TARTU ÜLIKOOL

LOODUS- JA TEHNOLOOGIATEADUSKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOINFORMAATIKA ÕPPETOOL

Mikk Puustusmaa

Papilloomiviirustes E8 valku kodeeriva ala tuvastamine *in silico*

Magistritöö

Juhendaja: vanemteadur Aare Abroi, PhD

Kaasjuhendaja: professor Maido Remm, PhD

TARTU 2014

SISUKORD

KASUTATUD LÜHENDID

BPV - veisepapilloomiviirus (*Bos taurus papillomavirus*)

CDEP – antud töö raames välja töötatud meetod hindamaks DNA konserveerumuse taset regiooniti (*conserved DNA element detection in protein coding sequence*)

CDS – geeni kodeeriv osa (*coding sequence*)

HPV – inimese papilloomiviirus (*human papillomavirus*)

LCR - pikk regulaatorala (*long control region*) ehk URR ehk NCR

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

NCR – mittekodeeriv regulatoorne ala (*non-coding region*) ehk URR ehk LCR

PV – papilloomiviirus

URR - ülesvoolu paiknev regulatoorne ala (*upstream regulatory region*) ehk LCR ehk NCR

SISSEJUHATUS

Papilloomiviirused on väikesed tsirkulaarse dsDNA genoomiga viirused, mis on võimelised põhjustama kasvajaid nahas ja limaskestas. Papilloomiviirused jaotatakse 35-te perekonda, sealjuures 90% kõikidest kirjeldatud HPV-dest kuuluvad kolme suuremasse: alfa-, beeta- ja gammapapilloomiviirusete perekonda. Inimese papilloomiviirused jagatakse veel eraldi kahte rühma, vastavalt nende potentsiaalile tekitada pahaloomulisi või healoomulisi kasvajaid - kõrge ja madala riskiga HPV-d. Kõrge riskiga HPV-de hulka kuuluvatest viirustest on HPV16 seotud ~50%, HPV18 ~20% ja HPV31 ~4% kõikide emakakaelavähi juhtumitega ning HPV16 on lisaks seostatud veel ~90% teiste anogenitaal ning suu- ja kurgukasvajatega.

Papilloomiviiruste genoomis leidub üldjuhul 8 geeni: 6 varajast (E1,E2 ja E4-E7) ning 2 hilist (L1 ja L2). Varajased geenid ekspresseeruvad papilloomiviiruse infektsiooni algfaasis, kuid hilised geenid viiruse produktiivses faasis, kus toimub virionide kokkupanek. Lisaks eelpool nimetatud geenidele leidub papilloomiviirustes ka mitmesuguseid alternatiivse splaissingu teel loodud transkripte ning nende põhjal kodeeritud valke.

E8^E2C on splaissingu teel saadud produkt, mille puhul on tegemist papilloomiviiruse varajaste geenide transkriptsiooni repressoriga ja genoomi replikatsiooni negatiivse regulaatoriga. Eksperimentaalselt on tõestatud E8^E2C olemasolu alfa-, delta-, beeta, müü- ja kappapapilloomiviirustes. Sealjuures ei ole E8^E2C olemasolu näidatud veel mitmetes teistes suurtes perekondades, näiteks gamma- või lambdapapilloomiviirustes. E8 CDS-i ennustamine papilloomiviiruste genoomis on keeruline, sest E8 puhul on tegemist lühikese järjestusega ~10 aminohapet ning E8 CDS paikneb E1 lugemisraami sees. Mõlemad punktid raskendavad E8 tuvastamist.

Antud töö eesmärkideks on uurida E8 CDS-i olemasolu teistes papilloomiviirustes *in silico* ning välja töötada meetod analüüsimaks DNA tasemel konserveerumuse olulisust.

Töö on valminud Tartu Ülikooli Molekulaar- ja Rakubioloogia Instituudis, bioinformaatika õppetooli juures koostöös Eesti Biokeskusega. Juhendamise ja nõuannete eest soovin tänada juhendajaid Aare Abroid ning Maido Remmi.

1. KIRJANDUSE ÜLEVAADE

1.1. Papilloomiviiruste üldiseloomustus

Papilloomiviirused (PV) on väikesed tsirkulaarse dsDNA genoomiga viirused ning kuuluvad sugukonda *Papillomaviridae.* PV genoom on ligikaudu 8000 aluspaari (bp), kusjuures suurim genoom on 8607 bp CPV1-el (*Canis familiaris papillomavirus* e koera PV) ja väikseim 6953 bp CmPV1-el (*Chelonia mydas papillomavirus* e rohelise merikilpkonna PV) (PaVe andmebaas 26.04.2014). Papilloomiviirused kodeerivad ligikaudu kaheksat erinevat valku, mis jagatakse varajasteks (E) ja hilisteks (L). PV-d on võimelised nakatama suurt hulka amnioote, sealhulgas ka inimesi (Van Doorslaer et al., 2013).

Esimene DNA tuumorviirus, mis avastati, oli CRPV (*Cottontail rabbit papillomavirus* e sooküüliku PV) ehk SfPV1 (*Sylvilagus floridanus* PV). Alles 1980-ndatel ehk ligikaudu viiskümmend aastat hiljem ilmusid esimesed tööd inimese papilloomiviirustega (HPV) HPV16 ja HPV18, mis näitasid tugevat seost emakakaelavähi tekkega. HPV-d nakatavad naha epiteel- või limaskesta rakke, põhjustades proliferatsiooni. Vohamise tagajärjeks on healoomulised papilloomid, mis võivad areneda vähkkasvajateks. Üldjuhul jagatakse inimese papilloomiviirused vastavalt nende potentsiaalile tekitada pahaloomulisi või healoomulisi kasvajaid kõrge ja madala riskiga HPV-deks. Hetkeseisuga on kõige paremini uuritud alfapapilloomiviiruseid, sest just sinna perekonda kuulub mitu kõrge riskiga HPV-d, näiteks eespool mainitud HPV16, HPV18 ning ka HPV31, mis põhjustavad anogenitaal kasvajaid (Sankovski et al., 2014). Sealjuures on teada, et HPV16 on seotud ~50% , HPV18 ~20% ja HPV31 ~4% kõikide emakakaelavähi juhtumitega üle maailma. Lisaks on HPV16 leitud veel ~90% kõikidest teistest anogenitaal ning suu- ja kurgukasvajatest (Fertey et al., 2011; Li, Li, Diaz, & You, 2014; Wang & Roden, 2013). Järjest enam pälvivad tähelepanu ka beetapapilloomiviirused, mis nakatavad naha epiteelrakke, sest on nähtud seost lamerakk-kartsinoomide (naha pahaloomuliste kasvajate alaliik) tekkimise ja beetapapilloomiviiruste vahel. Ligikaudu 90% juhtudest on täheldatud HPV5 ja HPV8 olemasolu, samas otsest seost veel leitud ei ole (Sankovski et al., 2014).

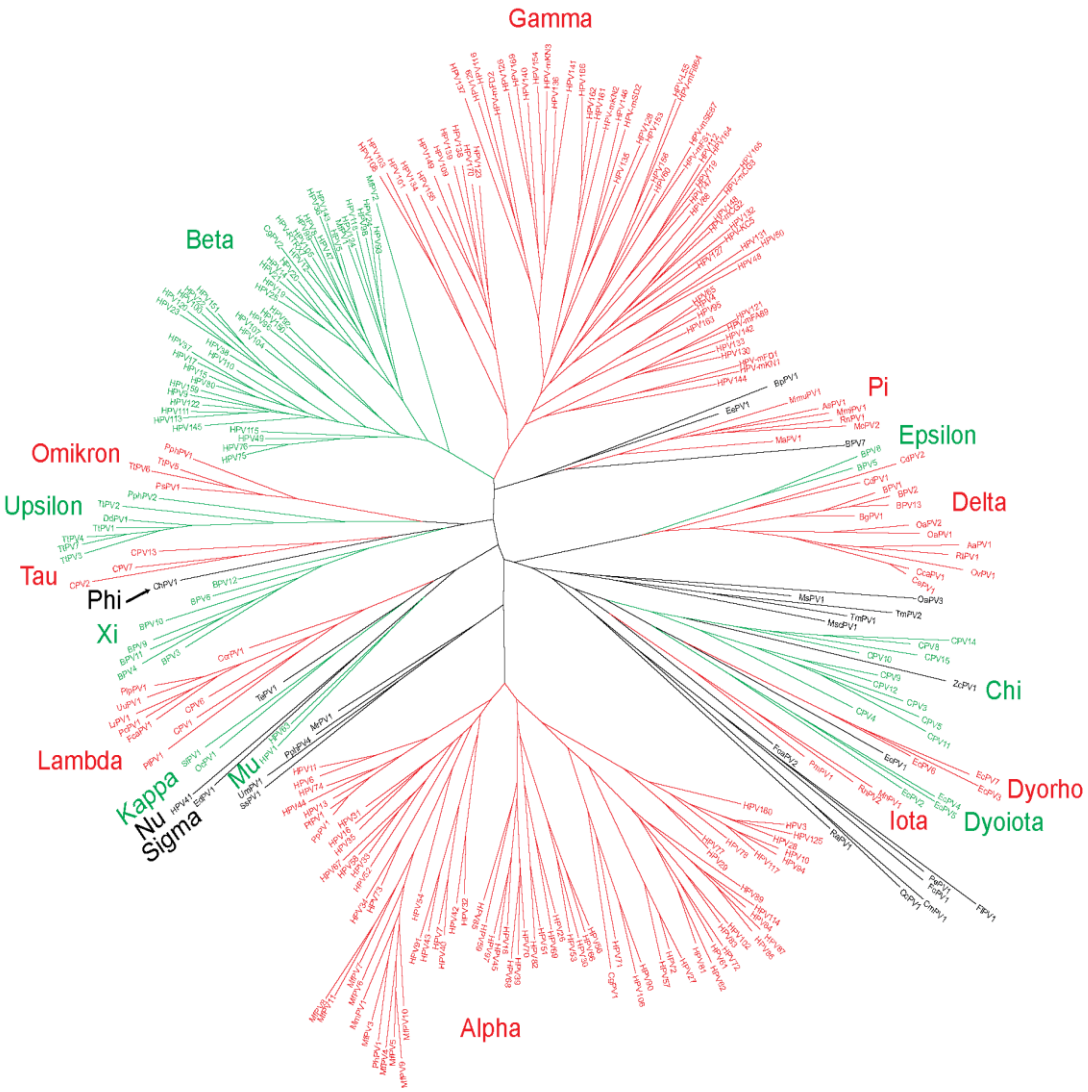
Inimese papilloomiviiruste kõrval on ka väga suur hulk loomade papilloomiviiruseid, hetkel on teada 112 täisgenoomi (Rector & Van Ranst, 2013). Viiruseid on leitud 54-st erinevast peremeesorganismist, millest enamuse moodustavad imetajad: lehm, koer, lõvi, jääkaru,

delfiin jne. Samas on tuvastatud ka kolm linnu (papagoi, metsvint ja mustfrankoliin) ning kolm roomaja (püüton ja 2 kilpkonna liiki) papilloomiviirust. PV-sid ei ole seni leitud kahepaiksetelt ning siinkohal võib spekuleerida, et papilloomiviirused piirduvadki amniootidega (roomajad, linnud ja imetajad) (Rector & Van Ranst, 2013; Shah, Doorbar, & Goldstein, 2010).

1.2. Papilloomiviiruste klassifikatsioon

Papilloomiviiruste uute tüüpide määramiseks rakendatakse järjestuste sarnasusel põhinevaid meetodeid. Klassifitseerimiseks kasutatakse L1 geeni (kapsiidivalk), kuna seda leidub kõikides PV-des ning on üpriski konserveerunud. Uue PV tüübi defineerimiseks peab L1 ORF olema vähem kui 90% sarnane temale taksonoomiliselt kõige lähema papilloomiviiruse L1 ORF-iga. Lisaks peab uue tüübi kinnitamisel eksisteerima ka viiruse täisgenoom (de Villiers, 2013).

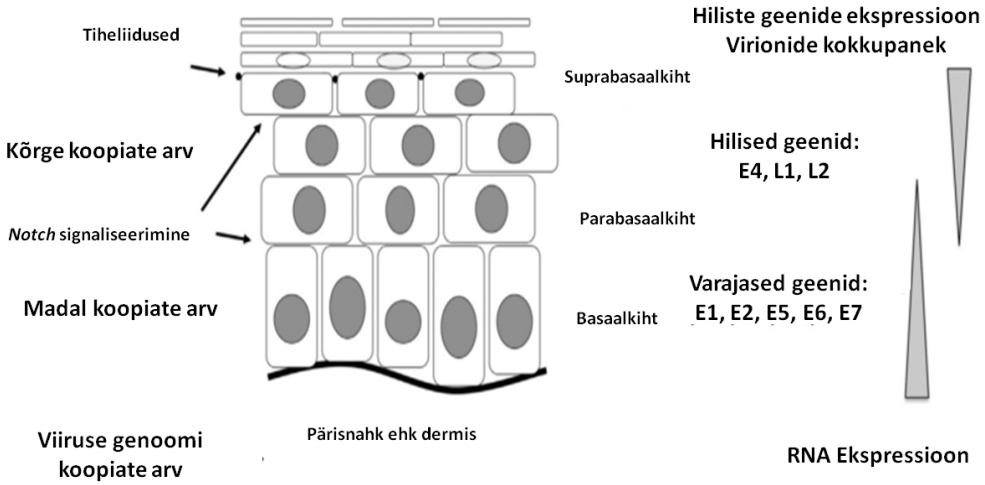
Papilloomiviiruste perekonnad tähistatakse kreeka tähtedega. Eesliide *dyo*, mis tähendab kreeka keeles number kahte, võeti kasutusele, sest klassifitseeritud perekondade arv ületas tähtede arvu kreeka tähestikus (joonis 1). Hetkel on aktsepteeritud ligikaudu 170 HPV tüüpi. Need kategoriseeritakse viide perekonda: alfa-, beeta-, gamma-, müü-, ja nüüpapilloomiviirused, sealjuures alfa-, beeta- ja gammapapilloomiviiruste perekonda kuulub 90% kõikidest kirjeldatud HPV-dest (joonis 1). Loomade papilloomiviirused on jaotunud 32-te perekonda, jättes puutumata ainult gamma-, müü- ja nüüpapilloomiviiruste perekonna, mis sisaldavad eranditult ainult HPV-sid (de Villiers, 2013; Rector & Van Ranst, 2013).



Joonis 1. Papilloomiviiruste fülogeneetiline puu. PV perekonnad nimetatakse kreeka tähtede järgi. Puul on perekonnad eristatud erivärvidega, tähistatud ainult suuremad perekonnad. Dyo eesliide tähendab kreeka keeles number kahte [http://pave.niaid.nih.gov/#prototypes?type=tree] (24.04.14).

1.3. Papilloomiviiruste infektsioon ja genoomi säilimine rakus

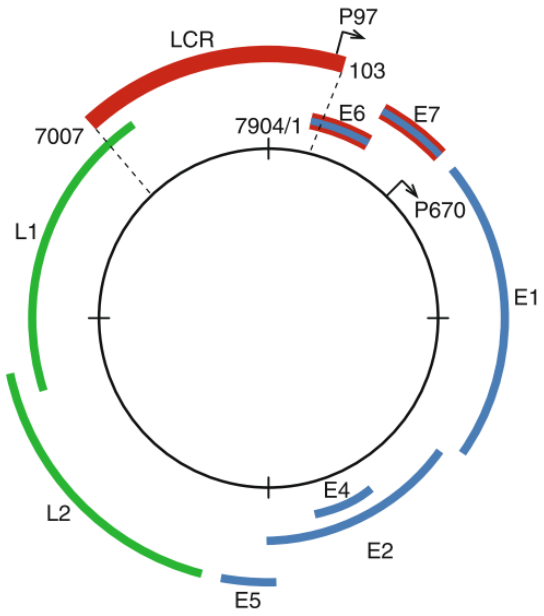
Papilloomiviiruste infektsioon saab alguse vigastusest (ka mikrovigastusest) nahas või limaskestas (Vande Pol & Klingelhutz, 2013). Pärast raku nakatamist transporditakse viiruse genoom (tsirkulaarne dsDNA) tuuma, kus see säilib ekstrakromosomaalse elemendina ehk episoomina. PV-de elutsükkel on tihedalt seotud keratinotsüütide (epidermise rakud) diferentseerumisega, sest viirus nakatab epidermaalseid või limaskesta basaalrakke, kuid virionide kokkupanek toimub alles terminaalselt diferentseerunud rakkudes (Longworth & Laimins, 2004). Arvatavasti on järk-järguline geenide aktiveerimine ja eriti just kapsiidivalkude ekspressioon alles terminaalselt diferentseerunud rakkudes oluline vältimaks immuunsüsteemi (Buck, Day, & Trus, 2013). Varajaste geenide ekspressioon toimub mittediferentseerunud epiteelrakkude basaal– ja parabasaalkihtides (joonis 2). Genoomi amplifikatsioon ja kapsiidivalkude ekspressioon ning ka virionide kokkupanek toimub suprabasaalkihis ning epiteeli diferentseerunud osas (Sankovski et al., 2014).



Joonis 2. Papilloomiviiruste elutsükkel ja geenide ekspressioon eri kihtides. Infektsioon saab alguse epiteelrakkude basaalkihis. Nakkuse algfaasis on viiruse genoomi arv madal. Raku diferentseerudes ehk liikudes basaalkihist ülesse, toimub *Notch* sõltuv diferentseerumine. Diferentseerumine viib viiruse produktiivsesse faasi. Onkovalkude (E5, E6 ja E7) produktsioon indutseerib rake minema uuesti S-faasi ning tagab viraalse genoomi amplifikatsiooni parabasaalkihtides**.** Järgnevalt diferentseerub üks alamhulk rakke ning koos sellega algab viiruse kapsiidivalkude ekspressioon ning virionide kokkupanek (Vande Pol & Klingelhutz, 2013).

1.4. Papilloomiviiruste genoomi struktuur

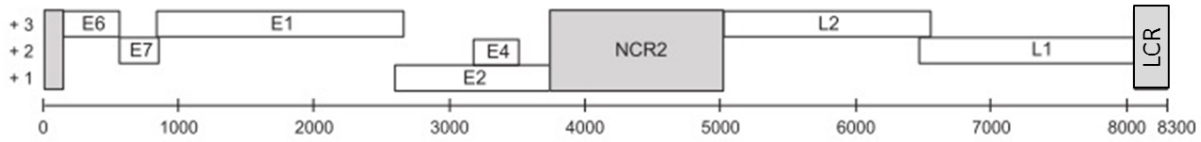
Papilloomiviiruste genoomis leidub kuus varajase valgu (E1, E2, E4, E5, E6, E7) ORF-i ning kaks hilise (L1 ja L2) ORF-i ning ülesvoolu paiknev regulaatorala ehk URR (*upstream regulatory region*) ehk LCR (*long control region*) ehk NCR (*non-coding region*) (joonis 3). LCR regioonis asub ka replikatsiooni alguspunkt (Ori). Transkriptsioon toimub hetketeadmiste põhjal ainult ühe ahela pealt, kuid lugemisraamid asuvad erinevates raamides. PV genoomi transkriptid on polütsistroonsed ehk omavad rohkem kui ühte ORF-i. Variatsioon luuakse alternatiivse RNA splaissimisega. Varajaste geenide ekspressiooni juhib varajane promootor, mis asub LCR regioonis (joonis 3) ja kontrollib peamiselt E6 ja E7 ekspressiooni. Hilise promootori ülesanne on reguleerida kapsiidivalkude (L1 ja L2) ekspressiooni viiruse elutsükli lõppfaasis (joonis 3) (Sankovski et al., 2014). Papilloomiviiruste genoome arhitektuur on üldjuhul tugevalt konserveerunud, kuigi peremeesorganismide mitmekesisus on lai. Samas 100% on esindatud ainult tuumikgeenid E1, E2, L2 ja L1 (Rector & Van Ranst, 2013).



Joonis 3. HPV16 genoomi struktuur. Punasega on tähistatud mittekodeeriv regulaatorala LCR, sinisega varajased geenid E1-E2 ja E4-E7 ning rohelisega hilised geenid. Lisaks on ära märgitud kaks promootorit - varajane P97 ning hiline P670 (Rautava & Syrjänen, 2012).

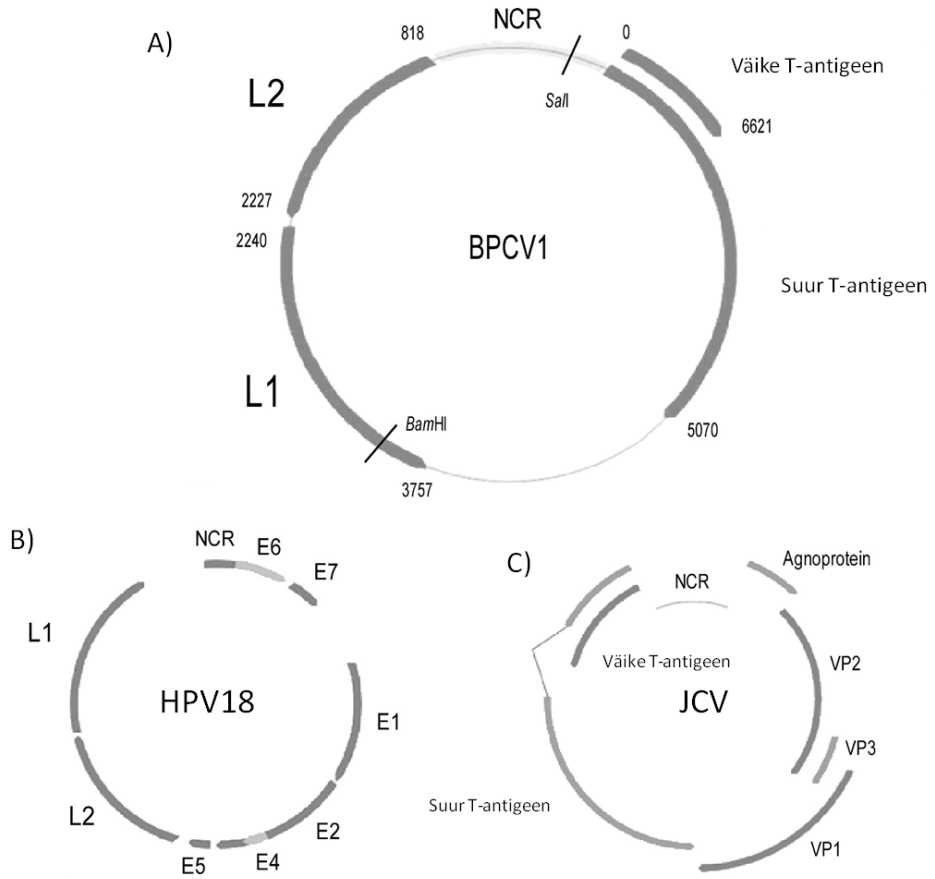
1.4.1. Kõrvalekalded klassikalisest PV genoomi struktuurist

Suuremaid kõrvalekaldeid genoomi struktuuris on täheldatud karnivooride papilloomiviiruste puhul (lambdapapilloomiviirus), kus esineb teine mittekodeeriv ala ehk NCR2 lisaks tüüpilisele LCR-le (joonis 4). Antud regioonis ei ole tuvastatud E2 seondumisjärjestust ning samuti ei ole leitud ühtegi promootorjärjestust või regulatsiooniga seotud elementi. Spekuleeritakse, et NCR2 regioon on tekkinud lambdapapilloomiviiruste ühisel eellasel, kas tundmatu funktsiooni ja päritoluga DNA lõigu integreerumisel viiruse genoomi või mingi genoomi ala duplikatsiooni käigus, mis on kaotanud oma funktsiooni. Viimane versioon on üpriski ebatõenäoline, sest ei ole leitud sarnasust ühegi teise PV regiooniga (Rector & Van Ranst, 2013).



