevalimi koostamine

NCBI vl andmebaas andis 31 viiruslikku järjestust: *Virgaviridae* sugukonna tobraviiruste, pekluviiruste, hordeiviiruste ja tobamoviiruste perekonnad ning *Potyviridae* sugukonnalt bümoviiruste perekond. Kõik järjestused ületasid SUPFAM-i lävendi E-väärtuse. Samuti ei olnud ükski järjestus lühem kui 131 aminohapet. Seega fülogeneetilise puu konstrueerimiseks ei kaotatud selle andmbeaasi järjestustest mitte ühtegi.

Pärast kõdususe eemaldamist jäid NCBI andmeid kasutades alles 34 järjestust (31 viiruslikku ja 3 *Drosophila* järjestust). Joondusest eemaldati 9 *Drosophila* järjestust liigse sarnasuse tõttu kõdususe eemaldamise etapis. Viiruslikud järjestused olid piisaval määral erinevad ning kõik jäid alles.

UNIPROT-i andmebaasist määras SUPFAM TMV-CP sarnaseks 468 järjestust. Pärast piirangute rakendamist (SUPFAM-i E-väärtuse lävend ja minimaalne järjestuse pikkus 131 aminohapet) jäi alles 354 järjestust, millest üks oli kärbseline *Glossina morsitans* (**Joonis 7**). Viiruslikud järjestused kuulusid kahte klassifitseerimata viiruste sugukonda: *Virgaviridae* perekonna esindajad tobraviirused, pekluviirused, hordeiviirused ja tobamoviirused ning *Potyviridae* perekonnast bümoviirused.

Joonis 7. *Glossina morsitans centralis*. Täiskasvanud isane laboratooriumi kolooniast. Pildi autor Steven Mihok, http://www.nzitrap.com/Biting/biting.htm, pilt alla laaditud 20.05.2013.

Eraldi tooks välja UP andmebaasis olevad kaks viiruslikku järjestust, millel esinesid probleemid:

- *Hibiscus chlorotic ringspot* viirus (Q6QDD3, HCRSV) on ametlikult klassifitseeritud kui karmoviirus (*carmovirus*). Igal konstrueeritud fülogeneetilisel puul grupeerus aga tobamoviiruste hulka. Viiruse klassifitseerimise kontrollimiseks teostasin NCBI BLAST (tblastn) otsingu, kasutades võrdluseks Huang *et al*. (2000) artiklis toodud HCRSV sekveneeritud järjestust (X86448). Otsing ei andnud usaldusväärset vastet. Lisaks teostasin üldise tblastn otsingu kasutades UNIPROT-i andmebaasist saadud HCRSV järjestust. Usaldusväärsete vastete hulgas esinesid

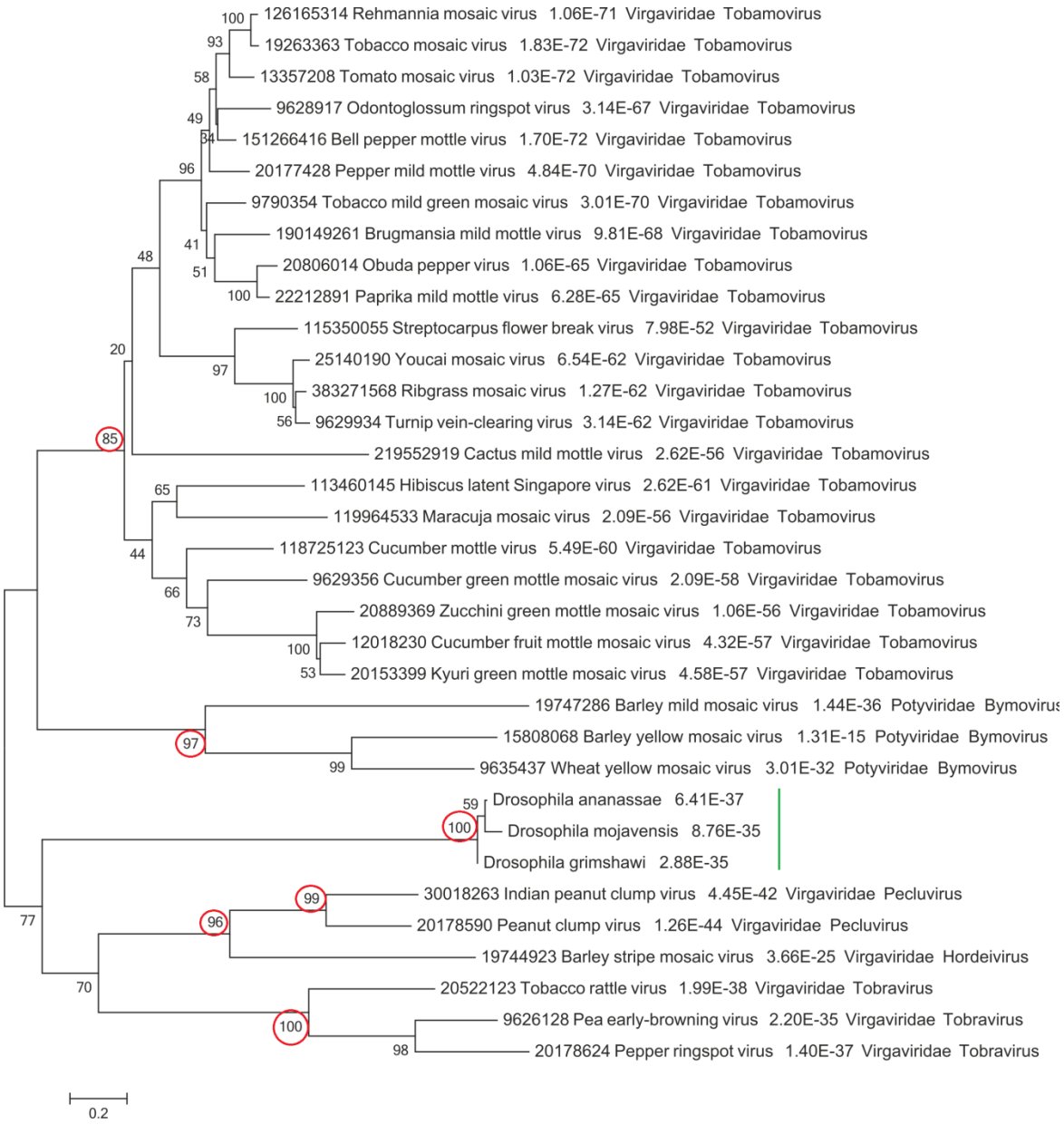
tobamoviirustest *Hibiscus latent Fort Pierce* ja *Hibiscus latent Singapore* viirused ning lisaks *Cucumber green mottle mosaic* viiruse erinevad tüved. Samas esines vastete hulgas ka *Hibiscus chlorotic ringspot* viirus (AY546633, AY546635 (andmed esitasid NCBI andmebaasi Chen TC, Liu FL ja Chen YK, 21.03.2004)). Järelikult võib arvata, kas viirus on identifitseerimisel valesti klassifitseeritud kui karmoviirus või viiruse sekveneerimiseks võetud proovis esines ka tobamoviiruse infektsioon ning tobamoviirusele omane kattevalgu järjestust assambleeriti karmoviiruse HCRSV genoomi.

- Vetika *Chara australis* viirus (Ca_F8ULT8; CAV; vana nimega *Chara corallina virus*) on klassifitseerimata viirus, kes sarnaneb struktuurilt tobamoviirustele. Siiski erineb viirus suurel määral, teostamaks lisauuringuid suguluse määramiseks. Esiteks prooviti vetikaviirusega nakatada teisi tubaka mosaiikviiruse (TMV) peremeestaimi, kuid infektsiooni ei toimunud. Samuti uuriti antiseerumi mõju viirusele. Sellega avastati nõrk sugulus TMV orhidee tüvedele, kuid mitte teistele TMV tüvedele (Gibbs *et al*., 1975). Genoomi sekveneeritud osa (genoomi ei õnnestunud täielikult sekveneerida) uurimine andis tõestust, et *Chara australis* viiruse ORF-ide produktid on suguluses nii tobamoviirustega, kui ka benyviirustega. See aga näitab, et CAV valkude ja lähemate sugulaste fülogeneetiline sugulus on vanem kui kumbki sugulusgrupp ise (Gibbs *et al*., 2011).

Kõdususega eemaldati UNIPROT-i valimit kasutades joondusest 290 järjestust: 281 viiruslikku järjestust ning 9 *Drosophila* liigi esindajat. Alles jäi 76 viiruslikku järjestust, 3 *Drosophila* ning üks *G. morsitans*-i järjestus.

2.3.5. Fülogeneetiliste puude konstrueerimine kasutades NCBI vl valimit

Kasutades NCBI vl andmebaasi valimit, konstrueeriti esimene fülogeneetiline puu (**Joonis 8**). *Drosophila* järjestused moodustasid hea bootstrap väärtusega toetatud haru. Samuti vastasid viiruste harude klasterdumised tegelike viiruste jaotumisega perekondadesse, olles toetatud bootstrap-väärtustega.