Joonis 4: Kodukassi papilloomiviiruse (FcaPV1) genoome ülesehitus lineaarsel kujul. Lisaks klassikalisele LCR regioonile leidub neil veel teine mittekodeeriv ala NCR2, mis asub varajase ja hilise regiooni vahel. (Rector & Van Ranst, 2013).

Leitud on ka hübriide papilloomiviiruse ja polüoomiviiruse vahel. Austraalias elavalt Lääne-kukkurloomalt (*Perameles bougainville*) leiti viirus BPCV1 (*Bandicoot Papillomatosis and Carcinomatosis Virus*), millel olid nii polüoomiviiruse kui ka papilloomiviiruse tunnused. Viirusel on olemas PV-de hiline regioon, kodeerides kanoonilisi L1 ja L2 valku, kuid ka varajane regioon, mis sisaldab suurt ja väikest t-antigeeni, mis on omane polüoomiviirustele (joonis 5) (Woolford et al., 2007). BPCV1 L1 ORF omab 60% sarnasust DNA tasemel talle kõige lähema papilloomiviirusega BpPV1 (*Bettongia penicillata papillomavirus* e pintselsabalise bettongi PV), mis kuulub dyokappapapilloomiviiruste perekonda. BpPV1 puhul on tegemist ühe ainukese kukkurloomade papilloomiviirusega, mis seni kirjeldatud. BPCV1 ei ole ainuke selline hübriidviirus, tuvastatud on ka teine - BPCV2 (Bennett et al., 2008). Arvestades nukleotiidset sarnasust ja L1 keskmist mutatsioonikiirust lambdapapilloomiviiruste seas, siis on leitud, et nende ühine eellane eksisteeris ligikaudu 34,4 miljonit aastat tagasi. Praeguse hüpoteesi järgi toimus sellal kahe viiruse vaheline rekombinatsioon ühises peremeesorganismis (Rector & Van Ranst, 2013).



Joonis 5. BPCV1 genoomne struktuur. (A) BPCV1 viiruse genoom omab papilloomiviiruse hilist regiooni, kodeerides L1 ja L2 valku, kuid sealjuures ka polüoomiviiruse varajast regiooni, kodeerides suurt ja väikest t-antigeeni. **(B)** Klassikaline papilloomiviiruse genoome struktuur (HPV18). **(C)** Polüoomiviiruse genoome ülesehitus (JCV) (Woolford et al., 2007).

1.4.2. Papilloomiviiruse genoomi evolutsioon

Papilloomiviiruste evolutsioon toimub suhteliselt aeglaselt ning peamiselt punktmutatsioonide kuhjudes (Rector & Van Ranst, 2013). Arvatakse, et papilloomiviirused koevolutsioneeruvad koos peremeesorganismidega (Vande Pol & Klingelhutz, 2013). Samas on näidatud hulgaliselt vastuolusid fülogeneetiliste puude vahel, mis on koostatud viiruse eri geenidest - E1, E2, E6, E7, L1 või L2. Lisaks on leitud erinevusi ka peremeesorganismide ja viiruse fülogeneesi vahel, mis on vastuolus kodivergeerumise põhimõttega. Sellised leiud näitavad, et ajalooliselt on toimunud papilloomiviiruste eellaste vahel rekombinatsioone, kuid nende ulatus on hetkel veel ebaselge (Shah et al., 2010).

Papilloomiviiruste evolutsiooniline kiiruse määramiseks võeti vaatluse alla lambdapapilloomiviirused. Nende puhul leiti, et kogu viiruse kodeeriva ala mutatsioonide

kiirus on ligikaudu $1{,}95\,x10^{-8}$ nukleotiidset asendust ühe saidi kohta aastas (Rector et al., 2007). Võrreldes imetajate genoomiga, kus mutatsioonikiirus on ligikaudu $2{,}2\,x10^{-9}$ punktmutatsiooni aastas, evolutsioneeruvad papilloomiviirused ligikaudu 8,6 korda kiiremini (Kumar & Subramanian, 2002). Lambdapapilloomiviirustel varieeruvad mutatsioonikiirused erinevates geenides, näiteks E1 valgus on see $1{,}76\,x10^{-8}$, kuid E2 valgus $2{,}11\,x10^{-8}$ ning L1 valgus $1{,}84\,x10^{-8}$ nukleotiidset asendust ühe saidi kohta aastas (Rector et al., 2007).

1.5. Kirjeldatud geenid

1.5.1. E1 valk

Ainuke ensüüm, mida papilloomiviirused kodeerivad, on heksameerne ATP-sõltuv DNA helikaas E1, mis kuulub SF3 helikaaside perekonda. Tegemist on äärmiselt konserveerunud valguga papilloomiviiruste hulgas ja see peegeldab tema funktsiooni olulisust ning tähtsust. E1 on vajalik kogu PV elutsükli jooksul: esmalt infektsiooni alguses basaalkihi keratinotsüütides genoomi arvu suurendamiseks, seejärel episoomide arvu konstantsena hoidmisel rakkude jagunemisel ning lõpuks viraalse genoomi amplifikatsioonis produktiivses faasis. Funktsiooni täitmiseks seondub E1 viiruse genoomis replikatsiooni alguspunkti (Ori-sse) ja moodustab topelt heksameerse struktuuri. Enamikel papilloomiviirusel asub Ori LCR regioonis ning hõlmab tüüpiliselt kahte kuni kolme E2 ja palindroomset E1 seondumissaiti. E1 interakteerub paljude rakus leiduvate faktoritega, et juhtida replisoomi kokkupanekut, mis on vajalik viiruse genoomi kahesuunaliseks replikatsiooniks. E1 on rangelt reguleeritud post-translatsiooniliste modifikatsioonidega, et vältida tema kuhjumist tuuma (Bergvall, Melendy, & Archambault, 2013).

E1 ORF on kõige pikem ja konserveerunuim lugemisraam papilloomiviirustel ning omab ülekatet E2 valguga C-terminaalses osas. Näiteks HPV16 (NC_001526) puhul asub E1 ORF piirkonnas 865..2813 ning E2 ORF 2755..3852, mis teeb ülekattuvaks ala suuruseks 58 nukleotiidi. BPV1-el (NC_001522) asub E1 ORF piirkonnas 849..2666 ning E2 2608..3840, mis teeb ülekatteks samuti 58 nukleotiidi. E1 valgu pikkus jääb vahemikku 600-650 aminohapet, olenevalt PV tüübist. Üldiselt saab valgu jaotada kolmeks funktsionaalseks segmendiks: N-terminaalne regulaatorala, DBD ehk DNA-ga seondumise domeen ja C-terminaalne ensümaatiline domeen (joonis 7A). N-terminaalne regulaatorala sisaldab mitmeid erinevaid motiive, näiteks NLS (tuuma lokalisatsiooni signaal), NES (Crm1-sõltuv tuuma ekspordi signaal). DBD regiooni funktsiooniks on replikatsiooni alguspunkti äratundmine. C-terminaalne osa on helikaas. E1 kaasab oma funktsiooni täitmises ka mitmeid rakulisi faktoreid, mille hulka kuuluvad DNA polümeraasi α-primaasi kompleks, replikatsiooni valk A ja topoisomeraas I (Bergvall et al., 2013).

1.5.2. E2 valk

Tuumavalk E2 on papilloomiviiruse geenide tsentraalne ekspressiooni koordinaator ning osaleb lisaks genoomi replikatsioonile ka selle säilitamises, transkriptsiooni aktivatsioonis ning repressioonis (Kurg, et al., 2005). Üheks E2 ülesandeks on transportida E1 valk Ori lähedusse, seda võimaldab tema võime üheaegselt siduda nii E1 valku kui ka seonduda viiruse genoomi LCR regiooniga. E1-E2-Ori kompleks initsieerib heksameerse E1 moodustumise ja replikatsiooni alguse (Bergvall et al., 2013).

E2 koosneb kahest erinevast domeenist: C-terminaalne dimerisatsiooni ja DNA-ga seondumise domeen (DBD) ning N-terminaalne transaktivatsiooni domeen (TAD) (joonis 7B). Mõlemat domeeni ühendab varieeruv *hinge* regioon. DBD domeeni ülesanne on seonduda E2 seondumissaidile Ori piirkonnas. Seondumine toimub järjestusspetsiifiliselt konsensusele $ACCN_6GGT$ (Lambert et al., 1989). TAD eesmärk on reguleerida viiruse geenide ekspressiooni ning antud domeeni kaudu toimub ka E1-ga seondumine (Bergvall et al., 2013). E2 valgu eluiga rakus on üldjuhul lühike, seda reguleerib ubikvitiin-sõltuv valkude degradatsiooni rada. E2 ubikvitineerimine toimub peamiselt tsütoplasmas. E2 stabiilsust mõjutavad veel fosforüleerimine ning ka sumolüülimine (Li et al., 2014).

Papilloomiviiruse elutsükli oluline osa on säilitada genoom tuumas ekstrakromosomaalse elemendina ning selles mängib suurt rolli E2 valk. E2 on võimeline seonduma Brd4 (bromodomeeni sisaldav valk 4) valguga, seondumine toimub E2 valgu N-terminaalse domeeni ja Brd4 C-terminaalse domeeni kaudu. Brd4 on seondub histoonide H3 ja H4 atsetüleeritud sabadega, seega võimeline kinnituma kromatiinile. Tänu antud interaktsioonile (peremeesorganismi DNA→ histoon→ Brd4→ E2→ viiruse genoom) säilib papilloomiviiruse genoom episoomina tuumas. Paljud hiljutised uuringud on näidanud, et Brd4 on oluline ka E2 stabiliseerimises, sest Brd4 üleekspressioon inhibeerib E2 lagundamist, kuna seondumine Brd4-ga takistab arvatavasti E2 eksporti tuumast (Li et al., 2014). Lisaks on E2 võimeline interakteeruma veel mitmete teiste rakuliste valkudega, näiteks Sp1, TBP, TFIIB, TFIID, AMF1, p300/CBP, p/CAF ja hNAP1(Kurg et al., 2005; Lambert, 1995).

1.5.3. E4 valk

Viiruse produktiivse faasi initsieerimine, mille alla kuulub nii genoomi amplifikatsioon kui ka virionide moodustumine, algab E4 ekspressiooniga ning selle kõrge tase rakus püsib kogu produktiivse faasi. E4 geeni ekspressioonimuster on sarnane pigem hiliste geenidega (joonis 2) (Wang & Roden, 2013; Wilson, Ryan, Knight, Laimins, & Roberts, 2007). Viiruse genoomis on küll olemas E4 lugemisraam, kuid eraldiseisvana seda valku ei esine. E4 valk saadakse splaissitud transkriptist E1^E4, see tähendab, et esimesed ~5 aminohapet on pärit E1 valgu algusest ning lõpp E4 lugemisraamist. Täpne E1^E4 funktsioon ei ole teada, kuid selle kõrge ekspressioon põhjustab raku G2 aresti. Kuna aga G2 → M ülemineku inhibeerimine stimuleerib raku mitte-replikatiivset DNA sünteesi, siis tänu sellele on viirusele genoomi amplifikatsiooniks vajaminevad faktorid kättesaadavad. Antud faktist saab järeldada, et E1^E4 funktsiooniks on viiruse genoomi amplifikatsiooni tagamine, samas ei ole see oluline genoomi säilitamises. Lisaks on teada, et E1^E4 indutseerib apoptoosi, mis on vajalik viiruse elutsükli hilisemas faasis (Wilson et al., 2007).

1.5.4. E5 valk

E5 on lühike transmembraane valk, mis on oluline raku transformatsioonis. Tegemist on onkogeeniga, kuid seda ei leidu kõikidel papilloomiviirustel (DiMaio & Petti, 2013; Rector & Van Ranst, 2013). E5 on võimeline transformeerima nii peremeesorganismi kui ka koekultuuri rakke. Valgu suurus jääb vahemikku 40 – 85 aminohapet, sisaldades palju hüdrofoobseid aminohappeid. Suure tõenäosusega puudub E5 valgul ensümaatiline aktiivsus, ta pigem moduleerib teisi rakulisi valke. Detailsemalt on uuritud BPV1 ja HPV16 E5 valku. Nende puhul on näidatud interaktsiooni PDGF β (inimese trombotsüütide kasvufaktor beeta subühik) ja EGF (epidermaalne kasvufaktor) retseptoritega, lisaks MHC (peamine koesobivuskompleks) klass I ja II retseptoritega, tsükliinsõltuvate kinaaside inhibiitoritega (p21 ja p27) ja NFκB-ga (raku tuuma faktor kappa). Huvitav on asjaolu, et kuigi HPV16 ja BPV1 E5 on mõlemad hüdrofoobsed, ei ole nad järjestuselt sarnased (DiMaio & Petti, 2013).

1.5.5. E6 valk

E6 puhul on samuti tegemist rakke transformeeriva valguga ning seda tüüpiliselt ekspresseeritakse varajaselt promootorilt. E6 suudab interakteeruda paljude rakuliste valkudega, sealhulgas ka p53-ga (tuumor-supressorvalk), viies selle lagundamisele. Lisaks on kõrge riskiga HPV-de E6 valgud võimelised aktiveerima telomeraasi. Põhjus võib olla keratinotsüütide eluea pikendamises, mis annab kõrge riskiga HPV-dele eelise. Uuringud on näidanud, et E6 on vajalik ka viiruse genoomi replikatsioonis - muteerides HPV16 E6-te kaotas antud viirus võime hoida stabiilset genoomi koopiaarvu (Vande Pol & Klingelhutz, 2013).

E6 puudumist on täheldatud HPV-de hulgas (HPV101, HPV103 ja HPV108 - gammapapilloomiviirused) ning veisepapilloomiviirustes (BPV3, BPV4, BPV6, BPV9, BPV10, BPV11 ja BPV12). Lisaks ei eksisteeri E6 ega ka E7 ORF-i kolmel lindude papilloomiviirustel: PePV1 (*Psittacus erithacus papillomavirus* e Aafrika halli papagoi PV) ja FcPV1 (*Fringilla coelebs papillomavirus* e metsvindi PV) (de Villiers, 2013; Rector & Van Ranst, 2013; Van Doorslaer et al., 2009).

1.5.6. E7 valk

Rakkude immortaliseerimiseks ei piisa ainult eelpoolkirjeldatud E5 ja E6 valgust, vaid on vaja ka E7 olemasolu (Roman & Munger, 2013; Vande Pol & Klingelhutz, 2013). E7 omab keskset rolli PV elutsüklis, reprogrammeerides raku keskkonda, et see oleks viirusele sobilik. E7 valk on ligikaudu 100 aminohappe suurune ning ta ei sarnane ühegi rakulise valguga. E7 valku on tuvastatud nii tsütoplasmas kui ka tuumas - omades nii NLS kui ka NES signaali. E7 valk on samuti võimeline interakteeruma paljude rakuliste valkudega (pRB, BRCA1, p21 jne.) ning seeläbi mõjutama nende stabiilsust (Roman & Munger, 2013).

E7 ORF-i ei leidu kõikidel papilloomiviirustel, selle puudumine on ühine tunnus vaalaliste PV-del: TtPV1, TtPV2, TtPV3, TtPV4, TtPV7, DdPv1, PhpPV2 (upsilonpapiloomiviirused), PsPV1, PhpPV1, TtPV5, TtPV6 (omikronpapilloomiviirused) ning PhpPV4 (dyopipapilloomiviirus). Lisaks vaalalistele, ei ole leitud kanoonilist E7 ORF-i veel omegapapilloomiviiruselt UmPV1 (*Ursus maritimus Papillomavirus* e jääkaru PV), ühelt

dyodeltapapilloomiviiruselt SsPV1 (*Sus scrofa domestica papillomavirus* e kodusea PV) ja ühelt klassifitseerimata viiruselt MrPV1 (*Myotis ricketti papillomavirus* e nahkhiire PV) ning eespool mainitud lindude papilloomiviirustelt (Rector & Van Ranst, 2013; Stevens, Rector, Bertelsen, Leifsson, & Van Ranst, 2008; Stevens, Rector, Van Der Kroght, & Van Ranst, 2008).

1.5.7. L1 ja L2 valk

Papilloomiviiruste ikosaeedrilise virioni diameeter jääb vahemikku 55-60 nm ning sisaldab 360 L1 ning kuni 72 L2 valku (Wang & Roden, 2013). L1 kui ka L2 on mõlemad ligikaudu 55kDa suurused, sealjuures on L1 võimeline iseseisvalt moodustama viiruse sarnaseid partikleid (VLP) ilma igasuguste *chaperon*-ide juuresolekuta, kuid L2 mitte (Buck et al., 2013; Wang & Roden, 2013). L1 vastutab ka infektsiooni esimese etapi eest – interaktsioon peremeesorganismi rakuga. Seondumine toimub heparaan sulfaat (HS) karbohüdraatide kaudu, mida leidub proteoglükaanides. Antud interaktsioon põhjustab kapsiidis konformatsioonilise muutuse, paljastades L2 valgu N-terminaalne osa. Järgnevalt toimub L2 lõikamine ning järjekordne konformatsiooniline muutus ning interaktsioon sekundaarse retseptoriga. Hetkel ei teata täpselt, mis on sekundaarne retseptor, kuid kahtlusaluseks on alfa-6-beeta-4 integriin (α6β4). Samuti ei ole hetkel veel teada ka täpsed mehhanismid, kuidas toimub endotsütoos rakku (Buck et al., 2013).

1.6. E8^E2 iseloomustus

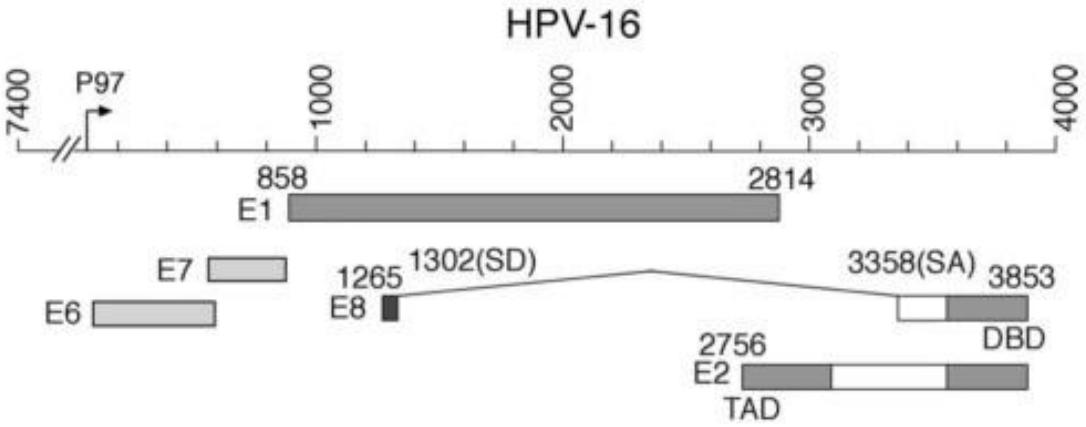
BPV1 ning erinevate HPV-de transkriptide analüüsid on näidanud, et lisaks kanoonilisele E2 valgu transkriptile eksisteerib ka E8^E2C splaissitud variant (joonis 6). Antud transkript erineb kanoonilisest E2-st N-terminaalse osa poolest (Fertey et al., 2011; Lambert et al., 1989). E8^E2C tekib splaissingu käigus, kui E1 lugemisraamis olev E8 CDS (joonis 7A) liidetakse E2 lugemisraami C-terminaalse osaga *hinge* regioonis (joonis 7B). E8 CDS-i all mõeldakse E8 kodeerivat ala ATG-st kuni splaissingu doonorsaidi G-nukleotiidini ehk viimase nukleotiidini, mis jääb E8 eksonisse. E2 ORF-is olevat splaissingu aktseptorsaiti kasutab ka papilloomiviirus E1^E4 transkripti konstrueerimiseks. Hetkeseisuga on E8^E2C transkripti leitud eksperimentaalselt: alfapapilloomiviirustes (HPV11, HPV16, HPV18, HPV31, HPV33), müüpapilloomiviiruses (HPV1a), kappapapilloomiviiruses (CRPV ehk SfPV1), deltapapilloomiviiruses (BPV1) ja beetapapilloomiviiruses (HPV5) (Chiang, Broker, & Chow, 1991; Choe et al., 1989; Hubbert et al., 1988; Kurg et al., 2010; Lace, et al., 2008; Lambert et al., 1989; Palermo-Dilts, Broker, & Chow, 1990; Sankovski et al., 2014; Snijders et al., 1992; Stubenrauch et al., 2000). E8^E2C on tuvastatud taksonoomiliselt kaugetes perekondades alfa- ja deltapapilloomiviirustes (Shah et al., 2010). Eeldusel, et E8 valgud on homoloogid (pärit ühisest eellasjärjestusest), siis võiks neid leiduda ka teistes suuremates perekondades. Deltapapilloomiviiruste ja alfapapilloomiviiruste, kellel on tuvastatud E8^E2C olemasolu, ühine eellane eksisteeris ~150 miljonit aastat tagasi (E1 järgi). Sealjuures näiteks lambdapapilloomiviirused lahknesid hiljem ~135 miljonit aastat tagasi (E1 järgi), seega võiks eeldada, et ka nendel eksisteerib E8^E2C (Shah et al., 2010).

E8^E2C valk on viiruse genoomi replikatsiooni negatiivne regulaator ning ka varajaste geenide transkriptsiooni repressor. E8^E2C mõju replikatsioonile võib põhjustada asjaolu, et E8^E2C omab funktsionaalset E2 valgu DBD domeen. Kuna DBD domeen ülesanne on seonduda E2 seondumissaidile LCR-regioonis, siis ta konkureerib kanoonilise E2 valguga, mis on vajalik E1 valgu kohaletoomiseks ja genoomi replikatsiooniks. HPV16 puhul on täpsemalt teada, et defektne E8^E2C viib viiruse genoomi arvu 8-15x tõusuni. Kuigi E8^E2C limiteerib HPV16-s viiruse genoomide arvu, ei ole seda vaja genoomi säilitamiseks rakus. Huvitava faktina on teada, et E8^E2C ekspressioon on sõltumatu varajasest promootorist P97, seega selle ekspressioon on teistsugune võrreldes varajaste geenidega (Lace et al., 2008).