Joonis 8. NCBI vl ja TMV-CP rakulisi järjestusi kasutades konstrueeritud fülogeneetiline puu. Puu koostati NJ meetoditega paketis MEGA5. Rohelisega on märgitud rakulised organismid. Punastega on märgitud harude perekondadesse jaotumise bootstrap väärtus.

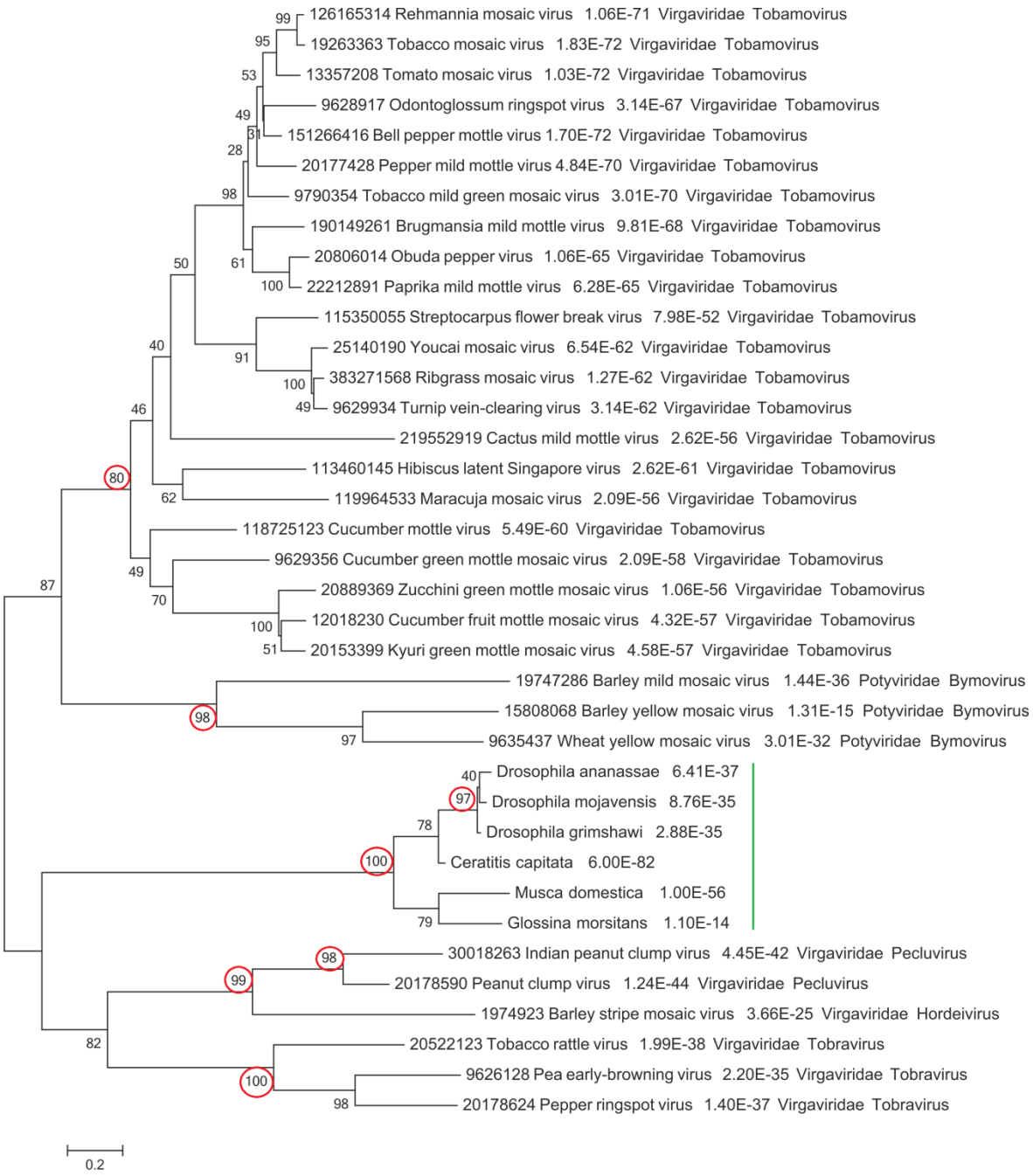
Lootes, et rohkem järjestusi parandab viiruste ja rakuliste organismide lahknemist fülogeneetilisel puul (**Joonis 8**), teostasin NCBI andmebaasis tblastn otsingu. Vasteteks sain suurel hulgal erinevaid *Drosophila* kärbeste liike ning lisaks ka kaks eukarüootset organismi, kes kõik ületasid filtreerimise lävendväärtust:

- Vahemere puuviljakärbes *Ceratitis capitata* (**Joonis 9**), tblastn-i E-väärtusega $6.00*10^{-82}$
- Harilik toakärbes *Musca domestica* (**Joonis 9)**, tblastn-i E-väärtusega $1.00*10^{-56}$

Joonis 9. Pildil vasakul *Ceratitis capitata* ja paremal *Musca domestica*. *C. capitata* pildi autor on Enio Branco (http://www.treknature.com/gallery/South_America/Brazil/photo185532.htm, 28.10.2008) ja *M. domestica* pilt pärit *Natural History Notebooks*, Kanada Loodusmuuseumi koduleheküljelt (http://nature.ca/notebooks/english/fly.htm, alla laetud 19.05.2013).

Drosophila kärbeste järjestusi ei arvestanud, kuna enamus neist ei ületaks kõdususe määra. *C. capitata* ja *M. domestica* järjestuste valgustruktuuride ennustamine LOMETS serveris kinnitas nende kuuluvust TMV-CP superperekonda.

Täiendavate andmetega konstrueeritud fülogeneeriline puu (**Joonis 10**) klasrerdas uued kärbseliste järjestused topoloogiliselt sarnaselt organismide evolutsioonilisele lahknemisele (**Joonis 11**). Ühtsed harud moodustasid *C. capitata* ja *Drosophila* liigid ning *G. morsitans* ja *M. domestica*. Viirusperekondade jaotumises ei esinenud olulisi erinevusi võrreldes esimese konstrueeritud puuga (**Joonis 8**). Viiruste lahknemist perekondadesse toetab kõrge bootstrap väärtus.



Joonis 10. NCBI vl, TMV-CP rakulisi ja lisatud kärbeste järjestustega konstrueeritud fülogeneetiline puu. Puu koostati NJ meetodiga paketis MEGA5. Rohelisega on märgitud rakulised organismid. Punastega on märgitud harude perekondadesse jaotumise bootstrap väärtus.

2.3.6. Fülogeneetiliste puude konstrueerimine kasutades UP valimit

SUPFAM-is oli välja toodud ka UNIPROT-i andmebaasist leitud järjestused, mis peaksid kuuluma uuritava valkude superperekonna hulka. Lootes, et suurem andmete maht

parandab järjestuste lahknemist, konstrueeriti UP valimit kasutades teine fülogeneetiline puu (**Lisa 1**).

Drosophila järjestused grupeerusid koos tse-tse kärbsega (*Glossina morsitans*) ühte harusse. Viirused klassifitseeriti perekonna tasemel, kuid enamus evolutsiooniliselt vanimad lahknemisi ei andnud usaldusväärseid bootstrap-i väärtuseid. Lisades tblastn-i otsingus saadud kahe kärbse järjestused (*Ceratitis capitata* ja *Musca domestica*; **Lisa 2**) paranesid tobamoviiruste ja bümoviiruste grupeerumine. Mõlemal fülogeneetilise puu puhul paiknes *Chara australis* viirus kärbseliste haru läheduses, kuid nõrga boostrap väärtusega. HCRSV karmoviirus paigutati hibiskus nakatava tobamoviiruse (*Hibiscus latent Fort Pierce* viirus) alaharusse, mis võib viidata HCRSV kuuluvusele tobamoviirus hulka.

2.4. ARUTELU

Antud bakalaureuse töö tõestab, erinevate meetoditega kontrollimise teel, superperekonna tasemel TMV-CP domeeni esinemist nii eukarüootsetes organismides, kui ka viirustes. Lisaks tõestatakse valgustruktuuri kaasamise tähtsust valgudomeenide otsingutes, kuna tavalised BLAST perekonna programmidega teostatud eukarüootsete järjestuste otsing viirustest ei anna usaldusväärseid vasteid ehk järjestusi ei leita viirustest üles.

Tõendid, et järjestust leidub vaid teatud taimeviirustel ning osadel *Diptera* esindajatel, mitte enamus organismidel, kinnitab toimunud geeniülekannet. Sel juhul tekib küsimus: kas järjestus levis organismidelt viirusele (H2V) või vastupidi (V2H)?