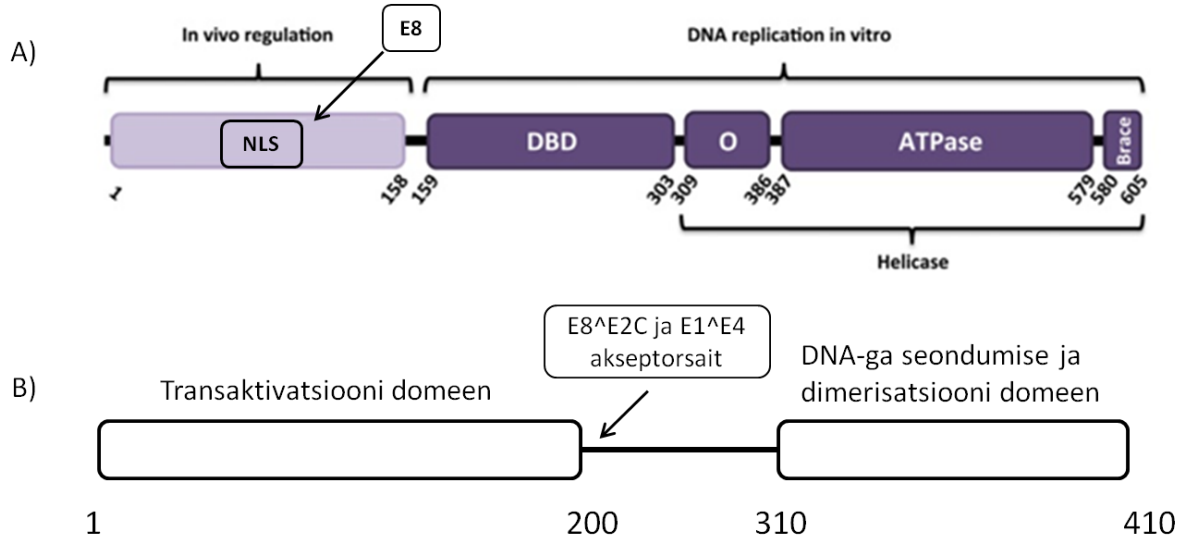
E8^E2C on ka viiruse varajaste geenide transkriptsiooni repressor. BPV1 puhul on näidatud, et E8^E2C surub alla E2 vahendatud geenide ekspressiooni 5-10 korda ning sarnast tulemust

on näidatud ka HPV16 korral (Lace et al., 2008; Lambert et al., 1990). E8^E2C suudab inhibeerida HeLa rakkude (emakakaelavähi rakuliin HPV18) kasvu, olles E6 ja E7 onkogeenide promootori inhibiitor (Fertey et al., 2011). Inhibeerimine toimub konkurentse seondumise tõttu promootorile – rakulised faktorid *versus* E8^E2C promootoralas või interaktsiooni tõttu erinevate rakuliste modulaatoritega (Fertey et al., 2010). Teadaolevalt on E2 valk transkriptsiooni regulaator. E2 valgus täidab transkriptsiooni regulatiivset rolli N-terminaalne TAD domeen, kuid E8^E2C-l see puudub. Seega peab E8^E2C kasutama teisi radu viiruse geenide represseerimiseks. E8^E2C võime represseerida viiruse geene kaob, kui inhibeerida klass I HDAC-d (histoon deatsetülaas), seega võib eeldada, et E8 regulatiivsed omadused on seotud klass I HDAC-ga. Täpsemalt on teada, et E8^E2C valgu E8 osal on võime interakteeruda HDAC3-ga, mis on põhiline komponent NCoR1 (*nuclear receptor corepressor 1*) transkriptsiooni vaigistavas kompleksis (Powell et al., 2010). E8^E2C on võimeline interakteeruma ka CHD6 valguga (näidatud HPV16, HPV18, HPV31 peal), mis kuulub rakulise helikaasi kompleksi. Antud juhul toimub interaktsioon E2C domeeni kaudu, seega on ka E2 võimeline interakteeruma CHD6-ga. Seondumine CHD6-ga on oluline onkogeenide E6 ja E7 promootori inhibeerimiseks (Fertey et al., 2010).

Pikka aega puudus eksperimentaalne info E8^E2C olemasolu kohta beetapapilloomiviirustes, sest ei eksisteerinud häid katsesüsteeme, kus beetapapilloomiviirused oleks replitseerunud. 2014. aastal demonstreeris Mart Ustavi uurimisgrupp, et inimese osteosarkooma koekultuuriliini (U2OS) on võimalik nakatada beetapapilloomiviirusega ning seal toimub ka viiruse genoomi replikatsioon. Nad leidsid, et E1 lugemisraamis, vahetult enne E8 ORF-i, asub seni kirjeldamata promootor. See kinnitab ka eelnevalt mainitud fakti, et HPV16 puhul on E8^E2C ekspressioon sõltumatu varajasest promootorist. Samuti näidati, et sarnaselt HPV16-le ja HPV18-le on HPV5 E8^E2C valk genoomi replikatsiooni repressor, viies alla E1 mRNA transkriptsiooni. HPV5 puhul nähti 120 tunni jooksul kuni 100x erinevust wt (*wild type*) ja defektse E8^E2C ORF-iga viiruste genoomide arvus. HPV5 E8^E2C valk omab lühikest 10-aa järjestust E8 ORF-ist (1328nt – 1359nt), mis on liidetud E2 valgu C-terminaalse osaga 315-aa (3322nt splaisingu aktseptorsait) (Sankovski et al., 2014).



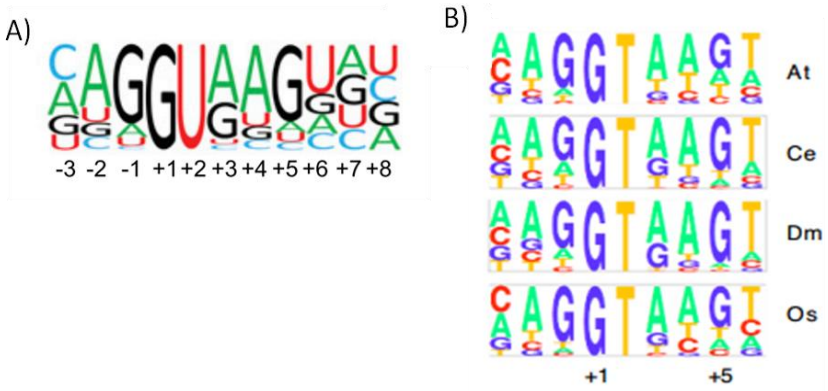
Joonis 6. HPV16 E8^E2C transkript. Joonisel on näidatud E8^E2 transkript ning selle asukoht E1, E2, E6 ja E7 lugemisraamide suhtes. E8^E2 saadakse splaissingu teel, kui E1 lugemisraamis olev lühike 31-40 nukleotiidi pikkune E8 järjestus splaissitakse E2C-terminaalse osaga kokku *hinge* regioonis asuva splaissingu aktseptorsaidi kaudu. Sama aktseptorsaiti kasutab ka E1^E4 (Lace et al., 2008).



Joonis 7. BPV1 E1 ja E2 valgu ülesehitus ning domeenid. Joonistel olevad numbrid tähistavad aminohappeid. **(A)** E1 valk koosneb viiest domeenist: N-terminaalne regulatsiooni domeen, DNA seondumisdomeen (DBD), oligomerisatsiooni domeen (O), ATP seondumise ja hüdrolüüsimise domeen (ATPase) ja klamber ehk piirkond, mis on oluline heksameeride kokkupanekul ja nende stabiliseerimisel (*Brace*). Domeenid, mis on vajalikud *in vitro* viiruse genoomi replikasiooniks, on tähistatud tumelillaga. Noolega märgitud E8 ORF-i asukoht E1 valgus. **(B)** E2 valk koosneb kahest domeenist: N-terminaalsest transaktivatsiooni domeenist ja C-terminaalsest DNA-ga seondumise ja dimerisatsiooni domeenist ning neid ühendab *hinge* regioon. Noolega märgitud E8^E2 splaissingu aktseptorsait *hinge* regiooni alguses, mida kasutab ühtlasi ka E1^E4 (Bergvall et al., 2013; Kurg et al., 2009).

1.7. Topeltkodeerivate alade tuvastamine *in silico*

Valgugeenide ennustamiseks kasutatavad meetodid *in silico* omavad tavaliselt kahte aspekti: informatsioon, mida kasutatakse ennustamisel ja matemaatiline algoritm, mis kombineerib selle. Informatsiooni saab jagada kolmeks: signaalid genoomis, koodonkasutus ning homoloogia. Signaalide alla kuuluvad start- ja stoppkoodon, transkriptsioonifaktorite seondumissaidid, *TATA box* promootoralad, CpG saared, cap-signaal, polüadenülatsiooni signaal, splaissingu aktseptor- ja doonorsait ning GC sisaldus (Sleator, 2010; Stormo, 2000). Topeltkodeerivate alade ennustamiseks ei saa kõiki signaale kasutada, sest sisemised lugemisraamid võivad rakendada primaarse raami elemente, nagu näiteks cap-signaal või polüadenülatsiooni signaal. Üheks suureks probleemiks, signaalide ennustamisel, on tõsiasi, et konsensused ei ole üldjuhul tugevalt konserveerunud. Vaadates näiteks inimese splaisingu doonorsaidi konsensust (joonis 8A) on näha, et konserveerunud on ainult nukleotiidid positsioonidel +1 G ja +2 U. Sama on näha ka müürlooga, hariliku äädikakärbse, ümarussi ning ka riisi genoomis (joonis 8B) (Korf, 2004; Roca, Krainer, & Eperon, 2013).



Joonis 8. Splaissingu doonorsaidi konsensus. Iga tähe kõrgus vastab selle nukleotiidi esinemissagedusele. Splaissosoomis toimub lõikamine -1 G ja +1 G nukleotiidi vahelt, sealjuures alates +1 G jääb intronisse. **(A)** Inimese doonorsaid konsensus on leitud 201541 inimese splaisssaidi põhjal. Jooniselt on näha, et täielikult on konserveerunud ainult guanosiin positsioonis +1 ja uridiin positsioonis +2, kõikides teistes positsioonides võib-olla mistahes nukleotiid. **(B)** Splaissingu doonorsaidi konsensus erinevates organismides. *Arabidopsis thaliana* (At), *Caenorhabditis elegans* (Ce), *Drosophila melanogaster* (Dm), *Oryza sativa* (Os). Ka antud juhul on näha, et täielikult konserveerunud ainult nukleotiidid positsioonides +1 G ja +2 U(T) (Korf, 2004; Roca et al., 2013).

Topeltkodeerivate alade ennustamiseks saab kasutada koodonite esinemise sagedusi, sest sünonüümsete koodonite kasutamise sagedused varieeruvad nii organismide kui valku kodeerivate geenide vahel (Camiolo, Farina, & Porceddu, 2012; Plotkin & Kudla, 2011). Eksisteerib kaks põhilist mudelit selle fenomeni seletamiseks. Neutraalne mudel väidab, et koodonkasutatuse kallutatus tekib lokaalsete mutatsioonide tõttu ning ei ole seotud mitte

millegi muuga. Selektiivne mudel väidab, et toimub koadapteerumine tRNA hulga suhtes, optimeerimaks translatsiooni efektiivsust ja täpsust (Plotkin & Kudla, 2011). Pagaripärmis on näidatud, et tRNA geenide arvu ning selle hulga vahel rakus on tugev korrelatsioon (Camiolo et al., 2012; Duret, 2002). Sünonüümsed asendused mõjutavad lisaks translatsioonile ka RNA protsessimist ning valgu pakkimist (Plotkin & Kudla, 2011) ning on näidatud, et koodonkasutus võib olla seotud ka geeni funktsiooniga – lähedase funktsiooniga geenid omavad sarnast koodonkasutust (Najafabadi, Goodarzi, & Salavati, 2009). Seega teades geeni koodonkasutust, on võimalik seda rakendada sisemiste lugemisraamide ennustamiseks, sest antud regioonis on nihkes geeni üldine koodonkasutus.

Kodeerivate alade ennustamiseks on võimalik ära kasutada ka järjestuste vahelist homoloogiat. Leides uuritavale järjestusele sarnase annoteeritud järjestuse, võimaldab see ennustada, kas tegemist on kodeeriva alaga või mitte ning annab infot ka funktsiooni kohta. Antud lähenemise probleemiks on andmebaasides leiduva info piiratus. Samuti on probleemiks lühikeste järjestuste otsimine, sest juhuslikult võib andmebaasist leida vaste (Stormo, 2000; Sleator, 2010). Topeltkodeerivate alade ennustamiseks viirustes on loodud võrdlemisi vähe programme, ühe näitena võib tuua MLOGD, kuid see ei ole võimeline leidma alasid, mis on lühemad kui 20 koodonit. (Firth & Brown, 2006). MLOGD on kättesaadaval kodulehelt [http://guinevere.otago.ac.nz/aef/MLOGD/]

2. EKSPERMENTAALOSA

2.1. Töö eesmärgid

E8^E2C on tuvastatud taksonoomiliselt kaugetes perekondades - alfapapilloomiviirustes ja deltapapilloomiviirustes, kelle ühine eellane eksisteeris ~150 miljonit aastat tagasi (Shah et al., 2010). Kuid paljud teised PV perekonnad lahknesid väiksemas ajalises sügavuses. Sealjuures näiteks lambdapapilloomiviirused lahknesid ülejäänutest ~110 miljonit aastat tagasi, kuid nende genoomis leidub omapärane NCR2, mida teistes ei eksisteeri (Rector & Van Ranst, 2013; Shah et al., 2010). Seega võiks eeldada, et ka teistes perekondades eksisteerib E8^E2C, kuid evolutsiooniliselt võib olla toimunud muutusi, nagu nägime lambdapapilloomiviiruste puhul.

Eesmärk nr 1: Uurida *in silico* E8 valku kodeeriva ala olemasolu papilloomiviiruste perekondades, lähtudes senikirjeldatud infost.

Eesmärk nr 2: Konstrueerida meetod hindamaks, kas DNA-s olev muster, näiteks E8 lugemisraam, püsib E1 valgulise järjestuse konserveerumuse tõttu või konserveerumus on tekkinud DNA tasemel ning on oluline viirusele muudel põhjustel kui E1 kodeerimine.

2.2. Materjal ja metoodika

2.2.1. E8 lugemisraami otsimine teadaoleva info põhjal

E8^E2C transkripti on eksperimentaalselt leitud üheksas papilloomiviiruses, teadaolevate andmete põhjal hangiti esialgne baasinformatsioon (tabel 1). Tabelis 1 on märgitud ka E1^E4 splaissingu aktseptorsait (3' ss), sest E8^E2C ja E1^E4 esimesed eksonid splaisitakse samasse doonorsaiti. Splaissingu atseptorsaidi asukoha saamiseks kasutati PaVe *(papillomavirus Episteme*) andmebaasi, kuhu on koondatud käsitsi kontrollitud papilloomiviiruste genoomne info (Van Doorslaer et al., 2013).

Tabel 1. Eksperimentaalselt kinnitatud E8^E2C informatsioon.*

PV	E8 kaugus E1-st [1]	E8 ATG[2]	E8 5' ss[3]	E8 pikkus[4]	E8 ORF[5]	5' ss[6]	E1^E4 3'ss[7]	ID[8]	Takson[9]
SfPV1	358	1720	1751	32	+ 1	caggta	3843	NC_001541	10623
BPV1	355	1204	1235	32	+ 1	caggta	3225	X02346	10559
HPV1a	388	1200	1231	32	+ 1	caggta	3200	V01116	10583
HPV5	367	1328	1359	32	+ 1	gaggta	3322	M17463	333923
HPV11	409	1241	1272	32	+ 1	caggta	3325	M14119	10580
HPV16	400	1265	1302	38	+ 1	caggta	3358	K02718	333760
HPV18	409	1323	1357	35	+ 1	caggta	3434	X05015	333761
HPV31	397	1259	1296	38	+ 1	caggta	3295	J04353	10585
HPV33	397	1276	1316	41	+ 1	caggta	3351	M12732	10586

* Tabel koostatud PaVe andmebaasi ning järgnevate artiklite põhjal (Chiang et al., 1991; Choe et al., 1989; Hubbert et al., 1988; Kurg et al., 2010; Lace et al., 2008; Lambert et al., 1989; Palermo-Dilts et al., 1990; Sankovski et al., 2014; Snijders et al., 1992; Stubenrauch et al., 2000)

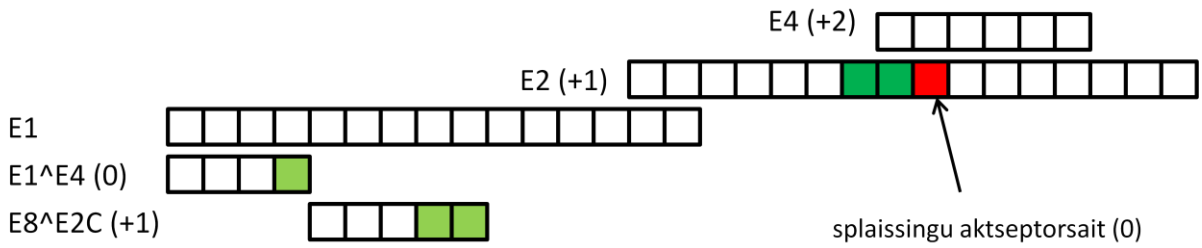
1. E8 lugemisraami ATG kaugus E1 ATG-st, arv on nukleotiidides
2. E8 ATG asukoht genoomis, A-nukleotiidi asukoht
3. E8 splaisingu doonorsaidi esimese G- asukoht genoomis (caGgta)
4. E8 CDS-i pikkus nukleotiidides
5. E8 lugemisraami E1 suhtes
6. Splaissingu doonorsait
7. E1^E4 splaissingu aktseptorsaidi asukoht (PaVe andmebaasi põhjal)
8. Genoomi ID
9. NCBI taksoni ID

Kogutud andmete põhjal leiti, et E8 ATG asub keskmiselt 386 nukleotiidi kaugusel E1 algusest, sealjuures minimaalne kaugus oli 355 ja maksimaalne 409 nukleotiidi. E8 CDS-i keskmine pikkus oli 34,6 nukleotiidi, sealjuures väikseim CDS on 32 ning suurim 41 nukleotiidi. Teadmata, millistel positsioonidel võib E8 asuda teistes PV tüüpides, otsustati

alustada E8 CDS-i otsimist 300 nt kauguselt E1 algusest. E8 CDS-i minimaalseks pikkuseks lubati 17 ja maksimaalselt 59 nukleotiidi, seda põhjusel, et mitte välistada pikemaid või lühemaid E8 CDS-e teistes tüüpides.

Kaheksal juhul üheksast on konserveerunud CAGGTA splaissingu doonorsait (5' ss), ainult HPV5-el on GAGGTA, mis on ainuke beetapapilloomiviirus (tabel 1). Inimese genoomi põhjal tehtud analüüsis on leitud, et splaissingu doonorsaidis on täielikult konserveerunud ainult kaks nukleotiidi (joonis 8A) ning on näha, et -3 positsioonis, kus leiti C või G, võib tegelikult peaaegu võrdselt olla kõiki nukleotiide. Lisaks on näidatud, et sarnane konsensus on levinud ka teistel organismidel (joonis 8B). Seega otsustati jätta splaissaidi otsingul välja -3 positsioon ning otsida rangeid splaissingu doonorsaidi konsensusi järjekorras AGGTA → AGGTG → AGGTT → AGGTC.

Kõik teadaolevad E8 ORF-id asuvad +1 lugemisraamis (lisa 4) E1 valgu suhtes (tabel 1). Eeldades, et E8 valgud on homoloogid, võiks arvata, et ka teistes papilloomiviirustes võib E8 ORF olla fikseerunud +1 lugemisraamis E1 valgu suhtes. Vaadates E1 ja E2 ORF-i, teame, et need on umbes 60 nukleotiidi ülekattes ning nad on omavahel fikseeritud eriraamides. E2 saab olla E1 suhtes +1 või +2 raamis, kuna aga papilloomiviiruste eellasel toimus see +1 raamis, siis see on evolutsiooniliselt fikseerunud. Samuti on E2 lugemisraamis fikseerunud E4 ORF +1 raamis E2 suhtes (ehk +2 E1 suhtes), kuid E1^E4 splaissingu aktseptorsait hoopis +2 raamis E2 suhtes, mida kasutab ka E8^E2C. Seda on näha ka E1^E4 DNA järjestuse E1 osa pikkusest, mis jagades 3-ga annab jäägiks 1 ehk panustab ühe nukleotiidi taastamaks E4 lugemisraam (joonis 9). Kuna evolutsiooniliselt on fikseerunud E1^E4 E1 CDS pikkus, E4 ORF-i asukoht, siis ei saa suure tõenäosusega muutuda ka E2 ORF-is asuv E1^E4 splaissingu aktseptorsaidi asukoht (raam), kuna see põhjustaks suuri ümberkorraldusi kogu viiruse genoomis. Sarnase järelduse võiks teha ka E8 CDS puhul, et see on fikseerunud +1 raamis E1 suhtes ning selle asukoha muutus võiks põhjustada suuri ümberkorraldusi. Samas teadmata kindlalt, kas see on ka reaalselt nii kõikides PV-des, vaadati pigem E8 CDS-i pikkust. Kokkuvõtvalt, kuna E1^E4 (ka E8^E2C) splaissingu aktseptorsait on E2 suhtes +2 raamis, siis peab E8 CDS panustama kaks nukleotiidi E2 lugemisraami taastamiseks ehk E8 pikkuse jagades 3-ga peab olema jäägiks 2.



Joonis 9. E1^E4 ja E8^E2C lugemisraamide asukohad E1 ORF-i suhtes skemaatiliselt. Kastid tähistavad nukleotiide, sulgudes märgitud number näitab lugemisraami asukohta E1 ORF- suhtes, sealjuures 0 tähistab sama raami E1-ga. Punasega on tähistatud E1^E4 ja E8^E2C splaissingu aktseptorsait. Tumerohelised ruudud koos punase ruuduga tähistavad E2 valgu lugemisraami koodonit. Helerohelisega on märgitud nukleotiidid, mis „antakse“ splaissingu käigus esimese eksoni viimase koodoni koosseisu.

Eelneval analüüsitud info põhjal kirjutati *Python*-is programmi *E8ORFSearch.py* , mis arvestab järgmisi piiranguid:

1. E8 CDS-i hakatakse otsima 300 nukleotiidi kauguselt E1 ORF-i algusest. (Võimalik rakendada ainult korrektselt annoteeritud N-terminusest E1 valgu puhul)
2. E8 CDS-i pikkus peab 3-ga jagades andma jäägiks 2.
3. E8 CDS-i otsing toimub 1000 nukleotiidi ulatuses. (akna suurus)
4. Splaissingu doonorsaiti vaadatakse järjekorras AGGTA→ AGGTG→ AGGTT→ AGGTC. (Esialgu otsitakse E8 CDS-i AGGTA splaissingu saidiga, kui ei leita, alles siis võetakse järgmine)
5. Minimaalne E8 CDS- i pikkus on 17 nukleotiidi ja maksimaalne 59 nukleotiidi.
6. E8 CDS ei tohi sisaldada stoppkoodonit.

2.2.2. PAL2NAL

Tegemist on veebis kättesaadava tööriistaga (http://www.bork.embl.de/pal2nal/), kuid on võimalik alla laadida ka *Perl*-i skripti ning jooksutada seda käsurealt. PAL2NAL-i eesmärk on konverteerida valgu mitme järjestuse joondus (MSA) vastavaks DNA (mRNA) joonduseks etteantud DNA järjestuste põhjal. Sisendiks on valgu MSA ning nende valgujärjestustele vastavad DNA järjestused *fasta* formaadis. Väljundiks on antud valgu joonduse põhjal genereeritud DNA joondus. (Suyama, Torrents, & Bork, 2006).

2.2.3. CDEP meetod

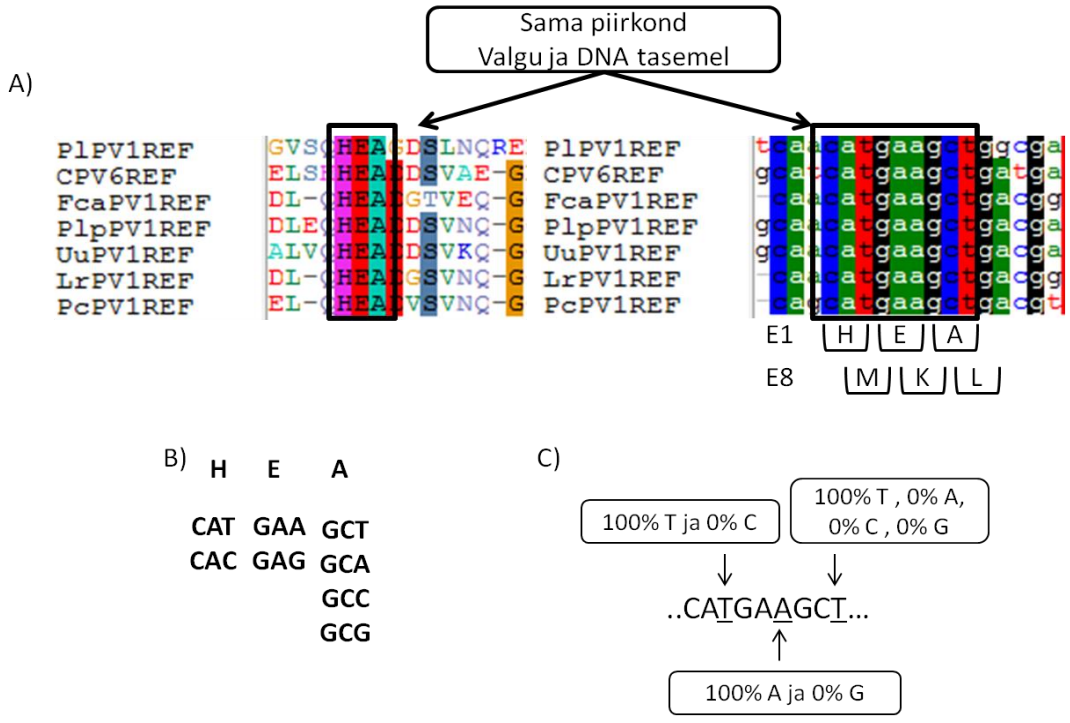
Valk oma funktsiooni säilitamiseks peab hoidma oma aminohappelist järjestust, see võib viia DNA tasemel juhuslikult mingi elemendi/konsensuse tekkeni (näiteks sisemine lugemisraam). Tekkinud konsensus kaob kiiresti evolutsiooni käigus, mutatsioonide tõttu, sest ühele aminohappele vastab mitu koodonit ja üldjuhul juhuslikult tekkinud element ei ole organismile vajalik. Võttes vaatluse alla homoloogsed valgu järjestused, näiteks ühe viiruse perekonna mingid valgud, siis näiteks juba 5 aminohappelise regiooni kodeerimiseks on $4^5 = 1024$ võimalust, kui kasutada aminohappeid, millel on 4 sünonüümset koodonit. Kui aga kõik antud perekonna viirused kasutavad mingis piirkonnas ainult ühte varianti nendest, siis võib eeldada, et antud järjestus on oluline ka DNA tasemel. Antud probleemi analüüsimiseks töötati välja meetod CDEP (*conserved DNA element detection in protein coding sequences*) hindamaks, kas valguline motiiv on konserveerunud funktsiooni tõttu või toimub positiivne surve DNA tasemel kindlate koodonite osas, mis hoiab rangelt muutumatuna ka aminohappelist järjestust. Tihti kasutatakse viirustes topeltkodeerivate piirkondade analüüsil MLOGD programmi, kuid see ei ole võimeline tuvastama regioone, mille pikkus on alla 20 koodoni. CDEP meetod suudab tuvastada lühikesi, alla 20 koodoni pikkuseid, topeltkodeerivaid alasid ning lisaks nendele ka teisi konserveerunud piirkondi DNA tasemel.

Probleemi paremaks mõistmiseks uurime ühte näidet lambdapapilloomiviiruste E1 valgu põhjal. Mutatsioonide tekkimiseks koodonites peab olema möödunud piisavalt aega. Teadaolevalt lahknesid lambdapapilloomiviiruste eellased teistest perekondadest ~100 miljonit aastat tagasi ning jagunemine tüüpideks toimus ~50-60 miljonit aastat tagasi (Shah et al., 2010). Samuti teame, et lambdapapilloomiviirustel kogu kodeeriva ala mutatsiooni kiirus on ligikaudu $1{,}95\,x10^{-8}$ nukleotiidset asendust ühe saidi kohta aastas. Kuna me uurime E1 valku, siis on targem kasutada ainult E1 valgu mutatsiooni kiirust ning see on ligikaudu $1{,}76\,x10^{-8}$ nukleotiidset asendust ühe saidi kohta aastas (Rector et al., 2007). Seega ühe positsiooni kohta tuleb $1{,}76\,x10^{-8}\ asendust/aastas\ x\ 50\,000\,000\ = 0{,}88\,asendust.$ Antud tulemusest näeme, et igas positsioonis võib leiduda mutatsioon 88% tõenäosusega. Võib eeldada, et on kulunud piisavalt aega papilloomiviiruste tüüpide lahknemisel lambdapapilloomiviiruste perekonnas, et saanuks tekkida mutatsioonid koodonites.

Järgnevalt võtame vaatluse alla ühe piirkonna lambdapapilloomiviiruste E1 valgust, antud

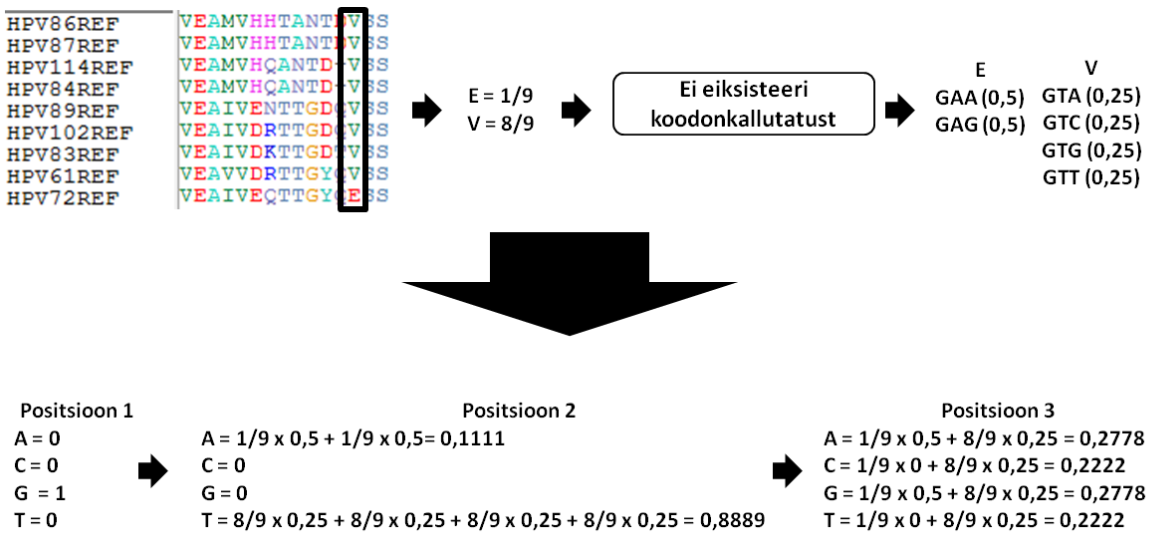
juhul regioon, kus võib asuda potentsiaalselt E8 ORF-i algus (joonis 10). Selgelt on näha, et valgulisel tasemel on konserveerunud HEA (histidiin-glutamiinhape-alaniin) muster (joonis 10A). Kui antud motiiv on vajalik ainult valgu funktsiooniks, siis võiks DNA tasemel kodeerida seda mistahes sünonüümsete koodonitega (joonis 10B). Samas on selgelt näha, et E8 alguses kasutatakse H kodeerimiseks ainult CAT koodonit, et säiliks E8 ATG. Samuti kasutab viirus glutamiinhappe kodeerimiseks ainult GAA ning ka alaniini kodeerimiseks ainult GCT koodonit (joonis 10C). Eksisteerib positiivne surve antud koodonite kasutamisele.

Eeldusel, et ka teistes papilloomiviiruste perekondades on E1 valgu mutatsioonikiirus sarnane, saab neid analüüsida samade kriteeriumite alusel. Vaadates näitena alfapapilloomiviirustes sama regiooni, kus on E8 algus, on näha, et näiteks G (glütsiini) kodeerimiseks enne E8 ORF-i algust kasutatakse 3 erinevat koodonit, kuid E8 alguses ainult ühte (lisa 3). Seega on möödunud piisavalt aega, et on saanud tekkida mutatsioonid koodonites, kuid mõnes positsioonis on kasutusel ikka ainult üks koodon. Eelpool toodust saab järeldada, et kui võrrelda vaadeldud sünonüümsete koodonite sagedust eeldatavatega, siis nende võrdlemisel on võimalik tuvastada kodeerivaid alasid.

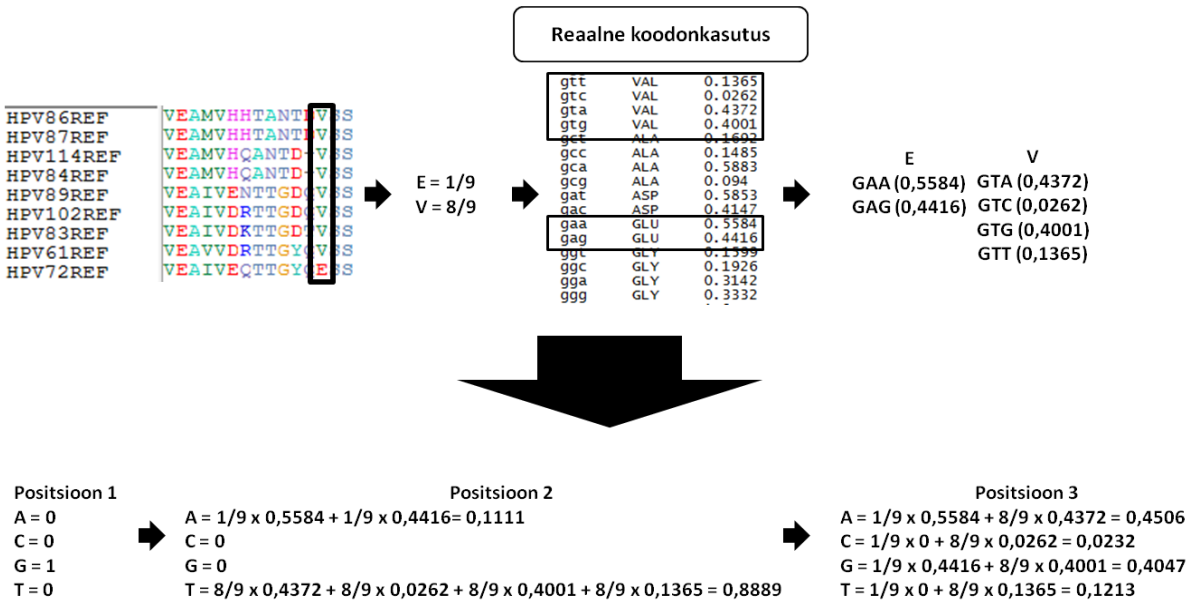


Joonis 10. Lambdapapilloomiviiruste E1 valgu joondus ning valgu joondusele vastav DNA joondus E8 ORF-i alguses. **(A)** Lambdapilloomiviiruste E1 valgu joondus vasakul ning paremal sellele vastav nukleotiidne joondus. E1 valgu joondus tehtud MUSCLE v3.8.31 ning valgu joonduse põhjal tehtud DNA joondus PAL2NAL-iga. **(B)** Selekteeritud regioonis olevate aminohapete sünonüümsed koodonid. **(C)** Antud regioonis kasutatud koodonite kolmanda positsiooni nukleotiidide sagedused.

CDEP meetodit algandmete genereerimiseks (vaadeldud ja eeldatud nukleotiidide sagedused) koguti esmalt kokku PaVe andmebaasist kõik papilloomiviiruste E1 valgu järjestused. Sealjuures hilisem analüüs teostati PV perekondade kaupa. E1 valgu joonduse tegemiseks kasutati, kas MUSCLE v3.8.31 või MAFFT v6.846b. MUSCLE puhul kasutati ka *refine* sätet, mis proovib parandada olemasolevat järjestust. Kasutades PAL2NAL-i tehti valgu joonduse põhjal DNA joondus. Järgnevalt kirjutati programm *Python*-is *PositionReader.py*, mis võtab sisendiks valgujoonduse ning antud joonduse põhjal genereeritud DNA joonduse. Programm leiab esmalt igas positsioonis esinevate nukleotiidide sagedused (vaadeldud sagedus antud joonduse positsioonis). Seejärel arvutab välja analüüsis oleva perekonna E1 valgu koodonkasutuse, seda leitakse ainult E1 lugemisraami põhjal, sest on näidatud, et koodonkasutus võib varieeruda geenide vahel. Kuna antud töö ülesanne on tuvastada topeltkodeerivat ala E1 lugemisraamis, siis on oluline rakendada just selle geeni koodonkasutust. Järgnevalt leiab programm ennustatud nukleotiidide sagedused igale positsioonile kahel viisil. Esiteks eeldades, et kõikide sünonüümsete koodonite esinemissagedused on võrdsed (joonis 11), teiseks võttes arvesse ka uurimise all olevate järjestuste koodonkasutust (joonis 12).



Joonis 11. Ennustatud nukleotiidide esinemissagedus, kui ühe aminohappe sünonüümsete koodonite esinemissagedused on võrdsed. Esmalt loetakse kokku kõik ühes positsioonis esinevad aminohapped ning leitakse nende sagedused. Järgnevalt vaadatakse, milliste koodonitega on võimalik neid aminohappes kodeerida ning eeldatakse, et iga koodoni esinemissagedus on võrdne. Järgnevalt leitakse igale positsioonile nelja nukleotiidi esinemissagedused. Näitena kasutatud üheksat alfapapilloomiviirust.



Joonis 12. Ennustatud nukleotiidide esinemissagedus võttes arvesse reaalset koodonkasutust E1 valgus. Esmalt loetakse kokku kõik ühes positsioonis esinevad aminohapped ning leitakse nende sagedused. Järgnevalt leitakse uurimise all olevate valkude koodonkasutus ehk E1 reaalne koodonkasutus üle kogu uurimise all oleva papilloomiviiruse perekonna. Edasi leitakse igale positsioonile nelja nukleotiidi esinemissagedused, rakendades eelnevalt leitud koodonkasutuse tabelit. Näitena kasutatud üheksat alfapapilloomiviirust.

2.2.4. CDEP meetod – andmete analüüs

Eelmises punktis genereeritud andmete (vaadeldud ja eeldatud nukleotiidide sagedused) analüüsiks ja visualiseerimiseks kirjutati skript *R*-is. Andmete analüüsis rakendati järgnevaid statistikuid:

1) $RMSD = \sqrt{\frac{1}{4}[(A_{obs} - A_{pre})^2 + (C_{obs} - C_{pre})^2 + (G_{obs} - G_{pre})^2 + (T_{obs} - T_{pre})^2]}$
2) $MaxDif = \max(|A_{obs} - A_{pre}|, |C_{obs} - C_{pre}|, |G_{obs} - G_{pre}|, |T_{obs} - T_{pre}|)$
3) $\chi^2\ test$

RMSD (*root-mean-square deviation*) kasutatakse laialdaselt andmete analüüsis. A_{obs} tähendab vaadeldud, A_{pre} ennustatud A nukleotiidi sagedust, sama ka teiste nukleotiidide puhul.

$\boldsymbol{MaxDif}$ puhul vaadati igas positsioonis vaadeldud ja eeldatud nukleotiidide sageduste erinevuste absoluutväärtust ning leiti nelja seast antud positsioonis suurima erinevusega variant. Leitud tulemus määrati statistiku väärtuseks.

$\boldsymbol{\chi^2\ test}$ analüüsib vaadeldud nukleotiidide arvu jaotuse erinevust ennustatud nukleotiidide jaotusest antud positsioonis. Statistiku väärtuseks võeti –log(p-väärtusest).

2.2.5. CDEP meetodi eeldused

CDEP meetodi ülesehitusest tingitult peavad olema täidetud järgnevad eeldused:

1. DNA piisav divergents - liikide/tüüpide vaheline lahknemine on toimunud piisavalt kaua aega tagasi, et oleks saanud toimuda mutatsioonid geenides.
2. Hea joondus valgu tasemel – joondus ei tohi sisaldada suuri tühimikke ning meid huvitavad piirkonnad peavad olema korrektselt joondatud (piisav konserveerumus valgu tasemel).
3. Uuritav valk ei tohi olla üle 30% topeltkodeeriv – tekitab nihke reaalses koodonkasutuses.
4. Analüüsiks peaks olema ühest perekonnast minimaalselt vähemalt 6 erinevat homoloogset valgu järjestust.

2.3. Tulemused

2.3.1. E8 CDS-i tuvastamine erinevates papilloomiviiruse tüüpides

PaVe andmebaasist hangiti 274 kättesaadavat PV genoomi. Kasutades enda loodud programmi *E8ORFSearch.py* leiti E8 CDS-i olemasolu 267-el papilloomiviirusel 274st (lisa 1). Tuvastatud E8 CDS-id asusid kõik +1 raamis E1 ORF-i suhtes ning nende keskmine pikkus oli 38,4 nukleotiidi, sealjuures lühemad olid 23 nukleotiidi (HPV88, HPV112, HPV119, HPV136, HPV147, HPV164, TmPV2) ja kõige pikemad 56 nukleotiidi (DdPV1, PsPV1, TtPV1, TtPV3, TtPV4, TtPV5, TtPV6). Leitud E8 CDS-ide keskmine kaugus E1 algusest oli 376,7 nukleotiidi, sealjuures väikseim kaugus oli 307 (BpPV1, BPV10, BPV12) ja suurim 415 (HPV40 ja HPV43).

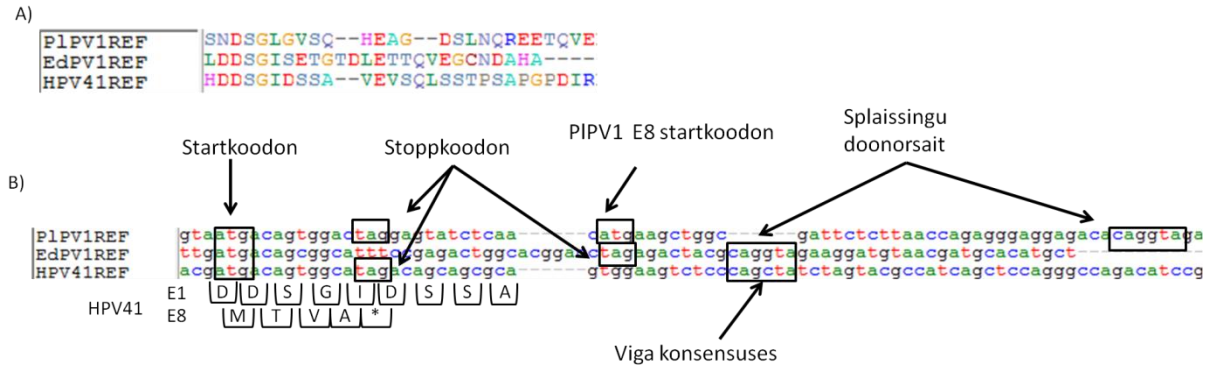
Etteantud piirangutega ei leitud E8 CDS-i 7-st E1 lugemisraamist (tabel 2). Nende hulka kuulusid kõik kolm teadaolevat linnupapilloomiviirust: PePV1, FcPV1 ja FlPV1 (*Francolinus leucoscepus papillomavirus* e kuldkurk-frankoliin PV) ning kaks merikilpkonna papilloomiviirust: CmPV1 (*Chelonia mudas papillomavirus* e. rohelise merikilpkonna PV) ja CcPV1 (*Caretta caretta papillomavirus e.* puupea merikilpkonna PV), lisaks veel EdPV1 (*Erethizon dorsatum papillomavirus* e Ursoni PV) ja HPV41. Muutes piiranguid lõdvemaks ehk lubades E1 pikkusel suuremalt varieeruda ning paikneda lähemal E1 algusele, siis ei suudetud ikka tuvastada nendes viirustes E8 lugemisraami.

Järgnevalt uuriti lähemalt HPV41-te, sest kõikidel teistel HPV-del tuvastati E8 CDS. Analüüsi kaasati temale taksonoomiliselt lähedane papilloomiviirus EdPV1, kellel samuti ei tuvastatud E8 CDS-i. Järgnevalt kasutati programmi *BLAST* leidmaks mõlema viiruse E1 valgule sarnaseim E1 valk papilloomiviiruste seast, kellel leiduks E1 CDS. Mõlema PV E1 valguga on sarnaseim PlPV1 (*Procyon lotor papillomavirus* e. pesukaru PV) (joonis 13A). Võrreldes nende kolme viiruse E1 DNA järjestusi, leiti, et kõigil on sarnases positsioonis ATG, kuid sellele järgneb varsti TAG stoppkoodon (joonis 13B), kuid PlPV1 on olemas pärast stoppkoodonit ka sekundaarne ATG, kust alustatakse E8 kodeerimist, kuid teistel mitte. Lisaks ei eksisteeri HPV41-l ka korrektset splaissingu aktseptorsaiti.