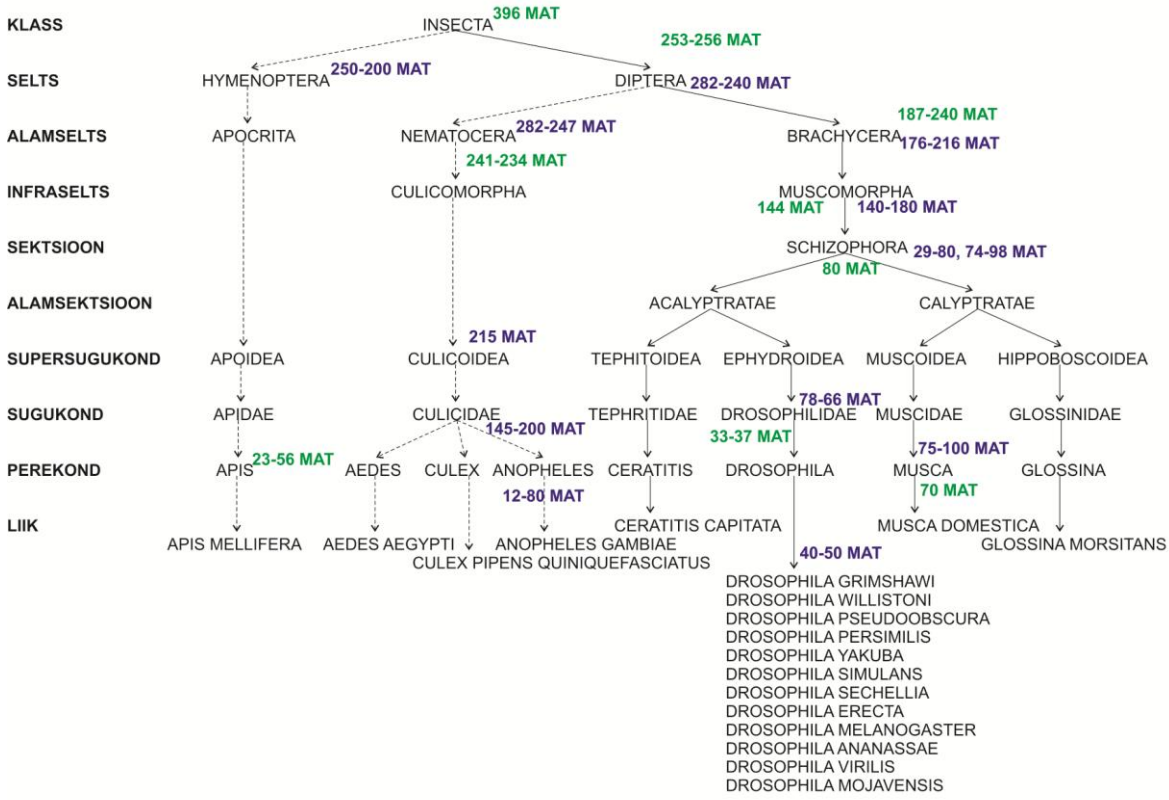
GEENIDE LEVIK VIIRUSTELE (H2V)

H2V toimumiseks peaks eukarüootse organismi mRNA sattuma kas viiruse kapsiidi või replikatsiooni ajal tsütoplasmaatilistesse vabrikutesse, kus peaks olema toimunud viiruse ja peremehe RNA-de vaheline rekombinatsioon. Selle tulemusel peaks ekspresseeritud valk andma evolutsioonilise eelise võrreldes eelneva kattevalguga, et järjestus kinnituks ning leviks kiiresti viiruste seas.

Probleeme tekitab aga *Virgaviridae* viiruste perekonnad tobamoviirused ja tobraviirused, mis levivad taimede piires ainult mehhaaniliselt – taimedesse satuvad viirused ainult taimevigastuse kaudu. Levimiseks võidakse kasutada putukvektorite abi, kuid sellisel juhul

ei pea viirus organismi rakku sisenema ning vektori RNA sattumine virioni on raskendatud.

Protsess oleks pidanud toimuma enne tobamoviiruste teket (120 MAT (Stobbe *et al*., 2012)), kuna järjestus esineb mitmetel *Virgaviridae* perekonnal. Praeguste tõendite põhjal võib väita, et järjestus on kärbselistes kindlasti olnud 50-80 MAT (**Joonis 11**). See teeks aga H2V ülekande suuna võimatuks. Samas aga kui arvestada, et sääselistel (*Nematocera*) ei esine insertsiooni ning kärbseliste (*Brachycera*) genoomides peaks järjestus olemas olema, teeks see järjestuse vanuseks organismides minimaalselt ca 200 MAT, mis näitab H2V ülekande võimalikkust.



Joonis 11. Putukate evolutsioon ning määratud evolutsioneerumise ligikaudsed ajad. Joonisel on välja toodud üldine taksonoomia töös käsitlevate kärbeste kohta: *Drosophila* liigid, *Ceratitis capitata, Musca domestica* ja *Glossina morsitans* ning lähimad sekveneeritud sugulased, kelle genoomist ei ole leitud meid huvitavat geenijärjestust– *Aedes aegypti, Culex pipens quiniquefasciarus, Anopheles Gambia* ja *Apis mellifera*. Joonisele on lisatud arvatavad divergeerumise ajad (sinisega molekulaarsete meetoditega dateeritud ajad ning rohelisega fossiilsete tõendite põhjal määratud ligikaudsed vanused (Gaunt ja Miles, 2002; Krzywinski *et al*., 2006; Wiegmann *et al*., 2003; Wiegmann *et al*., 2011; Engel ja Grimaldi, 2003)).

GEENIDE LEVIK VIIRUSTELT EUKARÜOOTIDE GENOOMI (V2H)

Teades, et taimeviirused võivad kasutada putukvektoreid, leidub võimalus, et viirus on transpordi ajal sattunud organismi sugurakkudesse, kus revertaasi ja rekombineerumiste abil on viiruslik järjestus integreerunud vektori genoomi. Protsessis võidakse kasutada kas organismi või näiteks mõne retroviiruse revertaasi abi.

Juhul kui järjestus oleks eukarüootset päritolu, peaks see olema säilinud paljudes organismides, eriti kärbseliste lähisugulastel. Võimalus, et uus geen tekkis *Brachycera, Muscomorpha* või *Schizophora* esindajal ei ole väga kõrge. Üldiselt tekivad eukarüootsetel organismidel valgud uue funktsiooni tekkimise teel – kas eelnevalt duplitseerunud pseudogeenist või mõnest teisest organismist saadud järjestusest. Seega peaks TMV-CP esinema ka teistes organismides, kellelt see järjestus on laenatud või peaks kärbseliste genoomid andma TMV-CP-le mitme erineva geeni vasteid (kuigi nõrgema skooriga).

Lisaks vajaks järjestuse ülekanne V2H vähem evolutsioonilisi etappe, kui see nõuaks viiruse kattevalgu väljavahetumisel ülekandega H2V.

Kuigi ajaliselt võiks toimuda TMV-CP ülekanne kärbestelt viirustele, on protsessi toimumise tõenäosus väike, samas aga ei saa seda täielikult välistada. Rohkemate putukaliste genoomide sekveneerimine võib pakkuda lisatõendeid V2H/H2V geeniülekandele. Antud töös tõendite baasil järeldan, et toimunud ülekanne toimus viirustelt eukarüootsetele organismidele.

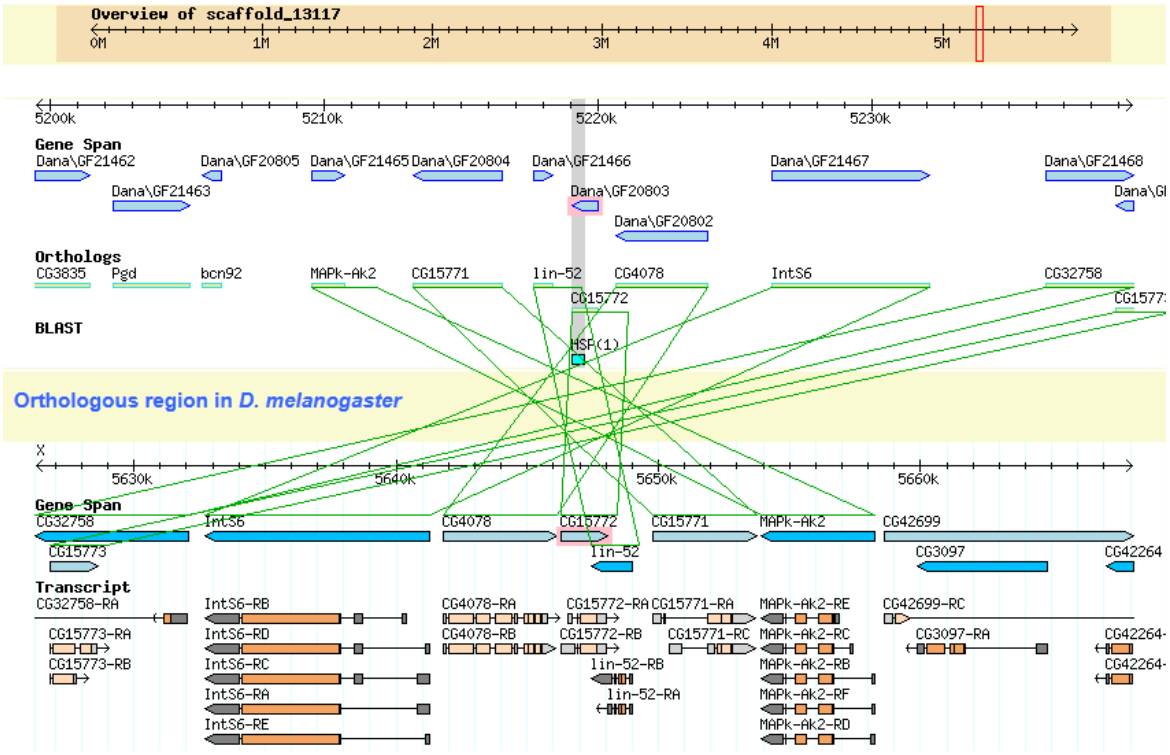
Järgnevalt tuleks välja selgitada kas on toimunud üks või mitu iseseisvat integratsiooni organismidesse. Vaadates konstrueeritud fülogeneetilisi puid (**Joonis 8, 10; Lisa 1, 2**) ning võrreldes neid kärbseliste lahknemistega (**Joonis 11**), võib järeldada, et toimunud on üks integratsioon mõnda kärbseliste eellase genoomi, kuna mõlemate fülogeneetiliste puude kärbeste lahknemised langevad kokku. Kui oleks toimunud vähemalt kaks iseseisvat integratsiooni, siis oleks pidanud selgelt näha olema eristunud kärbeste grupid.