Tabel 2. Papilloomiviirused, kust ei tuvastatud E8 CDS-i

	E1^E4*	Perekond	Peremeesorganism
PePV1	-	Teeta	Aafrika halli papagoi PV
EdPV1	+	Sigma	Ursoni PV
HPV41	+	Nu	Inimese PV
FcPV1	-	Eta	Metsvindi PV
FlPV1	+	Dyoepsilon	Kuldkurk-frankoliin PV
CmPV1	+	Dyozeeta	Rohelise merikilpkonna PV
CcPV1	+	Dyozeeta	Puupea merikilpkonna PV

* PaVe andmebaasi alusel



Joonis 13. EdPV1, HPV41 ja PlPV1 E1 valgu joondus ja sellel põhinev DNA joondus. Vaatluse all on piirkond, kus PlPV1 ennustati E8 CDS. **(A)** Valgu joondus tehtud MAFFT v6.846b **(B)** DNA joondus genereeritud valgu joonduse põhjal PAL2NAL-iga. PlPV1-l on ATG sarnases positsioonis EdPV1 ja HPV41-ga, kuid kõigil kolmel on pärast seda stardikoodonit terminatsioonikoodon (TAG). PlPV1 E8 kodeerimine hakkab järgmisest ATG-st (märgitud kastiga), kuid EdPV1 ja HPV41 ei eksisteeri teist ATG-d. Lisaks HPV41 puhul ei eksisteeri ka korrektset splaissaiti (tähistatud musta ristkülikuga).

Loodud programmi töö hindamiseks ehk positiivseteks kontrollideks on kõik eksperimentaalselt kinnitatud ja teadaolevad E8^E2C lugemisraamide asukohad (tabel 3). Tulemustest on näha, et viga tehti ainult BPV1 puhul, kuid genoomi kontrollimisel selgus, et PaVe andmebaasis oli BPV1 kirje ebakorrektne - 1204 positsioonis on ATG asemel ANG. Vaadates kirjet AB626705 on antud positsioonis ATG. Parandades genoomis antud vea, leidis programm korrektselt ka BPV1 puhul E8 CDS-i.

Tabel 3. Varem kirjeldatud E8 CDS-ide asukohtade võrdlus, antud töö raames ennustatutega.

PV[1]	E8 ATG[2]	Splaissingu doonorsait[3]	Takson[4]	Leitud E8 ATG[5]	Leitud doonorsait[6]
SFPV1	1720	1751	10623	1720	1751
BPV1*	**1204**	**1235**	**10559**	**1399**	**1451**
HPV1a	1200	1231	10583	1200	1231
HPV5	1328	1359	333923	1328	1359
HPV11	1241	1272	10580	1241	1272
HPV16	1265	1302	333760	1265	1302
HPV18	1323	1357	333761	1323	1357
HPV31	1259	1296	10585	1259	1296
HPV33	1276	1316	10586	1276	1316

1. Esimene tulp on papilloomiviiruse tüüp
2. Eksperimentaalselt kinnitatud E8 CDS-i algus genoomis
3. Splaissingu doonorsaidi asukoht genoomis
4. NCBI taksoni ID
5. *E8ORFSearch.py* skripti poolt tuvastatud E8 CDS-i alguse asukoht genoomis
6. *E8ORFSearch.py* skripti poolt tuvastatud E8 splaissingu doonorsait

* Punasega tähistatud BPV1 puhul ei kattunud minu poolt leitud tulemus kirjanduses leiduvaga

Järgnevalt vaadati, milliseid splaissingu doonorsaite kasutavad erinevate papilloomiviiruste perekonnad, sest juba kirjanduses leiduvast infost nägime, et ühel beetapapilloomiviirusel (HPV5) oli CAGGTA asemel GAGGTA. Analüüsi teostati perekondade kaupa ning vaatluse alla võeti perekonnad, kus on kirjeldatud 4 või rohkem PV tüüpi, väiksemad perekonnad liitsin kokku üheks (tabel 4). Tulemustest selgub, et enamjaolt kasutatakse CAGGTA splaissingu doonorsaidi konsensust, kuid beetapapilloomiviirused ja pipapilloomiviirused kasutavad eranditult ainult GAGGTA konsensust. Lisaks leidub ka üksikuid teisi splaissingu doonorsaidi variante.

Tabel 4. E8^E2C splaissingu doonorsaitide jaotus perekondade kaupa.

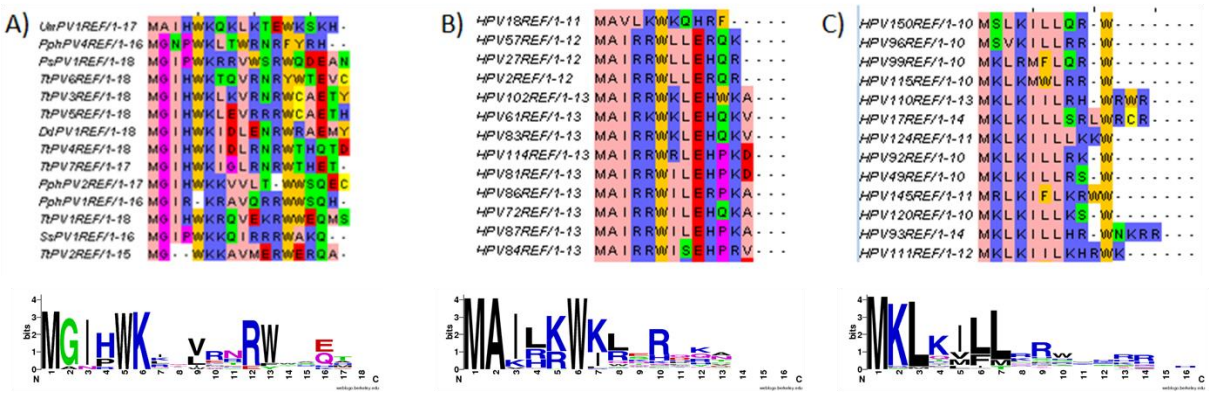
PV[1]	Kokku[2]	AGGTA[3]	AGGTG[4]	AGGTC[5]	AGGTT[6]	puudu[7]	Täpsustus[8]
Alfa	77	77	-	-	-	-	67 CAGGTA 10 gAGGTA
Beeta	47	47	-	-	-	-	47 gAGGTA
Gamma	50	49	-	-	1	-	45 CAGGTA 4 aAGGTA 1 CAGGTt
Delta	12	12	-	-	-	-	12 CAGGTA
Lambda	10	10	-	-	-	-	10 CAGGTA
Omikron	4	4	-	-	-	-	4 CAGGTA
Pi	6	6	-	-	-	-	6 gAGGTA
Xi	7	7	-	-	-	-	7 CAGGTA
Chi	10	10	-	-	-	-	10 CAGGTA
Upsilon	7	6	-	-	1	-	5 CAGGTA 1 CAGGTt 1 aAGGTA
Teised*	44	34	2	0	1	7	28 CAGGTA 4 gAGGTA 2 aAGGTA 1 CAGGTg 1 gAGGTg 1 CAGGTt

1. Papilloomiviiruse perekond
2. Perekonnas olevate papilloomiviiruste arv
3-6. Otsitud splaissingu doonorsaidid
7. Papilloomiviiruste arv, kellel antud perekonnas ei tuvastatud E8 CDS-i
8. Täpsemalt välja toodud splaissingu doonorsaidid ja nende arv.
* Teised on kõik ülejäänud papilloomiviiruste perekonnad

2.3.2. E8 perekonnad

Leidmaks, kas E8 valku saab klassifitseerida erinevatesse perekondadesse, koondati kokku kõik E8 valgujärjestused ning joondati need MUSCLE v3.8.31 vaikesätetega. Tulemustest saab selgelt eristada kolme E8 valgu perekonda (joonis 14). Esimese grupi ehk kõige väiksema rühma moodustavad vaalaliste ja delfiinide papilloomiviirused (joonis 14A), vastavalt siis omikron- ja upsilonpapilloomiviiruste E8 valgud (lisa 2) ning lisaks ka üks dyopipapilloomiviirus PphPV4 (*Phocoena phocoena papillomavirus* e harilik pringeli PV) ja eranditena dyodeltapapilloomiviirus SsPV1 (*Sus scrofa domestica papillomavirus* e kodusea PV) (Stevens, Rector, Van Der Kroght, et al., 2008) ning omegapapilloomiviirus UmPV1

(*Ursus maritimus Papillomavirus* e jääkaru PV) (Stevens, Rector, Bertelsen, et al., 2008). Antud grupis on iseloomulik GIHWK motiiv ja kaks osaliselt konserveerunud trüptofaani (W) positsioonides 5 ja 13 (joonis 14A). Teise rühma moodustavad alfapapilloomiviirustele kuuluvad E8 valgud (joonis 14B), seal on näha teises positsioonis täielikult konserveerunud alaniini (A) ning kuuendas positsioonis trüptofaani ning osaliselt konserveerunud isoleutsiini positsioonis 3. Kolmanda rühma moodustavad kõik ülejäänud papilloomiviirused (joonis 14C). Tegemist on kõige suurema perekonnaga, seal on näha osaliselt konserveerunud lüsiini (K) ja leutsiini (L) aminohappejääki E8 CDS-i alguses. Antud rühmas on paljudele tüüpidele iseloomulik KILL motiiv E8 keskosas. Üks klassifitseerimata papilloomiviirus MrPV1, kust leitud E8 järjestus (MRPKENWMSRWKK) oli erinev kõigist kolmest E8 perekonnast jäi antud puhul perekonda paigutamata.



Joonis 14. **E8 valkude perekonnad. (A)** Esimene perekond on kõige väiksem, sinna kuuluvad vaalaliste (omikron, dyopi) ja delfiiniliste (upsilon) papilloomiviirused. Antud rühmas on eranditena üks dyodeltapapilloomiviirus SsPV1 ja omegapapilloomiviirus UmPV1. **(B)** Teise perekonna moodustavad alfapapilloomiviiruste E8 valgud (joonisel näitena toodud osade alfapapilloomiviiruste E8 valgud). **(C)** Kolmas perekond on kõige suurem ja selle moodustavad kõik ülejäänud papilloomiviirused va üks klassifitseerimata papilloomiviirus MrPV1 (joonisel näitena toodud osade beetapapilloomiviiruste E8 valgud). Joondused tehtud iga perekonna siseselt eraldi MUSCLE v3.8.31, tulemused illustreeritud Jalview-ga ning zappo värvidega (hüdrofoobsed – roosa, hüdrofiilsed – rohelised, aromaatsed – oranž, positiivselt laetud – sinised, negatiivselt laetud – punased, süsteiin – kollane, proliin ja guaniin - lillad). Iga grupi all konsensuse logod, kasutatud veebiprogrammi weblogo (http://weblogo.berkeley.edu/).

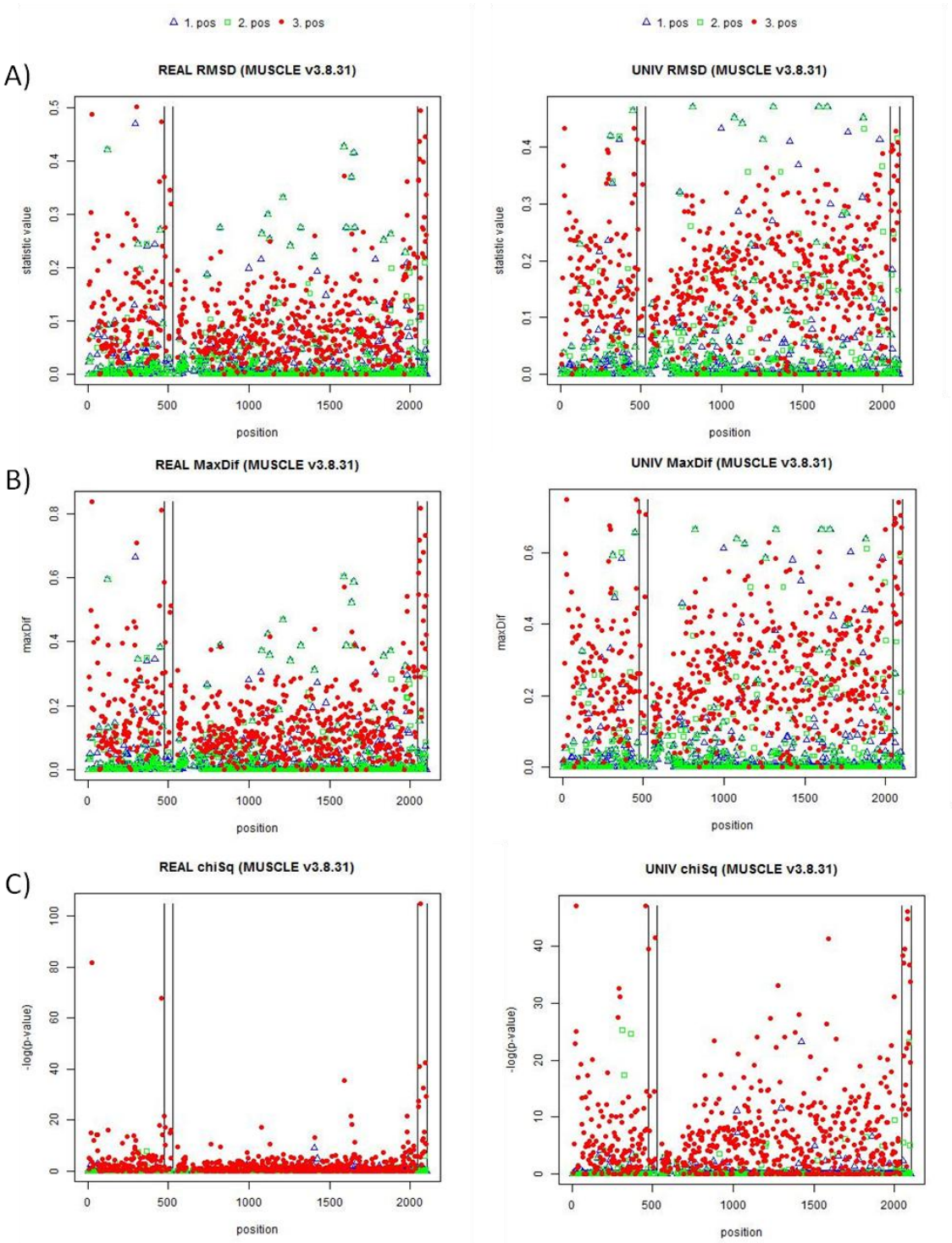
2.3.3. E8 CDS-i tuvastamine CDEP meetodiga E1 lugemisraamis alfapapilloomiviiruste näitel

Hindamaks, kas E8 lugemisraam on E1 valgulise järjestuse kõrvalprodukt või eksisteerib positiivne surve DNA tasemel kasutati CDEP meetodit. Kõigepealt hangiti PaVe andmebaasist kõikide alfapapilloomiviiruste E1 valkude järjestused (77 järjestust) ning joondati need MUSCLE-ga. MSA-s tekib peaaegu alati tühimikke (*gap*) erinevatesse positsioonidesse, kas halva joondamise tõttu või mõnes järjestuses esineva indel-i (*insertion* + *deletion*) tõttu. Suurematest insertsioonidest vabanemiseks ehk joonduse parandamiseks, tuleb vahel mõningad järjestused eemaldada. Alfapapilloomiviiruste E1 valgu joondusest eemaldasin HPV94, HPV117 ja HPV10 E1 valgu järjestused, sest need põhjustasid pikka indel-i E8 lugemisraami läheduses, mis omakorda põhjustas nihet mitmes teises järjestuses. Järgnevalt tehti valgujoondusest PAL2NAL-iga DNA joondus ning analüüsiti neid CDEP meetodiga. Analüüsist jätan välja joonduse positsioonid, kus esineb üle 20% järjestustest tühimikke. Nendes positsioonides pannakse statistiku väärtuseks 0.

Võrreldes tulemusi E1 valgu reaalse koodonkasutuse (REAL) korral ja olukorras, kus koodonkasutuse kallutatust ei arvesta (UNIV) ehk sünonüümsete koodonite sagedused on võrdsed, näeme et kõigi kolme statistiku puhul on REAL korral signaali müra on palju madalam (joonis 15). RMSD REAL puhul on statistikute aritmeetiline keskmine (0,046 < 0,072) kui ka mediaan (0,011 < 0,013) madalam võrreldes RMSD UNIV-ga. Sarnast erinevust näeme ka MaxDif juures, kus REAL korral on aritmeetiline keskmine (0,068 < 0,108) kui ka mediaan (0,016 < 0,018) madalamad. Samuti on ka REAL $\chi^2\ test$ puhul on statistiku [-log(p-väärtus)] aritmeetiline keskmine madalam (1,064 < 2,144). Sarnast müra taseme erinevust on näha ka teistes perekondades (lisa 5). Samuti on tulemustest näha, et kolmas positsioon on informatiivsem. Tegemist on ootuspärase tulemusega, sest üldjuhul varieerubki ainult koodoni kolmas positsioon. Teise või esimese nukleotiidi muutus koodonis põhjustab üldjuhul ka aminohappe muutust.

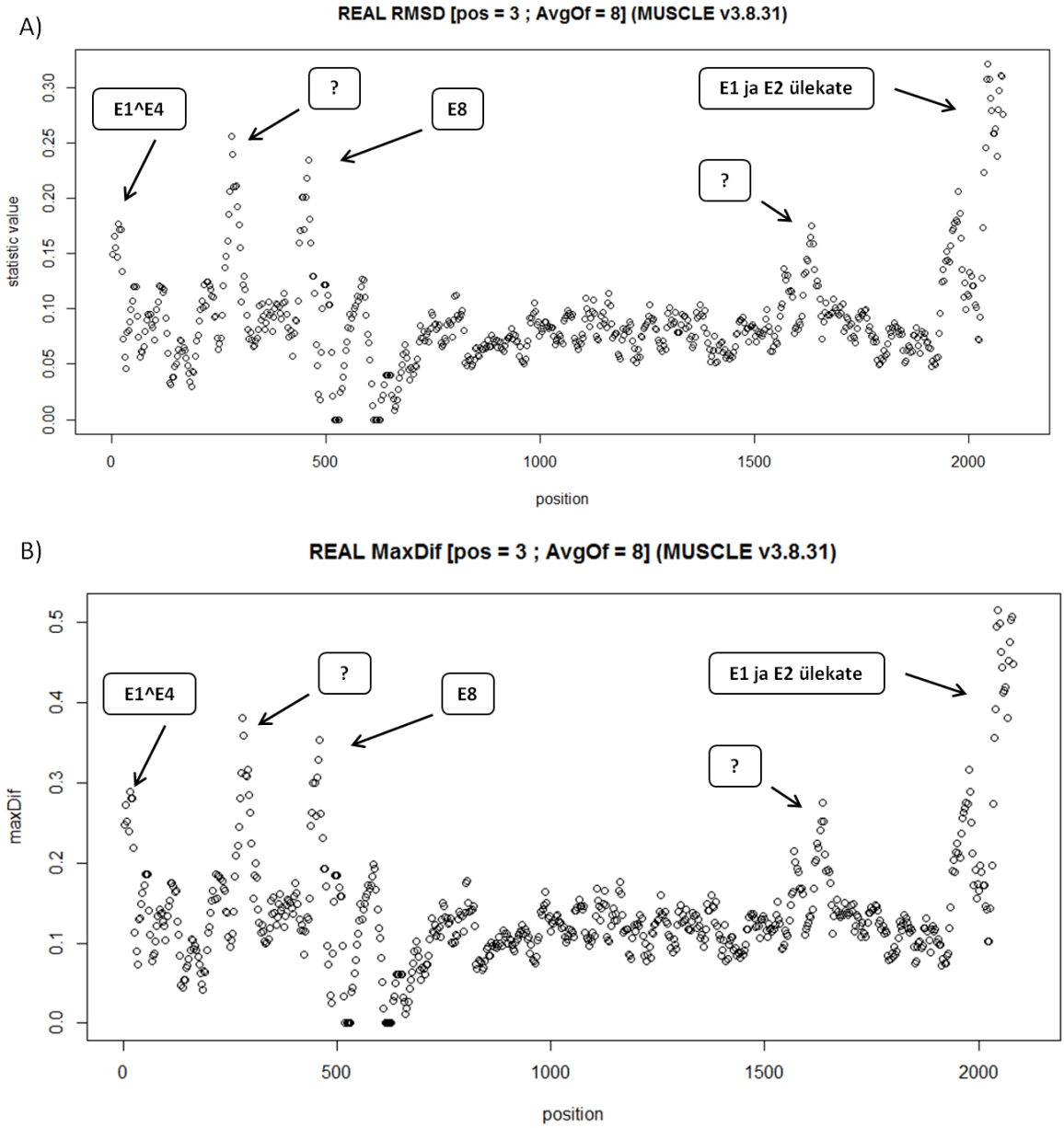
Joonis 15 graafikutelt on näha tugevamaid signaale E1 alguses ja lõpus. E1 ORF-i alguses, ligikaudu 15 nukleotiidi pärast ATG asub E1^E4 splaissingu doonorsait, mis põhjustab koodonite konserveerumust ning signaali teket. Ligikaudu 370 nukleotiidi kaugusel asub E8 CDS, mille ATG ja ka splaissingu doonorsait andsid signaali (MSA asukohaks 500 – joonis 15). Samuti saime ootuspäraselt signaali ka E1 valgu ORF-i lõpus, kus asub E1 ja E2 ORF-i ülekate. Antud piirkond on väga heaks positiivseks kontrolliks, kas CDEP meetod suudab

tuvastada topeltkodeerivaid alasid.



Joonis 15. **Alfapapilloomiviiruste E1 ORF-i analüüs CDEP meetodiga.** REAL (vasakul) tähistab olukorda, kus võetakse arvesse E1 kodeerimiseks kasutatud koodonite kallutatust. UNIV (paremal) puhul eeldatakse, et kõikide sünonüümsete koodonite esinemissagedus on võrdne. Teljed joonistel: y-telg tähistab statistiku väärtust ning x-telg positsiooni MSA-l. Sinised kolmnurgad koodoni esimene positsioon, rohelised ruudud teine positsioon ning punased ringid kolmas positsioon. Kahe paralleelse joonega vasakul pool tähistatud piirkond, kus asub E8 CDS ning paremal positiivse kontrollina E1 ja E2 ORF-i ülekate. **(A)** RMSD statistik**. (B)** MaxDif statistik **(C)** $\chi^2\ test$ –log(p-väärtusest).

Järgnevalt vaadati ainult koodonite kolmandate positsioonide CDEP statistikute väärtuseid ning analüüsiti E1 järjestust piirkondade kaupa, võttes arvesse reaalset koodonkasutust E1 järjestuses. Vaadeldava piirkonna ehk akna suuruseks valiti lühim eksisteeriv E8 CDS. Antud töö raames tuvastatud E8 CDS-idest lühim oli 23 nukleotiidi, seega akna suuruseks määrati 23/3 = ~8. Analüüsides andmeid kaheksase libiseva aknaga ehk liikudes edasi ühe koodoni võrra ja iga kord arvutades 8 järjestikuse koodoni kolmanda positsiooni CDEP statistiku aritmeetilise keskmise, on võimalik eristada olulisi piirkondi hõlpsamalt. Statistikute liitmist saab rakendada ainult RMSD ja MaxDif puhul ning mõlemad annavad sarnaseid tulemusi (joonis 16). Antud tulemustest on näha, et selline lähenemine on visuaalselt parem. Graafikult on väga hästi eristatav E1^E4 splaissingu doonorsaidi kui ka E1 ja E2 ülekatte signaal. Samuti näeme, et alfapapilloomiviirustel on konserveerunud E8 CDS DNA tasemel. Lisaks kolmele eelpool mainitud piirkonnale leiti ka kaks tundmatut. Uurides lähemalt joondust, siis leidsime, et E1^E4 splaissingu doonorsaidi ja E8 vahel asuvas piirkonnas (290 – 301) asub konsensus TAAAACGAAAGT, mis on konserveerunud 98% alfapapilloomiviirustes. Teine signaali, mis asub E8 ning E1 ja E2 ülekatte vahel, on joondusel piirkonnas 1569 – 1641. Antud regioonis leidub mitmeid konserveerunud alasid, kuid huvipakkuv võib olla lühike motiiv ATGAG positsioonis 1618-1622.



Joonis 16. Alfapapilloomiviiruste E1 ORF-ide analüüs CDEP meetodiga. Leitakse E1 lugemisraamis kaheksa järjestikuse koodoni kolmanda positsiooni CDEP statistikute aritmeetiline keskmine ning märgitakse graafikule piirkonna esimese koodoni positsioonile. Akent liigutatakse edasi ühe koodoni kaupa. Graafiku y-telg tähistab statistikute aritmeetilist keskmist ning x-telg positsiooni MSA-l. Joondused saadud 74 alfapapilloomiviiruse E1 valgu järjestuse põhjal. Mõlemal joonisel on näha viite tugevat signaali, kolm nendest on teadaolevad piirkonnad: E1^E4 splaissingu doonorsait, E8 CDS ja E1 ja E2 ORF-ide ülekate, kuid kahe piirkonna funktsioon on teadmata. **(A)** RMSD statistik kaheksase libiseva aknaga **(B)** MaxDif statistik kaheksase libiseva aknaga.

2.3.4. CDEP meetodi rakendamine teistes papilloomiviiruste perekondades

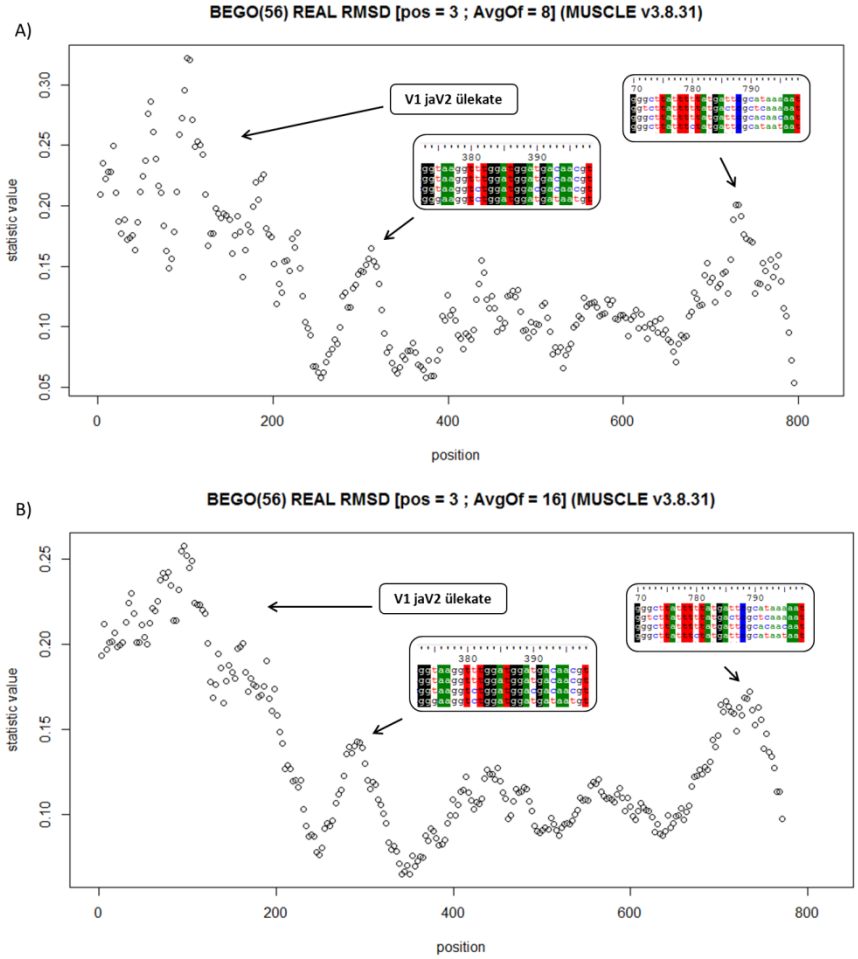
Analüüsides kõiki teisi perekondi, kus on vähemalt 6 liiget sain sarnased tulemused alfapapilloomiviirustega (lisa 6). Kõigil oli hästi eristatav E8 signaal, va xi- ja upsilonpapilloomiviirustel, kuigi nendel tuvastati E8 CDS eelnevalt, antud töö raames. Positiivse kontrollina oli kõikidel PV perekondadel tuvastatav E1 ORF-i lõpus ülekate E2 lugemisraamiga ning enamustel ka E1^E4 splaissingu doonorsaidi signaal. Lisaks oli delta-, chi-, lambda- ja gammapapilloomiviirustel tuvastatav alfapapilloomiviirustega sarnane signaal E1^E4 splaissingu doonorsaidi ja E8 signaali vahel. Beeta-, xi ja upsilonpapilloomiviiruste puhul on näha, et millegi tõttu on pidev kõrge signaal E1 valgu järjestuse algusosas. Samas beetapapilloomiviiruse puhul on siiski eristatavad kõik 3 põhilist signaali, kuid xi- ja upsilonpapilloomiviirustel mitte. Nende puhul on keerulisem eristada E8 signaali ning xipapilloomiviirustes ka E1^E4 splaissaidi signaali. Upsilon- ja lambdapapilloomiviirused on omapärased, sest neil on näha üle kogu E1 valgu ORF-i signaale. Erinevalt upsilonpapilloomiviirusest on lambdapapilloomiviirustes lihtne eristada E1^E4 splaissingu doonorsaiti ja E8 signaali.

2.3.5. CDEP meetodi rakendamine begomoviirustes

Erinevate papilloomiviiruste perekondade E1 valgu ORF-i analüüsimine näitas, et antud töö raames loodud CDEP meetod tuvastab hästi topeltkodeerivaid alasid ning ka teisi DNA tasemel konserveerunud piirkondi geeni kodeerivas osas, olenemata selle funktsioonist. Testimaks, kas antud meetod toimib ka teiste viiruste peal, võtsin vaatluse alla *Geminiviridae* sugukonna, tegemist on taimede ssDNA viirustega (Khatri et al., 2014). Nende seast valisin begomoviiruste perekonna, sest neil on V1 valgu (kattevalk) ORF-i 5’ algusosa ülekattes V2 valgu (pre-kattevalk) ORF-iga, seda umbes ~200nt ulatuses (näiteks *Soybean mild mottle* viiruse puhul 208nt). Ühtlasi on tegemist ka antud sugukonna suurima perekonnaga. Järgnevalt teostasin NCBI andmebaasis otsingu leidmaks RefSeq-i täisgenoome, kellel on annoteeritud V1 ja V2 geen (NCBI Otsing = *Begomovirus[Organism] AND V1[GENE] AND V2[GENE] AND srcdb_refseq[PROP] NOT wgs[PROP] NOT cellular organisms[ORGN]*). Antud kriteeriumitele vastas 59 begamoviiruse genoomi. Järgnevalt leidsin, et mingil põhjusel oli V1 valk annoteeritud V2-na 8 genoomis (NC_003896, NC_004654, NC_004626, NC_005348, NC_004569, NC_005347, NC_002817, NC_013639). Joonduse parandamiseks eemaldasin NC_013017, NC_013639 ning NC_000870 V1 järjestused, sest nende puhul oli

annoteeritud ainult V1 valgu mingi väike piirkond. Järgnevalt joondasin V1 valgud MUSCLE-ga ning kasutasin ka *refine* valikut.

Analüüsides begamoviiruseid libiseva aknaga tuleks esmalt paika panna kasutatava/otsitava piirkonna pikkus. Kasutades sama lähenemist, mida E8 puhul, peaks akna suuruseks võtma otsitava regiooni pikkuse koodonites - 208/3 = 69 (*Soybean mild mottle* viiruse näitel). Antud juhul on tõenäosus kaotada väiksemate regioonide signaale, seega otsustati esialgu proovida kaheksase libiseva aknaga, sarnaselt papilloomiviirustele (joonis 17A) ja hiljem 16 libiseva aknaga (joonis 17B). Mõlemal korral on tuvastatav joondusel ligikaudu 200 nukleotiidi ulatuses signaal. Lisaks otsitavale ülekattele tuvastati ka kaks tundmatut signaali.



Joonis 17. Begamoviiruste V1 ORF-i analüüs CDEP meetodiga, kasutades RMSD statistikut. Graafiku y-telg tähistab statistiku väärtust ning x-telg positsiooni joondusel. **(A)** RMSD kaheksase libiseva aknaga. **(B)**RMSD statistik 16-se libiseva aknaga. Mõlemal grafikul on näha ~200 nukleotiidi ulatuses signaali, mis tähistab V1 ORF-i ülekatet V2 ORF-iga. Lisaks välja toodud kaks piirkonda, mis mõlemal puhul andsid signaali.

2.4. Arutelu

Antud töö käigus leiti, et 274-st PaVe andmebaasist kättesaadavast papilloomiviiruse genoomist 267-el võib reaalselt eksisteerida E8 CDS E1 lugemisraamis. Tuvastatud E8 CDS-id olid fikseerunud +1 lugemisraamis E1 suhtes. Antud analüüsis selgus, et kõigil inimeste papilloomiviirustel leidub E8 CDS, välja arvatud HPV41-el. Uurides lähemalt HPV41-te ja temale taksonoomiliselt lähedast papilloomiviirust EdPV1, siis selgus et mõlema viiruse E1 lugemisraamis on E8 kriteeriumitele vastav järjestus, kuid nendes tuvastati sisemised stoppkoodonid ning HPV41 ei omanud kanoonilist splaissaidi konsensust. Nende puhul võib olla tegemist ka sekveneerimisveaga ning mõistlik oleks need üle kontrollida. Samas on üpriski ebatõenäoline, et mõlemal taksonoomiliselt lähedasel viirusel on korraga tekkinud E8 lugemisraamis sekveneerimisviga, mis tekitab üheaegselt stoppkoodoni. Pigem on tegemist evolutsioonis olulise muutusega. Võrreldes HPV41-e ja EdPV1-e E8 algussaiti PlPV1-ga näeme, et kõigil kolmel eksisteerib pärast esimest ATG-d stoppkoodon. Sealjuures omab PlPV1 sekundaarset ATG-d, kust teoreetiliselt algab ka E8 kodeerimine. Vaadates HPV41 E1 valgu järjestust, siis on näha, et positsioon, kus asub E8 stoppkoodon, ei ole oluline E1 valgule, sest isoleutsiini (I) võiks kodeerida ATA asemel ka ATT või ATC-ga, mis kaotaksid E8 lugemisraamist stoppkoodoni ning asendaksid selle vastavalt siis TTG→leutsiini või TCG→seriiniga (joonis 13).

Huvitav on ka fakt, et E8 CDS-i ei tuvastatud ühelgi teadaoleval linnupapilloomiviirusel (PePV1, FcPV1 ja FlPV1). Võib spekuleerida, et lindude epiteelkude erineb piisaval määral imetajate omast ning nende puhul on kasutusel teised rajad E8^E2C funktsiooni täitmiseks. Samas on ennustatud antud viirustel E1 lugemisraamis E9 ORF, mis võib täita E8^E2C funktsiooni, kuid hetkeseisuga ei teata, kas reaalselt antud valk üldse ekspresseerub (Van Doorslaer et al., 2009). PePV1 ja FcPV1 puhul ei ole annoteeritud ka E4 lugemisraami E2 valgus (PaVe andmebaasi põhjal), seega nendes viirustes ilmselt ei leidu ka kanoonilist E1^E4 valku, mis tähendab, et E2 lugemisraamis ei pruugi eksisteerida ka korrektset splaissingu aktseptorsaiti, mida kasutab ka E8^E2C. Samuti ei tuvastatud E8 CDS-i ka kahel teadaoleval merikilpkonna papilloomiviirusel: CmPV1 ja CcPV1, kuigi mõlemal on annoteeritud E1^E4 (Herbst et al., 2009). Antud viirustel ei ole leitud ka mõnda teist ORF, mis võiks asendada E8^E2C funktsiooni.

2013 aasta lõpus tuli PaVe välja uuendatud andmebaasiga, kus lisati papilloomiviiruse

kirjetesse nende poolt ennustatud E8 CDS-ide asukohad. Võrreldes nende ja antud töö tulemusi, leiti 80 papilloomiviiruse puhul erinevusi (lisa 7). Üldiselt oli neil erinevale kohale ennustatud splaissait, kuid vahel ka E8 ATG. Uurides lähemalt näiteks HPV104, siis näeme, et E8 ATG on ennustatud samasse kohta 1296, kuid antud töö raames ennustati splaissingu doonorsait positsioonile 1327 (GAGGTA), kuid PaVe-s on see positsioonil 1321 (GAGGTG). Mõlemad on splaissingu konsensused ning saadavad E8 järjestuse pikkused õiged, kuid HPV104 on beetapapilloomiviirus nagu ka HPV5, mille puhul on näidatud, et E8 kasutab GAGGTA splaissingu doonorsaiti, kuigi ka HPV5-el eksisteerib samas kohas GAGGTG konsensus. Seega võib eeldada, et korrektsem on doonorsait positsioonil 1327 (GAGGTA). Suurt ebausaldusväärsust tekitab asjaolu, et PaVe andmebaasis on valesti annoteeritud ka E8 CDS-id, mis on eksperimentaalselt juba eelnevalt tõestatud. Näiteks BPV1 puhul ennustavad nad E8 CDS-i positsioonidele 1270..1451, kuid eelnevalt on teada, et BPV1 E8 ORF asub positsioonis 1204...1235. Antud juhul võib probleem olla nende andmebaasis olevas vigases kirjes. Lisaks BPV1-le on nad aga valesti annoteerinud ka mitme teise teadaoleva papilloomiviiruse E8 splaissingu doonorsaidi, nende hulka kuulub HPV5, HPV31 ja HPV33 (täpsem info tabelis lisa 7). Teadmata nende annoteerimise kriteeriumeid on keeruline spekulaarida, mis põhjusel on neil valesti annoteeritud antud papilloomiviiruste E8 CDS-id. Nähtud vead aga sunnivad kriitilisemalt suhtuma ka teistesse nende poolt annoteeritud E8 asukohtadesse.

Uurides lähemalt, milliseid splaissingu doonorsaite ennustati erinevatele perekondadele, selgus, et kõik teadaolevad beeta- ja pipapilloomiviirused kasutavad ainult GAGGTA konsensust. Põhjus võib olla viiruste või koespetsiifilisuses, kuid keeruline on näha nende kahe perekonna vahel seost, sest pipapilloomiviirused nakatavad hamstrite limaskesta, kuid beetapapilloomiviirused peamiselt inimeste, kuid ka primaatide nahka (lisa 2). Samas vaadates nii beeta- kui ka pipapilloomiviiruste E1 valgu ja DNA joondust näeme, et GAGGTA moodustab glutamiinhappe koodoni ning see võib olla oluline hoopis E1 funktsioonis ning selle tõttu on konserveerunud GAGGTA konsensus CAGGTA asemel.

Papilloomiviiruste E8 CDS-i saab klassifitseerida 3 perekonda. Kõige väiksema rühma moodustavad vaalaliste papilloomiviirused ning erandina üks kodusea (SsPV1) ja jääkaru papilloomiviirus (UmPV1) (joonis 14B). UmPV1 E8 järjestus (MAIHWKQKLKTEWKSKH) on alguse osas küll sarnane alfapapilloomiviiruste E8 valguga (MAI), kuid HWK motiiv ja kaks trüptofaani on pigem vaalaliste E8 perekonna tunnuseks. Fülogeneetilisel puul (joonis 1) moodustavad SsPV1 ja UmPV1 koos vaalaliste seltsi kuuluv hariliku pringeli (PphPV4) ja

nahkhiire papilloomiviirusega (MrPV1) eraldi haru, seega on ka loogiline, et nad kuuluvad ühte perekonda ka E8 alusel. MrPV1 puhul on tegemist isesuguse E8 järjestusega (MRPKENWMSRWKK) ning esialgu jääb see klassifitseerimata. Kuna antud E8 perekonnas on enamus siiski vaalaliste papilloomiviirused uurisin lähemalt, kuidas tuvastati eelpoolnimetatud kodusea ja jääkaru papilloomiviirused. Selgus, et kodusea PV puhul saadi proovid kahelt nakatumata ehk tervelt kodusea naha pinnalt. Vatitikkudega võeti proovid 4 kohast: seljalt, vasakult esijalalt, pea ning nina pealt (Stevens, Rector, Van Der Kroght, et al., 2008). Kuna proove ei võetud nakatunud koest, vaid tervetelt isenditelt, siis võib spekuleerida, et nende nahale võis sattuda juhuslikult mõne veeorganismi papilloomiviirus ja reaalselt ei olegi tegemist kodusea papilloomiviirusega. Samuti puudub SsPV1 genoomist E7 ORF, mis on üldiselt vaalaliste papilloomiviiruste ühine tunnus (Stevens, Rector, Van Der Kroght, et al., 2008). Lisaks, kui genereerida fülogeneetiline puu E1 valgu järjestuse alusel, näeme, et SsPV1 klassifitseerub kokku ainult vaalaliste papilloomiviirustega (lisa 8) ning omavahel klassifitseeruvad kokku MrPV1 ja UmPV1. Sealjuures UmPV1 puhul eraldati proovid jääkaru suu limaskesta papilloomilt (Stevens, Rector, Bertelsen, et al., 2008). Lisaks on huvitav asjaolu, et ka UmPV1-el ja MrPV1-el puudub kanooniline E7 ORF. E8 teise grupi moodustavad alfapapilloomiviirused (joonis 14B), nende puhul võib olla tegemist spetsialiseerumisega primaatide epiteelkoele. Kõige suuremas perekonna moodustavad kõik ülejäänud E8 järjestused (joonis 14C). Antud rühmas on valgulise järjestuse varieeruvus suurem, kuid siiski on tegemist sarnaste järjestustega. Antud perekonnas on oluline konserveerunud lüsiin teises positsioonis. Muteerides antud positsioonis oleva lüsiini alaniiniks põhjustab see E8^E2C represseeriva funktsiooni kadumist (Shankovski, Karro, Ustav ja Abroi publitseerimata andmed).

Analüüsides alfapapilloomiviiruste perekonnas E1 valgu järjestust CDEP meetodiga nägime, et UNIV puhul oli üldine signaalitase kõrgem ning selle põhjustas asjaolu, et võeti aminohapete sünonüümsete koodonite esinemisesagedused võrdseks. Selline olukord tekitab aga tugevaid signaale, sest viiruse reaalne koodonkasutus võib olla märkimisväärselt teistsugune. Tulemus on ootuspärane, sest tänapäevase teadmiste alusel esineb koodonkasutuse kallutatus isegi geenide tasemel (Camiolo et al., 2012; Plotkin & Kudla, 2011). Samuti leiti, et koodonite kolmandad positsioonid on informatiivsemad, sest antud positsioonides saab toimuda mutatsioonid ilma, et valgujärjestus muutuks. Järjestuste paremaks analüüsimiseks ja visualiseerimiseks kasutati CDEP meetodit libiseva aknaga. Üheks meetodi probleemiks on akna suuruse määramine, see oleneb, milliseid motiive otsida. Mida väiksema aknaga otsid, seda paremini leiad lühemate regioonide signaale. Antud töös

võeti akna suuruseks otsitava regiooni pikkus (E8 CDS-i pikkus). Suuremate regioonide otsimisel võib selline lähenemine probleemseks osutuda, sest on võimalus, et kaotatakse väiksemate regioonide signaalid. Papilloomiviiruste puhul määrati akna suuruseks 8 koodonit, sest lühim E8 CDS, mis antud töö raames tuvastati oli 23 nukleotiidi, mis teeb ligikaudu 8 koodonit. Määratud akna suurusega oli võimalik tuvastada alfapapilloomiviirustes E8 signaal ning sellega tõestada, et antud järjestus ei ole tekkinud juhuslikult valgu aminohappelisest järjestusest, vaid regioonile on positiivne surve DNA tasemel. Positiivsete kontrollidena suutis CDEP meetod alfapapilloomiviirustes tuvastada nii E1^E4 splaissingu doonorsaidist kui ka E1 ja E2 ülekattest tingitud signaali. Uurides teisi papilloomiviiruste perekondi CDEP meetodiga (lisa 6) näeme, et E8 signaal on kõigis tuvastatav, va xi- ja upsilonpapilloomiviiruste perekonnas. Mõlemas perekonnas oli E1 alguses pikalt kõrge signaalitase ning see raskendas E8 CDS tuvastamist visuaalselt, kuigi eelnevalt antud töös tuvastati mõlemas perekonnas E8 CDS-id. Sellise pideva signaali päritolu nõuab lähemat uurimist.