Samuti viitab ühele toimunud insertsioonile *Drosophila*-de TMV-CP järjestuste piirkondade uurimine (**Joonis 12**). Neil esineb järjestus ühes lookuses (kromosoomis X, kui kontiigid on suudetud assambleerida kromosoomideks) ühe eksonina. Lisaks on järjestust ümbritsev ala kõigil *Drosophila*-del kõrge ortoloogsusega, erinedes vaid osaliselt geenide paiknemise järekorras.

Olles saanud kinnitusi toimunud ühest integratsioonist V2H, on võimalik hinnata toimunud integratsiooni aega. See oleks pidanud toimuma kärbseliste esivanemal. Arvestades, et

praeguse seisuga esineb kõigis *Schizophora* esindajatel insertsioon, võiks hinnata EVE minimaalseks vanuseks 50-80 MAT. Seda kinnitab ka tobamoviiruste hinnatav vanus – 120 MAT (Stobbe *et al*., 2012).

Maksimaalseks EVE insertsiooni aja määramiseks tuleks uurida insertsiooniga kärbseliste lähimaid sekveneeritud sugulasi, kellel insertsioon puudub. Nendeks on sääselised *Nematocera* alamseltsist ja mesilased *Hymenoptera* seltsist. Nende abil võiks hinnata insertsiooni ajaks *Insecta* või *Diptera*-de lahknemist (umbes 250 MAT; **Joonis 11**, vaata lk 36). Samas tuleks kindlasti arvestada ka võimalusega, et mingil põhjusel võib neist organismidest olla insertsioon kadunud ning maksimaalne EVE integreerumise aeg võib olla toimunud veel kaugemas ajas. Gibbs *et al*. (2011) hindasid *Chara australis* viiruse ja tobamoviiruste lahknemiseks 238-311 MAT, mis sobib kokku eeldatava TMV-CP maksimaalse insertsiooni ajaga.



Joonis 12. *D. ananassae* uuritava järjestuse ümbritsev ala võrreldes *D. melanogaster*-iga. Antud pilt on võetud FlyBase-i koduleheküljе *D. ananassae* genoomi brauserist. Näidatud on uuritava järjestuse (FBgn0097809 ehk GF20803) ja ümbritsevate geenide paigutuse võrdlust *D. melanogaster*-i ortoloogsete geenidega (FBgn0029799 ehk CG15722 ja ümbritsevad geenid).

Võttes arvesse tõendeid, saab hinnata TMV-CP insertsiooni ajaks keskmiselt 60-250 MAT. See on väga lai ajavahemik, kuid sekveneerides rohkemate putukaliste genoome ning uurides neist TMV-CP olemasolu, saab hakata hindama EVE insertsiooni järjest täpsemalt.

KOKKUVÕTE

Järjest rohkem avastatakse viiruslikke järjestusi hulkraksetest organismidest. Eriliseks üllatuseks oli RNA viiruste järjestuste avastamine. Põhjuseks, miks organism omastab viirusliku järjestuse, on pakutud kaitsemehhanismi. Viiruslikku järjestust võidakse ekspresseerida rakkudes madalal tasemel ning kui toimub tegeliku viiruse infektsioon, on organism võimeline koheselt viiruseid hävitama. Samuti võib organismis viiruslik järjestus omandada uue kasulikuma funktsiooni.

Viiruslike järjestuste uurimine hulkraksetes organismides kirjeldab viiruse-peremeesorganismi interaktsioonide uut tahku - võidurelvastumist, näidates nende suhete mitmekülgsust.

Antud töös uuriti TMV-sarnase viirusliku kattevalgu järjestuse esinemist eukarüootsete organismide genoomides. Töö aluseks on võetud SUPFAM-is kasutatavad HMM mudelid, mis otsivad täielikult sekveneeritud organismide genoomidest ühte superperekonda kuuluvaid valgujärjestusi.

Drosophila kärbeste liikidest ja lisaks kolmest *Schizophora* esindajatest avastati viirustele omane domeen, mida ekspresseeritakse ühe polüpeptiidina. *D. melanogaster*-il on leitud, et valk omab tähtsust pea ja kesknärvisüsteemi arengus, täpsemat funktsiooni ei teata (FlyBase, *D. melanogaster* geeni FBgn0029799 iseloomustus).

Järjestuse ülekande toimumist uuriti võrreldes kärbselistest leitud järjestusi viiruslikega. Konstrueeritud fülogeneetiliste puude lahknemised, organismide dateeritavad vanused ja protsessi toimumise võimalikkus viitavad *Virgaviridae* ja *Potyviridae* kattevalku omava eellase järjestuse integratsioonile mõnda kärbseliste eellase genoomi. Praeguseks mitteteadaolevatel põhjustel on järjestus organismi genoomis kinnistunud.

Uurides kärbseliste ja lähedaste liikide genoome, võib väita, et integratsioon on toimunud vähemalt *Scizophora* esindajatel ehk varem kui 50-80 MAT. Sekveneerides teiste putukaliste genoome ning kontrollides neist EVE-de olemasolu, saab hakata täpsemalt hindama toimunud ülekannet.

Mõistes EVE tuvastamisel esinevaid probleeme, võiks järgnevaks etapiks olla automatiseeritud programmi loomine, mis suudaks teostada töös esitatud etapid ning võimaldaks anda informatsiooni toimunud ülekande kohta.

Arvestades toimunud ülekande suunaks V2H ning teades, et *D. melanogaster*-il eskpresseeritakse valku, võib järeldada viiruste võimalikkusest hulkraksete loomade valgudomeenide allikana.

Could plant viruses be a new source of protein domains for multicellular animals?

Heleri Kirsip

SUMMARY

It is known that retroviruses can integrate into their host genomes – it is part of their life cycle. What came as a surprise was the finding of non-retroviral elements (ssRNA, dsRNA, ssDNA) in eukaryote genomes. They are known as endogenous viral elements (EVEs). The process how an RNA virus derived sequence can integrate into host genome is only hypothesized, not confirmed. In some cases, mostly based on retroviral elements, it has been found that EVEs can be beneficial for the host immune system. For example, they can help to block viral entry to the cell or they can participate in inhibiting viral replication. In non-retroviral EVEs, transcribed mRNAs have been confirmed to be present in the cell but not much is known of their function or protein production.

The aim of this bachelor thesis is to examine the transfer of genetic material between viruses and eukaryotes. It is achieved by following these steps:

- Confirming the actual transfer of genetic material.
- Confirming the direction of the transfer.
- Confirming whether there were one or two independent integrations into the genome.
- Evaluating the time of the integration.

The long-term aim is to automate the whole process by creating a program that can estimate the exchange of genetic material between viruses and organisms based on protein sequence and structure.

This research focuses on TMV-like viral coat protein (TMV-CP) found in several eukaryote genomes (*Drosophila* fly species). The approach is based on SUPFAM database, which uses HMM models for finding protein sequences that belong to one superfamily.

The transfer of genetic material was confirmed by examining eukaryote genomes and protein structures and finding TMV-CP in several viruses. By constructing a phylogenetic tree and examining the closest sequenced relatives of *Drosophila* it was concluded that there was one integration event from viruses to the host genomes. By examining the

phylogeny of the flies and the estimated time of their divergence it was concluded that the minimal time when the integration occurred 50–80 MYA – the time of the divergence of the section of true flies – *Schizophora*.

By examining the transfer of genetic material we can understand the complex relationship of viruses and their hosts. And thus we can understand the impact that viruses have on the evolution of multicellular organisms and their protein families.

KIRJANDUSE LOETELU

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

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. J. Mol. Biol. 215 (3): 403-410.

Ambrose HE, Clewley JP (2006). Virus discovery by sequence-independent genome amplification. Rev. Med. Virol. 16: 365-383.

Anderson RE, Brazelton WJ, Baross JA (2011). Is the genetic landscape of the deep subsurface biosphere affected by viruses? Frontiers in Microbiology. 2 (219): 1-16.

Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. PNAS. 108 (42): 17486-17491.

Aswad A, Katzourakis A (2012). Paleovirology and virally derived immunity. Trends in Ecology and Evolution. 27: 627-636.

Bejarno ER, 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 U S A. 93: 759-764.