CDEP meetod tuvastas E1 ORF-is lisaks kolmele teadaolevale ja oodatud regioonile ka kaks tundmatut ala. Esimene piirkond asus MSA-l positsioonis 290 – 301 ning teine 1654 – 1667 (joonis 16A). Esimene tundmatu regioon asub E1^E4 splaissingu doonorsaidi ja E8 signaali vahepeal. Tegemist on TAAAACGAAAGT konsensusega, mis on konserveerunud sellisena 98% alfapapilloomiviirustes. Lähemal uurimisel leiti, et kirjanduses on antud regiooni mainitud juba 1991. aastal, kuid funktsioon on täpselt teadmata. Spekuleeritakse, et see element võib olla vajalik tuumalokalisatsiooniks, kodeerides vajalikke aminohappeid, samas koodoni kolmas positsioon võiks sellisel juhul varieeruda. Tegemist võib olla ka erinevate rakuliste faktorite seondumiskohaga (Campione-Piccardo & Montpetit, 1991). Sarnases positsioonis oli signaal tuvastatav ka gamma-, lambda-, delta- ja chipapilloomiviiruste perekonnas (lisa 6). Järjestuste lähemal vaatlusel selgus, et samuti nendes perekondades oli konserveerunud sarnased motiivid: gammapapilloomiviirustes leidus TAAA[A/G]CGAAAGT, lambdapapilloomiviirstes T[A/T]AA[A/G][C/A]GAAA[G/A]T, deltapapilloomiviirustel T[G/C]AAAAGAAAA[C/G/T]T ning chipapilloomiviirustel T[A/G]AAA[A/C]GAAA[G/A][T/C] järjestus. Uurides lähemalt beeta-, xi-, pi- ja upsilonpapilloomiviiruste perekonda, kus antud signaal ei olnud nii hästi eristatav, selgus, et beetapapilloomiviirustes on see hästi konserveerunud, problemaatiliseks tegi antud piirkonna täpse tuvastamise asjaolu, et seal läheduses oli ka teisi konserveerunud regioone. Ülejäänud perekondades oli motiiv samuti tuvastatav, näiteks xipapilloomiviirustes oli järjestus [T/C]AAA[A/G]CGAAA[G/C][T/C], pipapilloomiviirustes TAAA[A/G]CGAAAG[T/C],

kuid upsilonpapilloomiviirustes oli konserveerunud ainult lõpuosa [C/G][A/C][A/G]AA[T/A/C]CGAAAGT. Teine signaali asub E8 ning E1 ja E2 ülekatte signaali vahepeal (joonis 16). Antud regioonis leidub mitmeid konserveerunud alasid, kuid huvipakkuv võib olla lühike motiiv ATGAG positsioonis 1618-1622, sealjuures stardikoodon ATG on E1 suhtes +1 raamis. Antud piirkonna kohta informatsioon puudub.

Vaadates deltapapilloomiviiruste CDEP tulemusi on näha tugevat signaali E1 ja E2 ORF-i ülekatte ees (lisa 6). BPV1-el on näidatud, et ~220 nukleotiidi enne E1 geeni lõppu asub P2443 promootor, mis aktiveerib nii E2 kui ka E5 transkriptsiooni (Hermonat, Spalholz, & Howley, 1988). Deltapapilloomiviiruste CDEP graafikult tuvastatud signaal enne E1 lõppu võib peegeldada antud promootori mingi elemendi signaali.

Vaatamaks, kas CDEP meetod ei ole kuidagi E1 valgu või papilloomiviiruste spetsiifiline ning toimiks ka teiste valkude / viiruste peal võeti uurimise alla begomoviiruste perekond. Begomoviiruste kattevalgu V1 ORF-i algusosa on ülekattes V2 pre-kattevalgu ORF-iga 200nt ulatuses. CDEP meetod oli võimeline tuvastama antud piirkonda ning lisaks ka kaks teist regiooni. Esimene tundmatu regioon asus DNA MSA positsioonides 373 – 494 ning teine 770-798. Antud juhul puudub informatsioon nende kahe piirkonna funktsiooni ning olulisuse kohta ning nõuavad edaspidist uurimist. Begomoviiruste V1 ORF-i analüüsimisel CDEP meetodiga on probleemiks topeltkodeeriva osa liiga suur osakaal, mis võib mõjutada tulemusi. *Soybean mild mottle* viiruse puhul on see näiteks 27%.

KOKKUVÕTE

Papilloomiviirused on võimelised nakatama suurt hulka organisme, mille sekka kuuluvad maismaa- ja veeimetajaid kui ka roomajad ning linnud. Olenemata suurest peremeesorganismide mitmekesisusest on nende genoome ülesehitus üpriski sarnane. Kõigil papilloomiviirustel on siiski esindatud ainult tuumikgeenid E1, E2, L1 ja L2. Varajased geenid E1 ja E2 osalevad genoomi replikatsioonis ning geenide ekspressioonis. Lisaks nendele kahele tuumikgeenile on kirjeldatud ka E8^E2C valk, mis samuti osaleb mõlemas protsessis. Eksperimentaalselt on E8^E2C kirjeldatud ainult mõningates papilloomiviirustes, kuid selle olulisuse tõttu võiks seda leiduda ka teistes papilloomiviirustes.

Antud töö raames loodud programmiga (*E8ORFSearch.py*) suutsime tuvastada E8 CDS-i 267-l papilloomiviirusel 274-st. E8 CDS-i ei tuvastatud ühelgi linnu ega merikilpkonna PV-l ning samuti ka ühel inimese ja ursoni papilloomiviirusel. Võrreldes antud töö tulemusi PaVe andmebaasis olevate andmetega, leiti erinevused kaheksakümnes E8 annotatsioonis.

Transleerides tuvastatud E8 CDS-id, saab need jaotada kolme perekonda. Esimese, kõige väiksema perekonna moodustavad vaalaliste papilloomiviirused. Erandina kuulub siia rühma veel üks jääkaru ja kodusea papilloomiviirus. Teise ehk keskmise suurusega perekonna moodustavad alfapapilloomiviiruste E8 valgud. Viimase ja kõige suurema E8 perekonna moodustavad kõik ülejäänud papilloomiviirused va MrPV1.

Antud töö käigus loodi CDEP meetod, millega on võimalik valku kodeerivas DNA järjestuses tuvastada piirkondi, kus eksisteerib positiivne surve DNA tasemel. E1 ORF-e analüüsiti CDEP meetodiga ning leiti, et E8 ei ole säilinud E1 aminohappelise järjestuse tõttu, vaid on konserveerunud DNA tasemel. Selline tulemus lubab meil oletada, et tegemist on olulise regulaator valguga papilloomiviirustes. CDEP meetod suudab enamustes papilloomiviiruste perekondades hästi tuvastada ka teisi konserveerunud alasid - E1^E4 splaissingu doonorsaiti ning E1 ja E2 ülekatet. Alfapapilloomiviirustel ja ka teistes perekondades tuvastati antud meetodiga kaks tundmatut piirkonda, kus on positiivne surve DNA tasemel. Üks nendest piirkondadest osutus varem kirjeldatud konsensuseks, kuid teine nõuab lähemat uurimist.

SUMMARY

Title: Detection of E8 protein coding sequence in papillomaviruses *in silico*

Papillomaviruses (PV) are small non-encapsulated viruses with a circular double-stranded DNA genome. The size of the genome is approximately 8000 bp. PVs are able to infect the squamous epithelia (mucosa, skin), inducing proliferation lesions. Papillomavirus DNA has been recovered from the skin and/or lesions of many mammalian species (from land and water) but also from birds and reptiles (a python and two turtle species). High risk human papillomaviruses (HPV) can cause anogenital cancers including cervical cancer.

PV life cycle and gene expression is tightly linked to the squamous cell differentiation program. Based on that, the genes are separated into two groups – early (E) and late (L). PV genome generally contains eight ORFs: E1, E2, E4, E5, E6, E7, L1 and L2. The E2 protein is a multifunctional regulatory protein. The full-length E2 is a transcriptional activator and repressor, but also important in virus genome replication. Besides full-length E2, papillomaviruses encode an alternatively spliced product called E8^E2C. Studies have shown that E8^E2C is a transcription repressor and a negative regulator of replication. Presently, E8^E2C has been detected only in alphapapillomaviruses (HPV11, HPV16, HPV18, HPV31, HPV33), müüpapillomavirus (HPV1a), kappapapillomavirus (CRPV), deltapapillomavirus (BPV1) and betapapillomavirus (HPV5). If E8^E2C exists in taxonomically distant genera like alpha- and deltapapillomaviruses, then it should also be present in other genera.

In this study we gathered 274 PV genomes from PaVe database and identified E8 ORF in 267 of them. We could not find E8 ORF in the following seven PV-s: PePV1, EdPV1, HPV41, FcPV1, FlPV1, CmPV1 and CcPV1. After that, all E8 ORFs were translated into amino acid sequences and classified into three groups. The smallest group consisted of E8 proteins from the cetaceans' papillomaviruses, but there were two exceptions – one polar bear (UmPV1) and one pig papillomavirus (SsPV1). Second group included E8 proteins from alphapapillomaviruses and third group consisted of all the rest except MrPV1 that was left out.

Second step was to develop a method to measure whether the motif in DNA is conserved because of the amino acid sequence of the E1 protein or there is positive evolutional selection affecting the DNA sequence. The method that we developed is called CDEP and the idea is to

convert protein multiple sequence alignment to DNA sequence alignment and then measure codon usage bias in every position. With CDEP, we could prove that E8 is conserved on DNA level in every larger PV genus. This method also clearly identified E1^E4 RNA splicing donor site and E1E2 overlap at the end of E1 ORF. In Alphapapillomaviruses, we saw two extra signals. One was previously described consensus TAAAACGAAAGT which is conserved in 98% of alphapapillomaviruses. Second signal was between the E8 and E1E2 overlap signal. There is no information about this consensus and it needs further investigation.

To assess if the CDEP method works on other viruses/proteins as well, we analyzed the begomovirus genus. They are plant viruses and belong to the taxonomic family *Geminiviridae*. Their ssDNA genome encodes V1 protein whose ORF overlaps with V2 ORF. CDEP method was able to detect this overlap perfectly and it also spotted two unknown conserved motifs.

This study was done in the Chair of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu and Estonian Biocentre. I would like to thank my supervisor Aare Abroi and co-supervisor Maido Remm for the help and good advice.

KASUTATUD KIRJANDUS

Bennett, M. D., Woolford, L., Stevens, H., Van Ranst, M., Oldfield, T., Slaven, M., … Nicholls, P. K. (2008). Genomic characterization of a novel virus found in papillomatous lesions from a southern brown bandicoot (Isoodon obesulus) in Western Australia. *Virology*, *376*(1), 173–82. doi:10.1016/j.virol.2008.03.014

Bergvall, M., Melendy, T., & Archambault, J. (2013). The E1 proteins. *Virology*, *445*(1-2), 35–56. doi:10.1016/j.virol.2013.07.020

Buck, C. B., Day, P. M., & Trus, B. L. (2013). The papillomavirus major capsid protein L1. *Virology*, *445*(1-2), 169–74. doi:10.1016/j.virol.2013.05.038

Camiolo, S., Farina, L., & Porceddu, A. (2012). The relation of codon bias to tissue-specific gene expression in Arabidopsis thaliana. *Genetics*, *192*(2), 641–9. doi:10.1534/genetics.112.143677

Campione-Piccardo, J., & Montpetit, M. (1991). A highly conserved nucleotide string shared by all genomes of human papillomaviruses. *Virus Genes*, 349–357

Chiang, C. M., Broker, T. R., & Chow, L. T. (1991). An E1M--E2C fusion protein encoded by human papillomavirus type 11 is asequence-specific transcription repressor. *Journal of Virology*, *65*(6), 3317–29

Choe, J., Vaillancourt, P., Stenlund, a, & Botchan, M. (1989). Bovine papillomavirus type 1 encodes two forms of a transcriptional repressor: structural and functional analysis of new viral cDNAs. *Journal of Virology*, *63*(4), 1743–55

De Villiers, E.-M. (2013). Cross-roads in the classification of papillomaviruses. *Virology*, *445*(1-2), 2–10. doi:10.1016/j.virol.2013.04.023

De Villiers, E.-M., Fauquet, C., Broker, T. R., Bernard, H.-U., & zur Hausen, H. (2004). Classification of papillomaviruses. *Virology*, *324*(1), 17–27. doi:10.1016/j.virol.2004.03.033

DiMaio, D., & Petti, L. M. (2013). The E5 proteins. *Virology*, *445*(1-2), 99–114. doi:10.1016/j.virol.2013.05.006

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, *12*(6), 640–9

Fertey, J., Ammermann, I., Winkler, M., Stöger, R., Iftner, T., & Stubenrauch, F. (2010). Interaction of the papillomavirus E8--E2C protein with the cellular CHD6 protein contributes to transcriptional repression. *Journal of Virology*, *84*(18), 9505–15. doi:10.1128/JVI.00678-10

Fertey, J., Hurst, J., Straub, E., Schenker, A., Iftner, T., & Stubenrauch, F. (2011). Growth inhibition of HeLa cells is a conserved feature of high-risk human papillomavirus E8^E2C proteins and can also be achieved by an artificial repressor protein. *Journal of Virology*, *85*(6), 2918–26. doi:10.1128/JVI.01647-10

Firth, A. E., & Brown, C. M. (2005). Detecting overlapping coding sequences with pairwise alignments. *Bioinformatics (Oxford, England)*, *21*(3), 282–92. doi:10.1093/bioinformatics/bti007

Firth, A. E., & Brown, C. M. (2006). Detecting overlapping coding sequences in virus genomes. *BMC Bioinformatics*, *7*, 75. doi:10.1186/1471-2105-7-75

Gottschling, M., Bravo, I. G., Schulz, E., Bracho, M. a, Deaville, R., Jepson, P. D., … Nindl, I. (2011). Modular organizations of novel cetacean papillomaviruses. *Molecular Phylogenetics and Evolution*, *59*(1), 34–42. doi:10.1016/j.ympev.2010.12.013

Herbst, L. H., Lenz, J., Van Doorslaer, K., Chen, Z., Stacy, B. a, Wellehan, J. F. X., … Burk, R. D. (2009). Genomic characterization of two novel reptilian papillomaviruses, Chelonia mydas papillomavirus 1 and Caretta caretta papillomavirus 1. *Virology*, *383*(1), 131–5. doi:10.1016/j.virol.2008.09.022

Hermonat, P. L., Spalholz, B. a, & Howley, P. M. (1988). The bovine papillomavirus P2443 promoter is E2 trans-responsive: evidence for E2 autoregulation. *The EMBO Journal*, *7*(9), 2815–22

Hubbert, N. L., Schiller, J. T., Lowy, D. R., Lerner, A. B., & Androphyt, E. J. (1988). Bovine papilloma virus-transformed cells contain multiple E2 proteins. *Biochemistry*, *85*(August), 5864–5868.

Khatri, S., Nahid, N., Fauquet, C. M., Mubin, M., & Nawaz-Ul-Rehman, M. S. (2014). A betasatellite-dependent begomovirus infects ornamental rose: characterization of begomovirus infecting rose in Pakistan. *Virus Genes*. doi:10.1007/s11262-014-1076-6

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59. doi:10.1186/1471-2105-5-59

Kumar, S., & Subramanian, S. (2002). Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(2), 803–8. doi:10.1073/pnas.022629899

Kurg, R., Sild, K., Ilves, A., Sepp, M., & Ustav, M. (2005). Association of bovine papillomavirus E2 protein with nuclear structures in vivo. *Journal of Virology*, *79*(16), 10528–10539. doi:10.1128/JVI.79.16.10528

Kurg, R., Uusen, P., Sepp, T., Sepp, M., Abroi, A., & Ustav, M. (2009). Bovine papillomavirus type 1 E2 protein heterodimer is functional in papillomavirus DNA replication in vivo. *Virology*, *386*(2), 353–9. doi:10.1016/j.virol.2009.01.025

Kurg, R., Uusen, P., Võsa, L., & Ustav, M. (2010). Human papillomavirus E2 protein with single activation domain initiates HPV18 genome replication, but is not sufficient for long-term maintenance of virus genome. *Virology*, *408*(2), 159–66. doi:10.1016/j.virol.2010.09.010

Lace, M. J., Anson, J. R., Thomas, G. S., Turek, L. P., & Haugen, T. H. (2008). The E8--E2 gene product of human papillomavirus type 16 represses early transcription and replication but is dispensable for viral plasmid persistence in keratinocytes. *Journal of Virology*, *82*(21), 10841–53. doi:10.1128/JVI.01481-08

Lambert, P. F. (1995). Bovine papillomavirus type 1 E2 transcriptional regulators directly bind two cellular transcription factors , TFIID and TFIIB . These include : Bovine Papillomavirus Type 1 E2 Transcriptional Regulators Directly Bind Two Cellular Transcription Factors , T. *Virology*, *69*(10), 6323–6334.

Lambert, P. F., Hubbert, N. L., Howley, P. M., & Schiller-, J. T. (1989). Genetic Assignment of Multiple. *Virology*, *63*(7), 3151–3154.

Lambert, P. F., Monk, B. C., & Howley, P. M. (1990). type 1 E2 repressor mutants . Phenotypic Analysis of Bovine Papillomavirus Type 1 E2 Repressor Mutants. *Virology*, *64*(2), 960–956.

Li, J., Li, Q., Diaz, J., & You, J. (2014). Brd4-mediated nuclear retention of the papillomavirus E2 protein contributes to its stabilization in host cells. *Viruses*, *6*(1), 319–35. doi:10.3390/v6010319

Longworth, M. S., & Laimins, L. A. (2004). Pathogenesis of Human Papillomaviruses in Differentiating Epithelia Pathogenesis of Human Papillomaviruses in Differentiating Epithelia. *Microbiology and Molecular Biology Reviews*, *68*(2), 362–372. doi:10.1128/MMBR.68.2.362

Najafabadi, H. S., Goodarzi, H., & Salavati, R. (2009). Universal function-specificity of codon usage. *Nucleic Acids Research*, *37*(21), 7014–23. doi:10.1093/nar/gkp792

Palermo-Dilts, D. a, Broker, T. R., & Chow, L. T. (1990). Human papillomavirus type 1 produces redundant as well as polycistronic mRNAs in plantar warts. *Journal of Virology*, *64*(6), 3144–9

Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews. Genetics*, *12*(1), 32–42. doi:10.1038/nrg2899

Powell, M. L. C., Smith, J. a, Sowa, M. E., Harper, J. W., Iftner, T., Stubenrauch, F., & Howley, P. M. (2010). NCoR1 mediates papillomavirus E8;E2C transcriptional repression. *Journal of Virology*, *84*(9), 4451–60. doi:10.1128/JVI.02390-09

Rautava, J., & Syrjänen, S. (2012). Biology of human papillomavirus infections in head and neck carcinogenesis. *Head and Neck Pathology*, *6 Suppl 1*, S3–15. doi:10.1007/s12105-012-0367-2

Rector, A., Lemey, P., Tachezy, R., Mostmans, S., Ghim, S.-J., Van Doorslaer, K., … Van Ranst, M. (2007). Ancient papillomavirus-host co-speciation in Felidae. *Genome Biology*, *8*(4), R57. doi:10.1186/gb-2007-8-4-r57

Rector, A., Tachezy, R., Van Doorslaer, K., MacNamara, T., Burk, R. D., Sundberg, J. P., & Van Ranst, M. (2005). Isolation and cloning of a papillomavirus from a North American porcupine by using multiply primed rolling-circle amplification: the Erethizon dorsatum papillomavirus type 1. *Virology*, *331*(2), 449–56. doi:10.1016/j.virol.2004.10.033

Rector, A., & Van Ranst, M. (2013). Animal papillomaviruses. *Virology*, *445*(1-2), 213–23. doi:10.1016/j.virol.2013.05.007

Roca, X., Krainer, A. R., & Eperon, I. C. (2013). Pick one, but be quick: 5' splice sites and the problems of too many choices. *Genes & Development*, *27*(2), 129–44. doi:10.1101/gad.209759.112

Roman, A., & Munger, K. (2013). The papillomavirus E7 proteins. *Virology*, *445*(1-2), 138–68. doi:10.1016/j.virol.2013.04.013

Sankovski, E., Männik, A., Geimanen, J., Ustav, E., & Ustav, M. (2014). Mapping of betapapillomavirus human papillomavirus 5 transcription and characterization of viral-genome replication function. *Journal of Virology*, *88*(2), 961–73. doi:10.1128/JVI.01841-13

Shah, S. D., Doorbar, J., & Goldstein, R. a. (2010). Analysis of host-parasite incongruence in papillomavirus evolution using importance sampling. *Molecular Biology and Evolution*, *27*(6), 1301–14. doi:10.1093/molbev/msq015

Sleator, R. D. (2010). An overview of the current status of eukaryote gene prediction strategies. *Gene*, *461*(1-2), 1–4. doi:10.1016/j.gene.2010.04.008

Snijders, P. J., van den Brule, a J., Schrijnemakers, H. F., Raaphorst, P. M., Meijer, C. J., & Walboomers, J. M. (1992). Human papillomavirus type 33 in a tonsillar carcinoma generates its putative E7 mRNA via two E6* transcript species which are terminated at different early region poly(A) sites. *Journal of Virology*, *66*(5), 3172–8

Stevens, H., Rector, A., Bertelsen, M. F., Leifsson, P. S., & Van Ranst, M. (2008). Novel papillomavirus isolated from the oral mucosa of a polar bear does not cluster with other papillomaviruses of carnivores. *Veterinary Microbiology*, *129*(1-2), 108–16. doi:10.1016/j.vetmic.2007.11.037

Stevens, H., Rector, A., Van Der Kroght, K., & Van Ranst, M. (2008). Isolation and cloning of two variant papillomaviruses from domestic pigs: Sus scrofa papillomaviruses type 1 variants a and b. *The Journal of General Virology*, *89*(Pt 10), 2475–81. doi:10.1099/vir.0.2008/003186-0

Stormo, G. D. (2000). Gene-Finding Approaches for Eukaryotes. *Genome Research*, *10*(4), 394–397. doi:10.1101/gr.10.4.394

Stubenrauch, F., Hummel, M., Iftner, T., & Laimins, L. a. (2000). The E8E2C protein, a negative regulator of viral transcription and replication, is required for extrachromosomal maintenance of human papillomavirus type 31 in keratinocytes. *Journal of Virology*, *74*(3), 1178–86

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(Web Server issue), W609–12. doi:10.1093/nar/gkl315

Van Doorslaer, K., Rector, A., Vos, P., & Van Ranst, M. (2006). Genetic characterization of the Capra hircus papillomavirus: a novel close-to-root artiodactyl papillomavirus. *Virus Research*, *118*(1-2), 164–9. doi:10.1016/j.virusres.2005.12.007

Van Doorslaer, K., Sidi, A. O. M. O., Zanier, K., Rybin, V., Deryckère, F., Rector, A., … Travé, G. (2009). Identification of unusual E6 and E7 proteins within avian papillomaviruses: cellular localization, biophysical characterization, and phylogenetic analysis. *Journal of Virology*, *83*(17), 8759–70. doi:10.1128/JVI.01777-08

Van Doorslaer, K., Tan, Q., Xirasagar, S., Bandaru, S., Gopalan, V., Mohamoud, Y., … McBride, A. a. (2013). The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Research*, *41*(Database issue), D571–8. doi:10.1093/nar/gks984

Vande Pol, S. B., & Klingelhutz, A. J. (2013). Papillomavirus E6 oncoproteins. *Virology*, *445*(1-2), 115–37. doi:10.1016/j.virol.2013.04.026

Wang, J. W., & Roden, R. B. S. (2013). L2, the minor capsid protein of papillomavirus. *Virology*, *445*(1-2), 175–86. doi:10.1016/j.virol.2013.04.017

Wilson, R., Ryan, G. B., Knight, G. L., Laimins, L. a, & Roberts, S. (2007). The full-length E1E4 protein of human papillomavirus type 18 modulates differentiation-dependent viral DNA amplification and late gene expression. *Virology*, *362*(2), 453–60. doi:10.1016/j.virol.2007.01.005