Belyi VA, Levine AJ, Skalka AM (2010). 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: 12458-12462.

Bergh O, Borsheim KY, Bratbak G, Heldal M (1989). High abundance of viruses found in aquatic environments. Nature. 340: 467-468.

Breitbart M (2012). Marine viruses: truth or dare. Annu. Rev. Mar. Sci. 4: 425-448.

Breitbart M, Rohwer F (2005). Here a virus, there a virus, everywhere the same virus? TRENDS in Microbiology. 13 (6): 278-284.

Breitbart M, Rohwer F (2005). Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. Biotechniques. 39: 151-156.

Breitbart M, Salamon P, Andersen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F (2002). Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. USA. 99: 14250-14255.

Brüssow H, Hendrix RW (2002). Phage genomics: small is beautiful. Cell. 108(1): 13-16.

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

Crochu S, Cook A, Attoui H, Charrel RN, De Chesse R, Belhouchet M, Lemasson JJ, de Micco P, de Lamballerie X (2004). Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of *Aedes* ssp. mosquitoes. J. Gen. Virol. 85: 1971-1980.

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

Delwart EL (2007). Viral metagenomics. Rev. Med. Virol. 17: 115-131.

Dong J, Olano JP, McBride JW, Walker DH (2008). Emerging pathogens: challenges and successes of molecular diagnostics. J. Mol. Diagn. 10: 185-197.

Duffy A, Shackelton LA, Holmes EC (2008). Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genetics. 9: 267-276.

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

Engel MS, Grimaldi DA (2004). New light shed on the oldest insect. Nature. 427 (6975): 627-630.

Falkow S (2004). Molecular Koch's postulates applied to bacterial pathogenicity – a personal recollection 15 years later. Nat. Rev. Microbiol. 2(1): 67-72.

Feschotte C, Gilbert C (2012). Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Microbiol. 13: 283-296.

Finsterbusch T, Mankertz A (2009). Porcine circoviruses – Small but powerful. Virus Research. 143: 177-183.

Firth C, Charleston MA, Duffy A, Sharpire B, Holmes EC (2009). Insights into the evolutionary history of an emerging livestock pathogen: porcine circovirus 2. J. Virol. 83: 12813-12821.

Fouts DE (2006). Phage_Finder: Automated identification and classification of prophade regions in complete bacteria genome sequence. Nucleic Acids Research. 34 (20): 5839-5851.

Fuhrman JA, Noble RT (1995). Viruses and protists cause similar bacterial mortality in coastal seawater. Limnol. Oceanogr. 40: 1236-1242.

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

Geuking MB, Weber J, Dewannieux M, Gorelik E, Heidmann T, *et al.* (2009). Recombination of retrotransposons and exogenous RNA virus results in nonretroviral cDNA integration. Science. 323: 393-396.

Gibbs A, Skotnicki AH, Gardiner JE, Walker ES (1975). A Tobamovirus of a Green Alga. Virology. 64: 571-574.

Gibbs AJ, Torronen M, Mackenzie AM, Wood JT, Armstrong JS, Kondo H, Tamada T, Keese PL (2011). The enigmatic genome of *Chara australis* virus. Journal of General Virology. 92: 2679-2690.

Goddard TD, Huang CC, Ferrin TE (2005). Software Extensions to UCSF Chimera for Interactive Visualization of Large Molecular Assemblies. Structure. 13: 473-482.

Gordana A, Gough J, Teichmann SA (2001). An insight into domain combinations. Bioinformatics. 1(1): 1-7.

Gough J (2002). Hidden Markov models and their application to genome analysis in the context of protein structure. PhD thesis. Sidney Sussex College.

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

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, 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.

Hanada K, Suzuki Y, Gojobori T (2004). A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Mol. Biol. Evol. 21 (6): 1074-1080.

Harris JR (1991). The evolution of placental mammals. FEBS Lett. 295: 3-4.

Heldal M, Bratbak G (1991). Production and decay of viruses in aquatic environments. Mar. Ecol. Prog. Ser. 72: 205-212.

Holmes EC (2003). Molecular Clocks and the Puzzle of RNA Virus Origins. Journal of Virology. 77 (7): 3893-3897.

Holmes EC (2009). The Evolutionary Genetics of Emerging Viruses. Annu. Rev. Ecol. Evol. Syst. 40: 353-372.

Holmes EC (2011). The evolution of endogenous viral elements. Cell Host Microbe. 10: 368-377.

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

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

Huang M, Koh DC, Weng LJ, Chang ML, Yap YK, Zhang L, Wong SM (2000). Complete nucleotide sequence and genome organization of hibiscus chlorotic ringspot virus, a new member of the genus Carmovirus: evidence for the presence and expression of two novel open reading frames. J. Virol. 74(7): 3149-3155.

Jenkins GM, Rambaut A, Pybus OG, Holmes EC (2002). Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis. Journal of Molecular Evolution. 54 (2): 156-165.

Katoh K, Misawa K, Kuma K, Miyata T (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl. Acids Res. 30 (14): 3059-3066.

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

Klenerman P, Hengartner H, Zinkernagel RM (1997). A non-retroviral RNA virus persists in DNA form. Nature. 390: 298-301.

Krzywinski J, Grushko OG, Besansky NJ (2006). Analysis of the complete mitochondrial DNA from *Anopheles funetus*: an improved dipteran mitochondrial genome

annotation and a temporal dimension of mosquito evolution. Molecular Phylogenetics and Evolution. 39: 417-423.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, *et al.* (2001). Initial sequencing and analysis of the human genome. Nature. 409: 860-921.

Lang AS, Zhaxybayeva O, Beatty JT (2012). Gene transfer agents: phage-like elements of genetic exchange. Nat. Rev. Microbiol. 10: 472-482.

Lawrence JG, Hatfull GF, Hendrix RW (2002). Imbroglios of Viral Taxonomy: Genetic Exchange and Failings of Phenetic Approaches. Journal of Bacteriology. 184 (17): 4891-4905.

Lee S, Hallam SJ (2009). Extraction of high molecular weight genomic DNA from soils and sediments. J. Vis. Exp. 1596.

Middleboe M, Lyck P (2002). Regeneration of dissolved organic matter by viral lysis in marine microbial communities. Aquat. Microb. Ecol. 27: 187-194.

Mokili JL, Rohwer F, Dutilh BE (2012). Metagenomics and future perspectives in virus discovery. Current Opinion in Virology. 2: 63-77.

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.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thorton JM (1997). CATH a hierarchic classification of protein structure. Structure. 5: 1093-1098.

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.

Riley M, Labedan B (1997). Protein evolution viewed through Eschrichia coli protein sequences: introducing the notion of a structural segment of homology, the module. J. Mol. Biol. 268: 857-868.

Rivers TM (1937). Viruses and Koch's Postulates. J. Bacteriol. 33(1): 1-12.

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

Roossinck MJ (2011). The good viruses: viral mutualistic symbioses. Nat. Rev. Micriobiol. 9 (2): 99-108.

Salanoubat M, GEnin S, Artguenave F, Gouzy J, Mangenot S, Arlat M, Billault A, Brottier P, Camus JC, Cattolico L, Chandler M, Choisne N, Claudel-Renard C, Cunnac S, Demange N, Gaspin C, Lavie M, Moisan A, Robert C, Saurin W, Schiex T, Siquier P, Thébault P, Whalen M, Wincker P, Levy M, Weissenbach J, Boucher CA (2002). Genome sequence of the plant pathogen *Ralstonia solanacearum*. Nature. 415: 497-502.

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA. 74: 5463-5467.

Schadt EE, Turner S, Kasarskis A (2010). A window into third-generation sequencing. Hum. Mol. Genet. 19: 227-240.

Schatz MC, Delcher AL, Salzberg SL (2010). Assembly of large genomes using second-generation secuencing. Genome Res. 20: 1165-1173.

Schmieder R, Edwards E (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PloS ONE. 6(3): e17288.

Short CM, Shuttle CA (2005). Bearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. Appl. Environ. Microbiol. 71: 480-486.

Stein J, Liang P (2002). Differential display technology: a general guide. Cell. Mol. Life. Sci. 59 (8): 1235-1240.

Stobbe AH, Melcher U, Palmer MW, Roossinck AJ, Shen G (2012). Co-divergence and host-switching in the evolution of tobamoviruses. Journal of General Virology. 93: 408-418.

Suzuki Y, Gojobori T (1997). The origin and evolution of Ebola and Marburg viruses. Mol. Biol. Evol. 14 (8): 800-806.

Suttle CA (2005). Viruses in the sea. Nature. 437: 356-361.

Suttle CA (2007). Marine viruses – major players in the global ecosystem. Nat. Rev. Microbiol. 5: 801-812.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distances, and Maximum Parsimont Methods. Molecular Biology and Evolution. 28: 2731-2739.

Tang P, Chiu C (2010). Metagenomics for the discovery of novel human viruses. Future Microbiol. 5: 177-189.

Taylor DJ, Bruenn J (2009). The evolution of novel fungal genes from non-retroviral RNA viruses. BMC Biol.

Taylor DJ, Leach RW, Bruenn J (2010). Filoviruses are ancient and integrated into mammalian genomes. BMC Evlolutionary Biology. 10: 193

Thingstad TF, Lignell R (1997). Theoretical models for the control of bacterial growth rate, abundance and carbon demand. Aquat. Microb. Ecol. 13: 19-27.

Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009). Laboratory procedures to generate viral metagenomics. Nature Protocols. 4: 470-483

Vandamme AM, Bertazzoni U, Salemi M (2000). Evolutionary strategies of human T-cell lymphotrophic cirus type II. Gene. 261: 171-180.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009). Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. Bioinformatics. 25(9): 1189-1191.

Weinbauer MG, Fuks D, Puskaric S, Peduzzi P (1995). Diel, seasonal, and depth-related variability of viruses and dissolved DNA in the northern Adriatic Sea. Microb. Ecol. 30: 25-41.

Weinbauer MG, Rassoulzadegan F (2004). Are viruses driving microbial diversification and diversity? Environmental Microbiology. 6 (1): 1-11.

Whon TW, Kim MS, Roh SW, Shin NR, Lee HW, Bae JW (2012). Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. Journal of Virology. 86 (15): 8221-8231.

Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim JW, Lambkin C, Bertone MA, Cassel BK, Bayless KM, Heimberg AM, Wheeler BM, Peterson KJ, Pape T, Sinclair BJ, Skevington JH, Blagoderov V, Caravas J, Kutty SN, Schmidt-Ott U,

Kampmeier GE, Thompson FC, Grimaldi DA, Beckenbach AT, Courtney GW, Friedrich M, Meier R, Yeates DK (2011). Episodic radiations in the fly tree of life. PNAS. 108 (14): 5690-5695.

Wiegmann BM, Yeates DK, Thorne JL, Kishino H (2003). Time flies, a new molecular time-scale for Brachyceran fly evolution without a clock. Syst. Biol. 52 (6): 745-756.

Wilhelm SW, Suttle CA (1999). Viruses and nutrient cycles in the sea. Bioscience. 49:781-788.

Williamson KE, Wommack KE, Radosevich M (2003). Sampling Natural Viral Communities from Soil for Culture-Independent Analyses. Appl. Environ. Microbiol. 69 (11): 6628-6633.

Wommack KE, Colwell EE (2000). Virioplankton: viruses in aquatic ecosystems. Microbiol. Mol. Biol. Rev. 64: 69-114.

Wu S, Zhang Y (2007). LOMETS: A local meta-threading-server for protein structure prediction. Nucleic Acids Research. 35(10): 3375-3382.

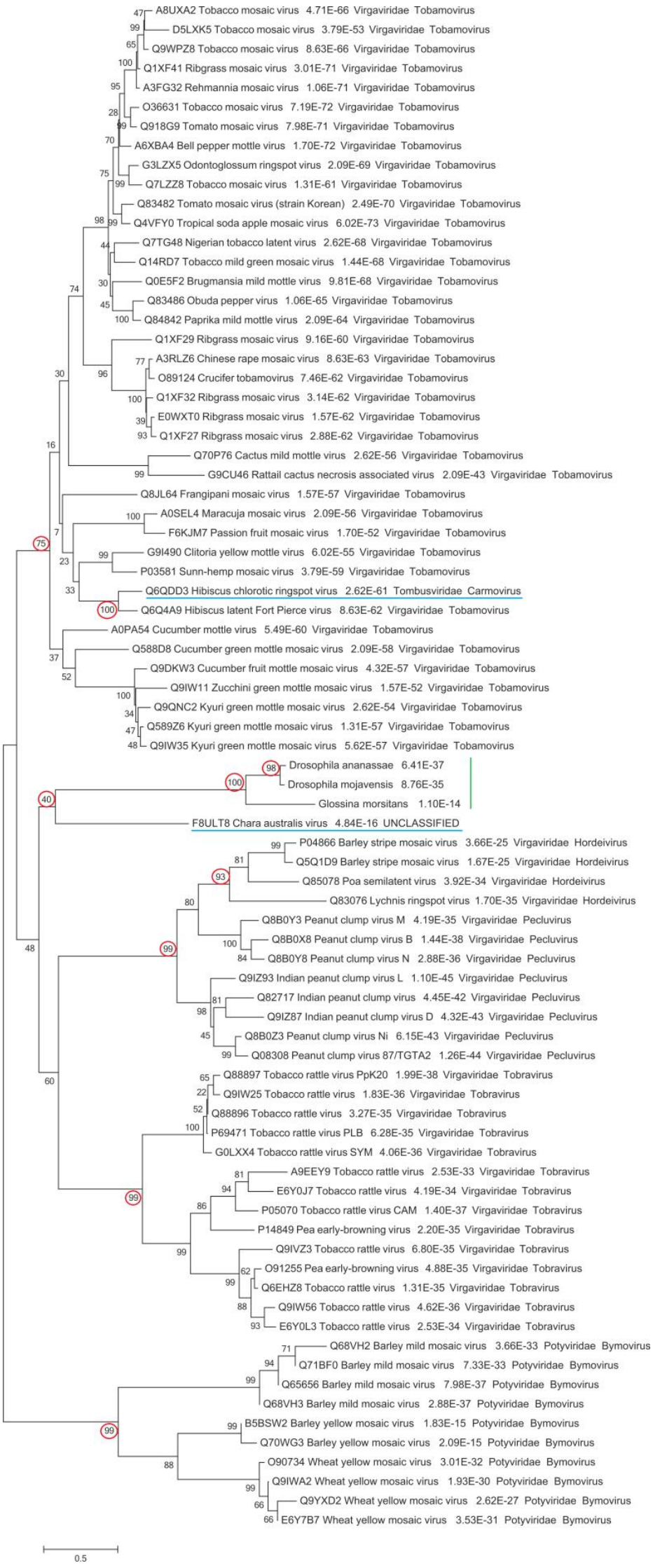
Xu P, Chen F, Mannas JP, Feldman T, Sumner LW, Roossinck MJ (2008). Virus infection improves drought tolerance. New Phytologist. 180: 911-921.

KASUTATUD VEEBIAADRESSID

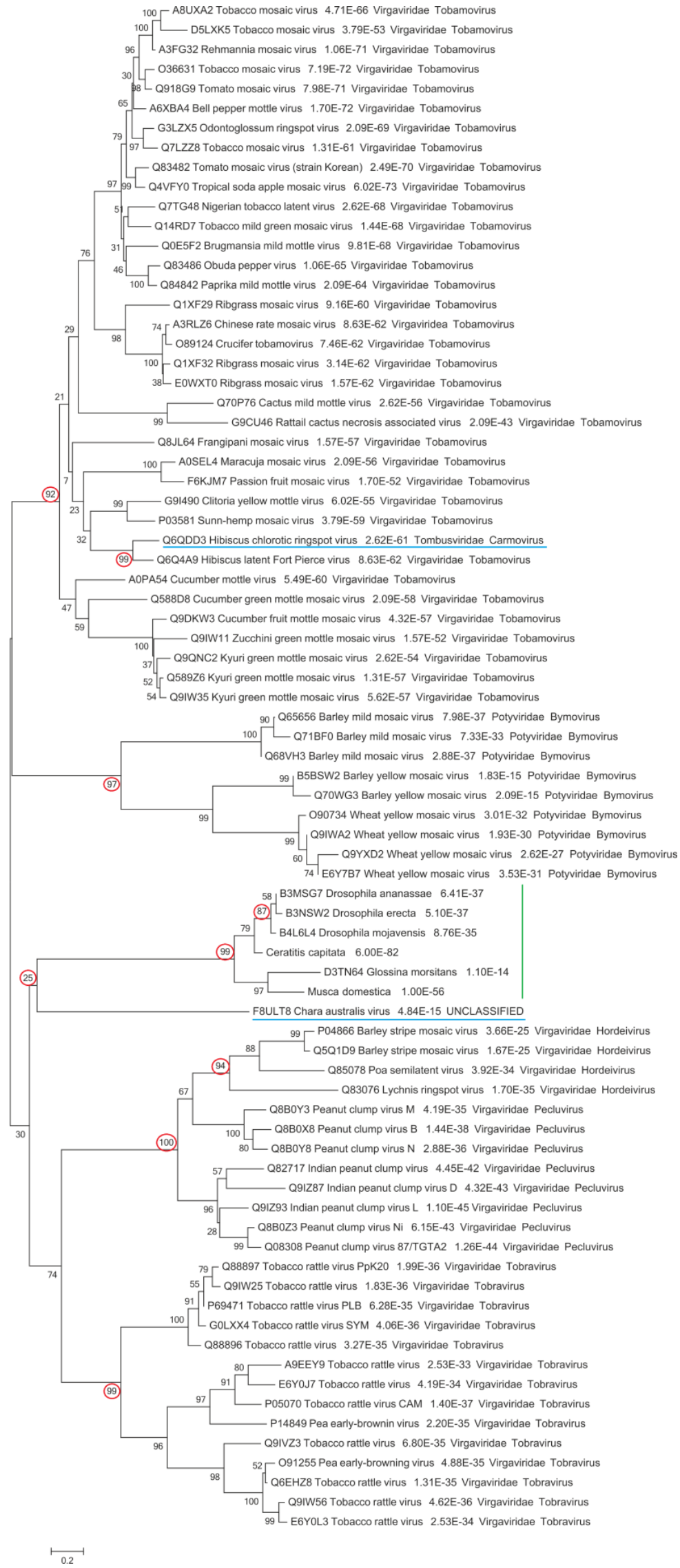
- http://blast.ncbi.nlm.nih.gov/Blast.cgi
- http://flybase.org/
- http://flybase.org/cgi-bin/gbrowse/dana/
- http://genome.jgi-psf.org/Mucci1/Mucci1.home.html
- http://genome.jgi-psf.org/Mucci2/Mucci2.home.html
- http://genome.jgi-psf.org/Phybl1/Phybl1.home.html
- http://genome.jgi-psf.org/Phybl2/Phybl2.home.html
- http://health.howstuffworks.com/human-body/systems/respiratory/question98.htm
- http://metazoa.ensembl.org/Pristionchus_pacificus/Info/Index?db=core;h=BLAST_NEW:BLA_CkcmC4gbO!!;r=Ppa_Contig9:1623875-1758876
- http://nature.ca/notebooks/english/fly.htm
- http://scop.berkeley.edu/
- http://www.ebi.ac.uk/genomes/virus.html
- http://www.ensembl.org/Drosophila_melanogaster/Info/Index
- http://www.nzitrap.com/Biting/biting.htm
- http://www.phytozome.net/search.php
- http://www.rcsb.org/pdb/home/home.do
- http://www.supfam.org/SUPERFAMILY/
- http://www.treknature.com/gallery/South_America/Brazil/photo185532.htm
- http://zhanglab.ccmb.med.umich.edu/LOMETS/

LISAD

Lisa 1. UP valimit ja TMV-CP rakulisi järjestusi kasutades konstrueeritud fülogeneetiline puu. Puu koostati NJ meetodiga paketis MEGA5. Rohelisega on märgitud rakulised organismid. Sinisega on märgitud kaks viirust, millega esinesid probleemid (vaata teemat 2.3.4) ning punasega on märgitud tähtsamate lahknemiste bootstrap-väärtused.



Lisa 2. UP valimit, TMV-CP rakulisi ja lisatud kärbeste järjestusi kasutades konstrueeritud fülogeneetiline puu. Puu koostati NJ meetodiga paketis MEGA5. Rohelisega on märgitud rakulised organismid. Sinisega on märgitud kaks viirust, millega esinesid probleemid (vaata teemat 2.3.4) ning punasega on märgitud tähtsamate lahknemiste bootstrap-väärtused.



Lisa 3. TMV-CP ja lisatud kärbeste järjestused koos genoomsete andmetega.

Organism	Järjestuse nimi	Regioon	Regiooni pikkus	SUPFAM-i E-väärtus	Genoomi vaste	Asukoht genoomis		Kontiigi/kromosoomi pikkus	Geenide lookused	Kodeerivate eksonite arv
Pristinochus pacificus	PPA28277	46-105	60	3,27E-05	PPA28277	Kontiig 24	1 274 085 - 1 274 344	1 693 104	Sama	3
Phycomyces blakesleaanus	jgi|Phybl1|62563|fgeneshPB_pg.5__401	2-147	146	1,27E-06	gfgenesh1_pg.3_#_501	Kontiig 3	1 780 021 - 1 780 470	3 118 726		1
	jgi|Phybl1|73235|fgeneshPB_pg.67__9	2-146	145	5,02E-05	Genmark1.15211_g	Kontiig 40	340 586 - 341 038	370 815		1
Mucor circinelloides	jgi|Mucci1|79906|fgeneshMC_pg.3_#_667	3-154	152	5,76E-06	Mucci1.fgeneshMC_pg.3_#_667	Kontiig 3	2 168 507 - 2 168 974	4 868 387		1
Panicum virgatum	Pavirv00017506m|PACid:23804625	49-187	139	8,76E-07	Pavirv00017506m	Kontiig 50355	1 771 - 2 337	5 693		1
Drosophila qrimshawi	FBpp0158041	119-267	149	2,88E-35	FBgn0131591	Scaffold_14853	7 445 178 - 7 445 981	10 151 454		1
Drosophila willistoni	FBpp0254067	125-271	147	3,14E-36	FBgn0226883	Scaffold2_1100000004909	11 920 554 - 11 921 372	12 416 693		1
Drosophila pseudoobscura	FBpp0272130	106-253	148	1,29E-36	FBgn0073979	XL_group3a	1 870 773 - 1 871 494	2 692 213		1
Drosophila persimilis	FBpp0178572	58-205	148	6,93E-37	FBgn0152070	Scaffold 11	2 815 250 - 2 815 867	2 846 995		1
Drosophila yakuba	FBpp0261436	100-247	148	5,10E-37	FBgn0233952	X	2 825 862 - 2 826 605	21 770 863		1
Drosophila simulans	FBpp0215150	104-251	148	5,36E-37	FBgn0188336	X	4 386 169 - 4 386 924	17 042 790		1
Drosophila sechellia	FBpp0193910	103-250	148	2,09E-36	FBgn0167370	Scaffold_4	1 116 487 - 1 117 239	6 179 234		1
Drosophila melanogaster	FBpp0070774	112-259	148	5,89E-37	FBgn0029799	X	5 646 282 - 5 648 074	22 422 827		1
Drosophila erecta	FBpp0137327	100-247	148	5,10E-37	FBgn0110989	Scaffold_4690	2 995 706 - 2 996 449	18 748 788		1
Drosophila ananassae	FBpp0123995	131-279	149	6,41E-37	FBgn0097809	Scaffold_13117	5 219 017 - 5 219 968	5 790 199		1
Drosophila virilis	FBpp0231425	91-238	148	1,57E-34	FBgn0204185	Scaffold_12928	6 416 473 - 6 417 189	7 717 345		1
Drosophila mojavensis	FBpp0165607	115-262	148	8,76E-35	FBgn0139138	Scaffold_6328	2 090 316 - 2 091 104	4 453 435		1
Glossina morsitans	D3TN64	24-174	151	1.10E-14						
Ceratitis capitata					LOC1010450673 regulator of telomere elongation helicase 1 homolog	Kontiig15288	16 895 - 17 440	33 758		
Musca domestica						Kontiig	24 649 - 25 167	25 568		

Lisa 4. NCBI vl andmebaasist saadud järjestused.

Valgu ID	Taksonoomia ID	Viiruse nimi
9626128	12294	Pea early-browning virus
9628917	12238	Odontoglossum ringspot virus
9629356	12235	Cucumber green mottle mosaic virus
9629934	29272	Turnip vein-clearing virus
9635437	75746	Wheat yellow mosaic virus
9790354	12241	Tobacco mild green mosaic virus
12018230	146499	Cucumber fruit mottle mosaic virus
13357208	12253	Tomato mosaic virus
15808068	12465	Barley yellow mosaic virus
19263363	12242	Tobacco mosaic virus
19744923	12327	Barley stripe mosaic virus
19747286	12466	Barley mild mosaic virus
20153399	111970	Kyuri green mottle mosaic virus
20177428	12239	Pepper mild mottle virus
20178590	28355	Peanut clump virus
20178624	31750	Pepper ringspot virus
20522123	12295	Tobacco rattle virus
20806014	31749	Obuda pepper virus
20889369	111418	Zucchini green mottle mosaic virus
22212891	35281	Paprika mild mottle virus
25140190	228578	Youcai mosaic virus
30018263	32629	Indian peanut clump virus
113460145	185955	Hibiscus latent Singapore virus
115350055	335187	Streptocarpus flower break virus
118725123	388038	Cucumber mottle virus
119964533	368736	Maracuja mosaic virus
126165314	425279	Rehmannia mosaic virus
151266416	368735	Bell pepper mottle virus
190149261	402399	Brugmansia mild mottle virus
219552919	229030	Cactus mild mottle virus
383271568	51680	Ribgrass mosaic virus

Lisa 5. UP andmebaasi viiruslikud järjestused.

Valgu ID	Taksonoomia ID	Viiruse nimetus
A0PA54	388038	Cucumber mottle virus
A0SEL4	368736	Maracuja mosaic virus
A3FG32	425279	Rehmannia mosaic virus

A3RLZ6	42007	Chinese Rape Mosaic Virus
A6XBA4	368735	Bell pepper mottle virus
A8UXA2	12242	Tobacco mosaic virus
A9EEY9	12295	Tobacco rattle virus
B5BSW2	12465	Barley yellow mosaic virus
D5LXK5	12242	Tobacco mosaic virus
E0WXT0	51680	Ribgrass mosaic virus
E6Y0J7	12295	Tobacco rattle virus
E6Y0L3	12295	Tobacco rattle virus
E6Y7B7	75746	Wheat yellow mosaic virus
F6KJM7	1032457	Passion fruit mosaic virus
F8ULT8	1051671	Chara australis virus
G0LXX4	12298	Tobacco rattle virus-SYM
G3LZX5	12238	Odontoglossum ringspot virus
G9CU46	1123754	Rattail cactus necrosis associated virus
G9I490	1128119	Clitoria yellow mottle virus
O36631	12242	Tobacco mosaic virus
O89124	78276	Crucifer tobamovirus
O90734	75746	Wheat yellow mosaic virus
O91255	12294	Pea early-browning virus
P03581	12240	Sunn-hemp mosaic virus
P04866	12327	Barley stripe mosaic virus
P05070	12296	Tobacco rattle virus-CAM
P14849	12294	Pea early-browning virus
P69471	33766	Tobacco rattle virus-PLB
Q08308	652837	Peanut clump virus 87/TGTA2
Q0E5F2	402399	Brugmansia mild mottle virus
Q14RD7	12241	Tobacco mild green mosaic virus
Q1XF27	51680	Ribgrass mosaic virus
Q1XF29	51680	Ribgrass mosaic virus
Q1XF32	51680	Ribgrass mosaic virus
Q1XF41	51680	Ribgrass mosaic virus
Q4VFY0	327387	Tropical soda apple mosaic virus
Q588D8	12235	Cucumber green mottle mosaic virus
Q589Z6	111970	Kyuri green mottle mosaic virus
Q5Q1D9	12327	Barley stripe mosaic virus
Q65656	12466	Barley mild mosaic virus
Q68VH2	12466	Barley mild mosaic virus
Q68VH3	12466	Barley mild mosaic virus
Q6EHZ8	12295	Tobacco rattle virus
Q6Q4A9	233051	Hibiscus latent Fort Pierce virus
Q6QDD3	53181	Hibiscus chlorotic ringspot virus
Q70P76	229030	Cactus mild mottle virus
Q70WG3	12465	Barley yellow mosaic virus

Q71BF0	12466	Barley mild mosaic virus
Q7LZZ8	12242	Tobacco mosaic virus
Q7TG48	208293	Nigerian tobacco latent virus
Q82717	32629	Indian peanut clump virus
Q83076	44421	Lychnis ringspot virus
Q83482	138313	Tomato mosaic virus (strain Korean)
Q83486	31749	Obuda pepper virus
Q84842	35281	Paprika mild mottle virus
Q85078	12328	Poa semilatent virus
Q88896	12295	Tobacco rattle virus
Q88897	652939	Tobacco rattle virus PpK20
Q8B0Z3	188887	Peanut clump virus Ni
Q8B0X8	188884	Peanut clump virus B
Q8B0Y3	188885	Peanut clump virus M
Q8B0Y8	188886	Peanut clump virus N
Q8JL64	99585	Frangipani mosaic virus
Q918G9	12253	Tomato mosaic virus
Q9DKW3	146499	Cucumber fruit mottle mosaic virus
Q9IZ87	119104	Indian peanut clump virus D
Q9IZ93	119102	Indian peanut clump virus L
Q9IW11	111418	Zucchini green mottle mosaic virus
Q9IW25	12295	Tobacco rattle virus
Q9IW35	111970	Kyuri green mottle mosaic virus
Q9IW56	12295	Tobacco rattle virus
Q9IWA2	75746	Wheat yellow mosaic virus
Q9IVZ3	12295	Tobacco rattle virus
Q9QNC2	111970	Kyuri green mottle mosaic virus
Q9WPZ8	12242	Tobacco mosaic virus
Q9YXD2	75746	Wheat yellow mosaic virus

Lisa 6. LOMETS tulemused, välja toodud organismid *D. ananassae, D. erecta ja D. mojavensis* ning teised kärbselised *G. morsitans, M. domestica* ja *C. capitata.* SCOP-i klassifitseerimist kasutades TMV-CP superperekonda kuuluvad valgumudelid on alla joonitud rohelisega. Heleroheliseg on alla joonitud valgud, mis peaksid kuuluma TMV-CP sf-i, kuid mida ei ole veel klassifitseeritud SCOP-is. Joonised on võetud LOMETS serverite valgustruktuuri ennutuse tulemustest.

Drosophila ananassae

Rank	Template	Align_length	Coverage	Zscore	Seq_id	Confidence Score	Program
1	1ei7A	147	0.986	8.374	0.184	High	MUSTER
2	1vtmP	146	0.979	63.000	0.192	High	FFAS
3	1rmvA	142	0.953	14.900	0.134	Medium	PRC
4	3pdm_P	147	0.986	14.637	0.197	Medium	HHsearch
5	1ei7A	147	0.986	8.374	0.184	Medium	SP3
6	1ei7A	147	0.986	8.374	0.184	Low	SPARKS2
7	4j5uA2	83	0.557	9.464	0.229	Low	SAM
8	1ei7a	149	1	2.623	0.168	Low	PROSPECT2
9	1ei7_A	147	0.986	13.192	0.177	Low	HHsearch
10	3pdmP	146	0.979	7.441	0.199	Low	MUSTER

Drosophila erecta

Rank	Template	Align_length	Coverage	Zscore	Seq_id	Confidence Score	Program
1	1ei7A	146	0.986	8.503	0.185	High	MUSTER
2	1vtmP	144	0.972	63.100	0.187	High	FFAS
3	1rmvA	142	0.959	14.900	0.134	Medium	PRC
4	3pdm_P	146	0.986	14.643	0.199	Medium	HHsearch
5	1ei7A	146	0.986	8.503	0.185	Medium	SP3
6	1ei7A	146	0.986	8.503	0.185	Medium	SPARKS2
7	4j5uA2	82	0.554	9.436	0.220	Low	SAM
8	1r4wA	148	1	2.517	0.122	Low	PROSPECT2
9	1ei7_A	146	0.986	13.198	0.178	Low	HHsearch
10	3pdmP	145	0.979	7.567	0.200	Low	MUSTER

Drosophila mojavensis

Rank	Template	Align_length	Coverage	Zscore	Seq_id	Confidence Score	Program
1	1ei7A	147	0.993	8.037	0.184	High	PPA-I
2	1vtmP	144	0.972	64.900	0.167	High	FFAS
3	3pdm_P	138	0.932	18.702	0.196	High	HHsearch
4	1rmvA	142	0.959	14.300	0.120	Medium	PRC
5	1ei7A	146	0.986	7.710	0.192	Medium	MUSTER
6	1ei7A	146	0.986	7.710	0.192	Low	SP3
7	1ei7A	146	0.986	7.710	0.192	Low	SPARKS2
8	1wfpA	41	0.277	9.055	0.220	Low	SAM
9	1ei7a	148	1	2.731	0.155	Low	PROSPECT2
10	1cgmE	142	0.959	13.100	0.190	Low	PRC

Glossina morsitans

Rank	Template	Align_length	Coverage	Zscore	Seq_id	Confidence Score	Program
1	1ei7A	150	0.993	7.317	0.160	High	PPA-I
2	1vtmP	148	0.980	67.300	0.142	High	FFAS
3	1ei7A	149	0.986	9.664	0.161	Medium	MUSTER
4	1rmvA	144	0.953	12.900	0.111	Medium	PRC
5	1ei7A	149	0.986	9.664	0.161	Medium	SPARKS2
6	1ei7_A	147	0.973	10.746	0.143	Low	HHsearch
7	1ei7a	142	0.940	7.264	0.148	Low	SP3
8	1wfpA	41	0.271	8.796	0.244	Low	SAM
9	1cy0A2	146	0.966	2.622	0.034	Low	PROSPECT2
10	1rmvA	147	0.973	8.774	0.122	Low	MUSTER

Musca domestica

Rank	Template	Align_length	Coverage	Zscore	Seq_id	Confidence Score	Program
1	1ei7A	149	0.866	9.311	0.154	High	PPA-I
2	1vtmP	148	0.860	57.700	0.149	High	FFAS
3	3pdmP	142	0.825	14.600	0.176	Medium	PRC
4	1ei7_A	140	0.813	16.199	0.157	Medium	HHsearch
5	1ei7a	147	0.854	9.163	0.163	Medium	SP3
6	1ei7A	149	0.866	7.324	0.161	Medium	MUSTER
7	1ei7A	149	0.866	7.324	0.161	Low	SPARKS2
8	4j5uA2	106	0.616	10.080	0.208	Low	SAM
9	1h1oA	163	0.947	2.719	0.055	Low	PROSPECT2
10	3pdm_P	141	0.819	15.158	0.170	Low	HHsearch

Ceratitis capitata

Rank	Template	Align_length	Coverage	Zscore	Seq_id	Confidence Score	Program
1	1ei7A	150	0.824	10.292	0.187	High	MUSTER
2	1vtmP	148	0.813	61.000	0.182	High	FFAS
3	3pdmP	142	0.780	16.300	0.204	High	PRC
4	1ei7_A	148	0.813	14.794	0.189	Medium	HHsearch
5	1ei7A	150	0.824	10.292	0.187	Medium	SP3
6	1ei7A	150	0.824	10.292	0.187	Medium	SPARKS2
7	1ei7a	151	0.829	3.177	0.172	Low	PROSPECT2
8	4j5uA2	99	0.543	9.492	0.192	Low	SAM
9	1cgmE	150	0.824	9.660	0.180	Low	MUSTER
10	3pdm_P	143	0.785	14.784	0.217	Low	HHsearch

LIHTLITSENTS

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Heleri Kirsip (05.04.1991)

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

„Kas taimeviirused võivad olla uute valgudomeenide allikaks hulkraksetele loomadele?“,

mille juhendajad on vanemteadur Aare Abroi ja doktorant Tõnu Margus,

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 kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

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

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

Tartus, 25.05.2013