Woolford, L., Rector, A., Van Ranst, M., Ducki, A., Bennett, M. D., Nicholls, P. K., … O'Hara, A. J. (2007). A novel virus detected in papillomas and carcinomas of the endangered western barred bandicoot (Perameles bougainville) exhibits genomic features of both the Papillomaviridae and Polyomaviridae. *Journal of Virology*, *81*(24), 13280–90. doi:10.1128/JVI.01662-07

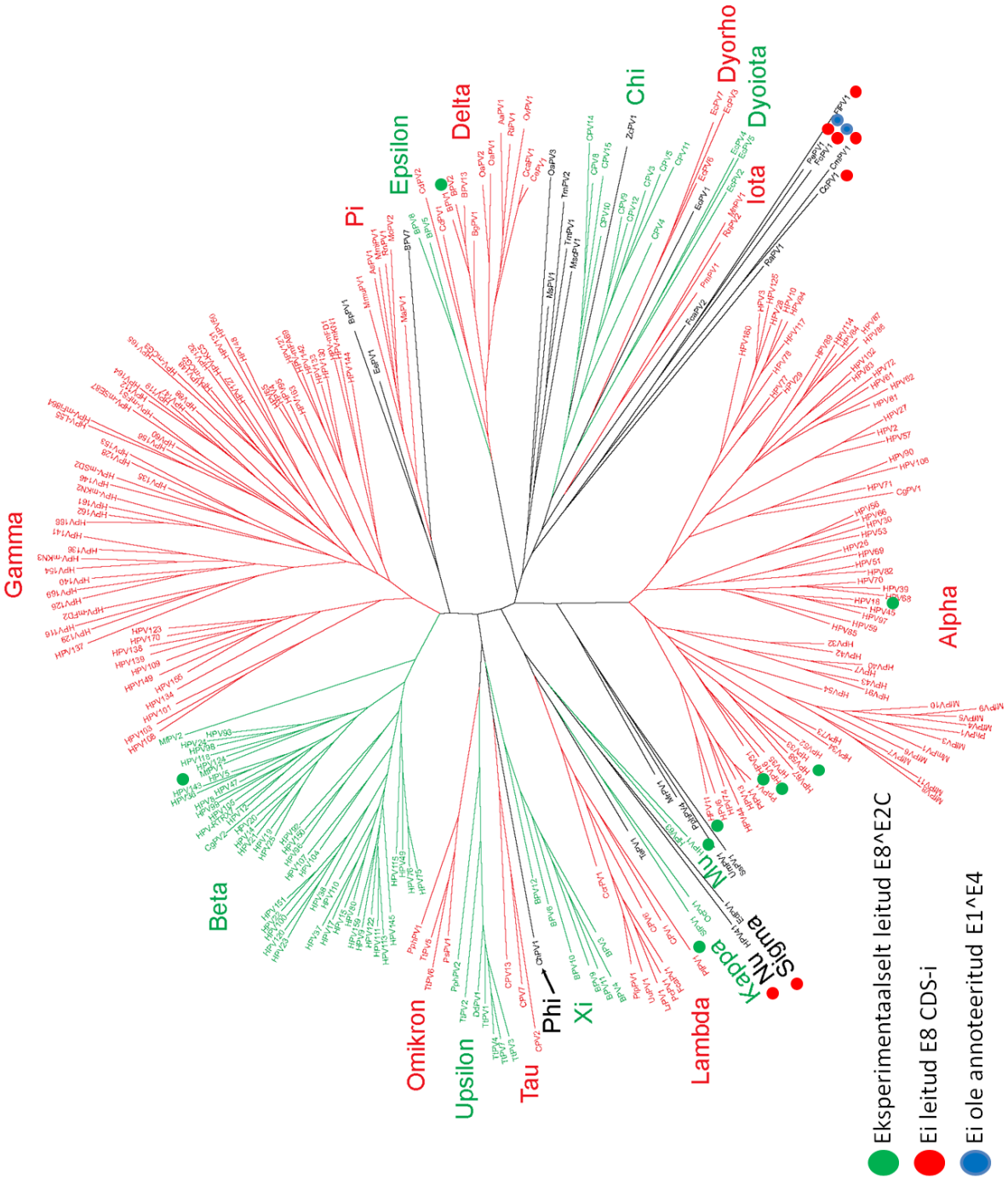
Yuan, H., Ghim, S., Newsome, J., Apolinario, T., Olcese, V., Martin, M., … Schlegel, R. (2007). An epidermotropic canine papillomavirus with malignant potential contains an E5 gene and establishes a unique genus. *Virology*, *359*(1), 28–36. doi:10.1016/j.virol.2006.08.029

Yuan, H., Luff, J., Zhou, D., Wang, J., Affolter, V., Moore, P., & Schlegel, R. (2012). Complete genome sequence of canine papillomavirus type 9. *Journal of Virology*, *86*(10), 5966. doi:10.1128/JVI.00543-12

KASUTATUD VEEBIAADRESSID

1. http://pave.niaid.nih.gov/
2. http://www.bork.embl.de/pal2nal/
3. http://weblogo.berkeley.edu/
4. http://www.ncbi.nlm.nih.gov/genbank/

LISAD

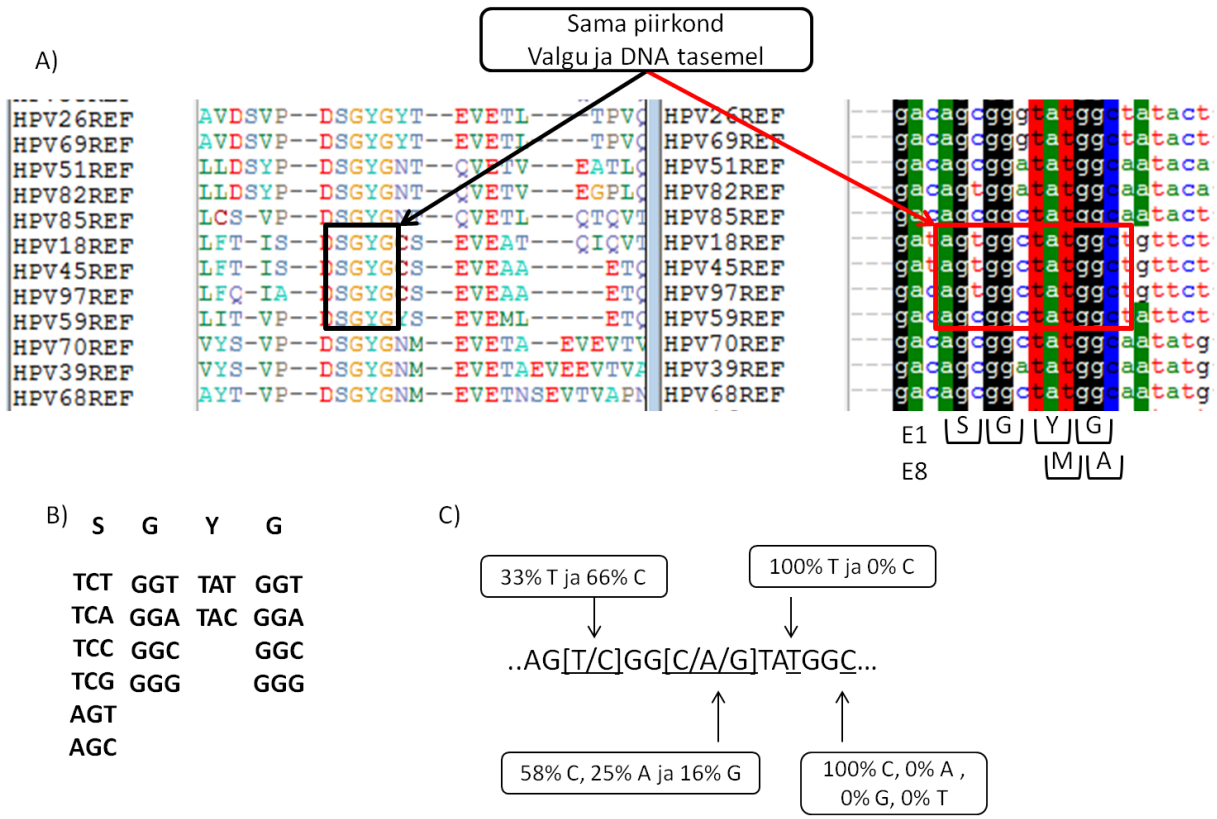


Lisa 1. Papilloomiviiruste fülogeneetiline puu. Roheliste ringidega on märgitud viirused, kellel on varasemalt eksperimentaalselt leitud E8^E2C, punaste ringidega viirused, kellel antud töö käigus ei tuvastanud E8 CDS-i ning sinisega need viirused, kellel ei olnud annoteeritud E1^E4 PaVe andmebaasis. [http://pave.niaid.nih.gov/#prototypes?type=tree] (24.04.14)

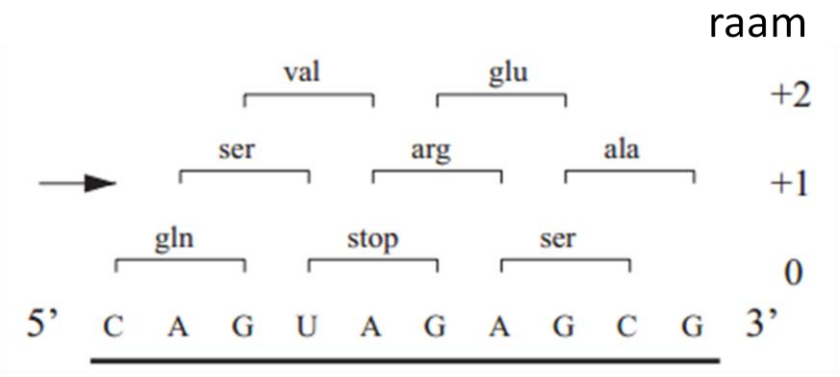
Lisa 2. Papilloomiviiruste suuremad perekonnad, nende peremeesorganism ja koe nakatamise spetsiifika*.

Perekond	Peremeesorganism	Koe spetsiifika
Alfa	inimene ja primaat	limaskest ja nahk
Beeta	Inimene ja primaat	nahk
Gamma	inimene	nahk
Mu	inimene	nahk
Nu	inimene	Nahk
Delta	kabjalised	nahk
Xi	veised	limaskest ja nahk
Epsilon	veised	nahk
Zeta	hobused	nahk
Dyorho	hobused	nahk
Phi	kits	nahk
Lambda	erinevad loomad	limaskest ja nahk
Chi	koerad	nahk ja limaskest
Tau	koerad	Nahk
Sigma	Urson e Põhja-Ameerika okassiga	nahk
Kappa	küülik	limaskest ja nahk
Iota	närilised	nahk
Pi	hamster	limaskest
Eta	linnud	nahk
Teeta	linnud	nahk
Omikron	vaalalised	limaskest
Upsilon	delfiinilised	limaskest
Dyozeeta	merikilpkonnad	nahk

* (de Villiers, Fauquet, Broker, Bernard, & zur Hausen, 2004; Gottschling et al., 2011; Herbst et al., 2009; Rector et al., 2005; Van Doorslaer, Rector, Vos, & Van Ranst, 2006; Yuan et al., 2007, 2012).

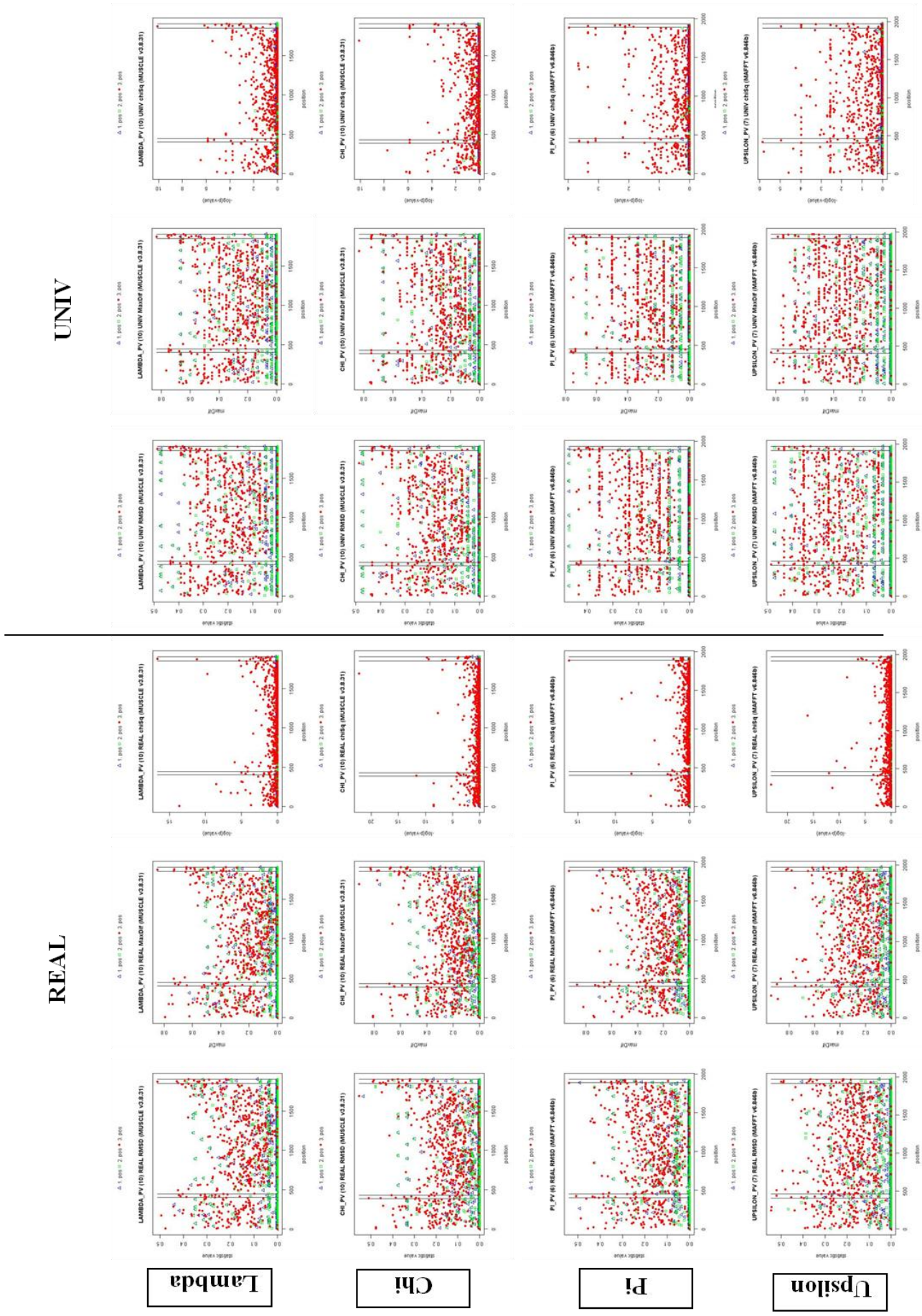


Lisa 3. Alfapapilloomiviiruste E1 valgu ning DNA joondus E8 ORF-i alguses. (A) E1 valgu joondus vasakul ning paremal sellele vastav nukleotiidne joondus. E1 valgu joondus tehtud MUSCLE v3.8.31 ning valgu joonduse põhjal tehtud DNA joondus PAL2NAL-iga**. (B)** Selekteeritud regioonis leiduvatele aminohapetele vastavad koodonid. **(C)** Antud regioonis kasutatud sünonüümsete koodonite kolmanda positsiooni nukleotiidide sagedused. Antud jooniselt on näha, et enne E8 lugemisraami kasutab E1 glütsiini (G) kodeerimiseks kolme koodonit, kuid E8 alguses ehk teise glütsiini jaoks ainult ühte (GGC)– toimub positiivne valik kindla koodoni osas.

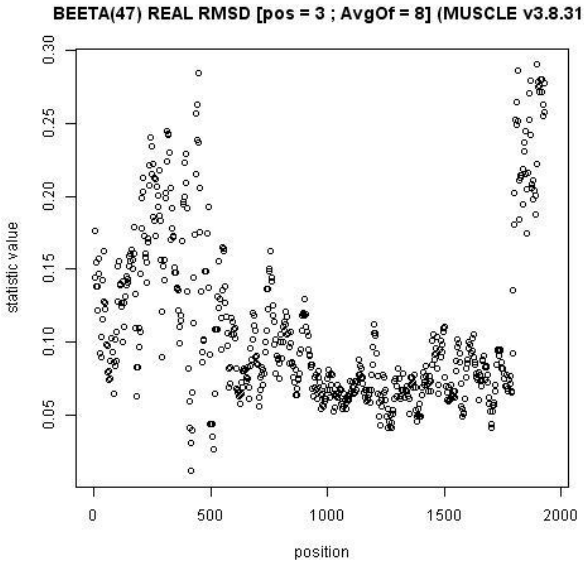


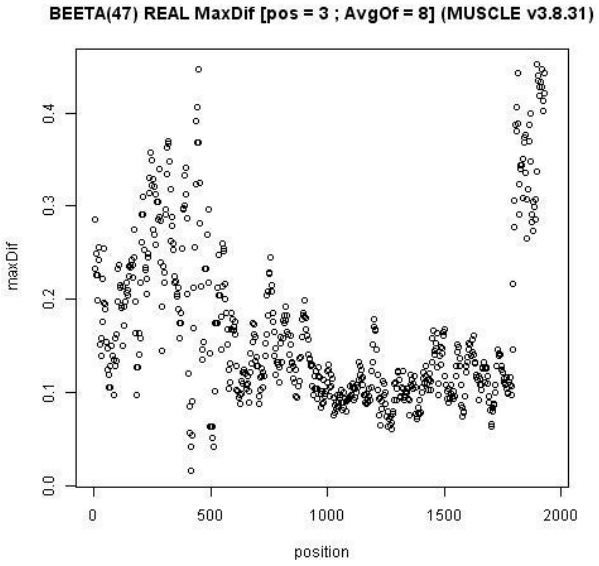
Lisa 4. Lugemisraamide nimetamine antud töös. Peamine lugemisraam on 0. Liikudes DNA järjestusel edasi ühe nukleotiidi võrra ehk alustades triplettide lugemist peamise lugemisraami esimese koodoni teisest positsioonist, saame +1 raami. Alustades lugemist peamise lugemisraami esimese koodoni kolmandast positsioonist saame +2 raami (Firth & Brown, 2005).

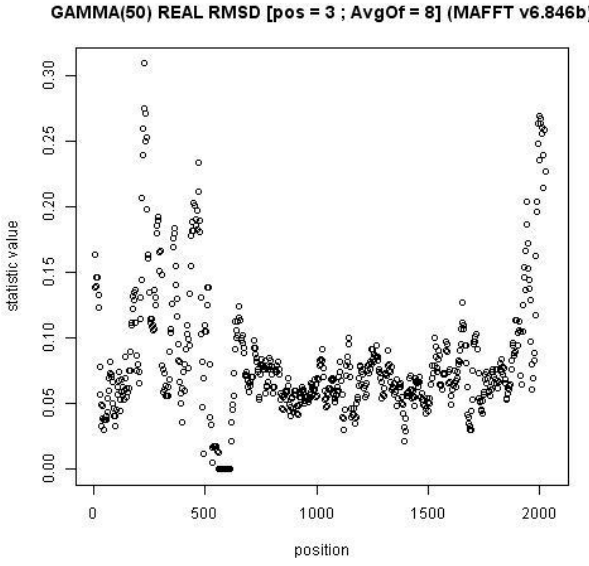
REAL
UNIV
Beeta
Gamma
Delta
Xi

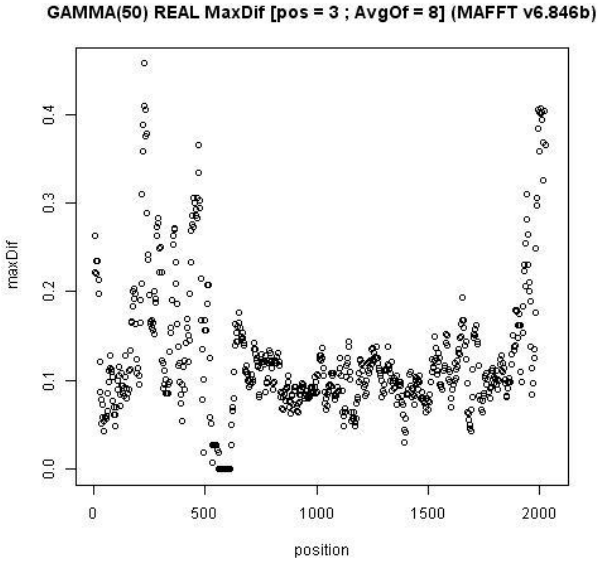


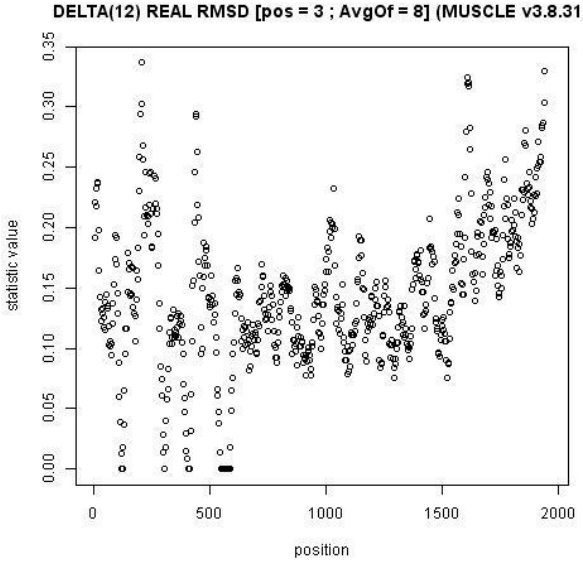
Lisa 5. E1 ORF-ide analüüs CDEP meetodiga. Vaatluse all suuremad PV perekonnad

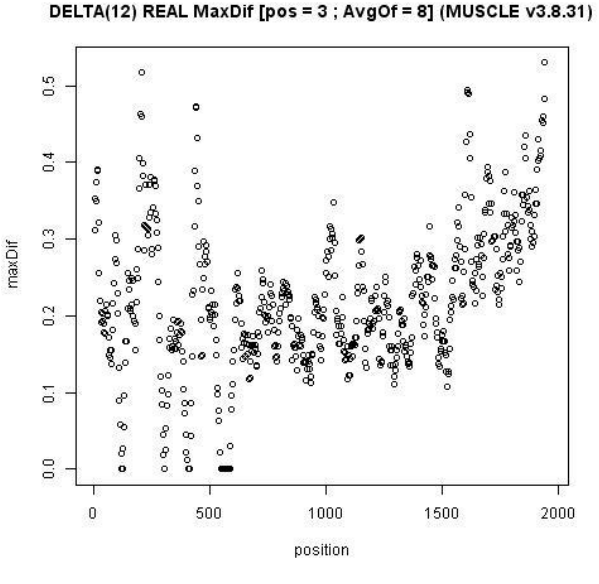

BEETA(47) REAL RMSD [pos = 3 ; AvgOf = 8] (MUSCLE v3.8.31)
statistic value
position

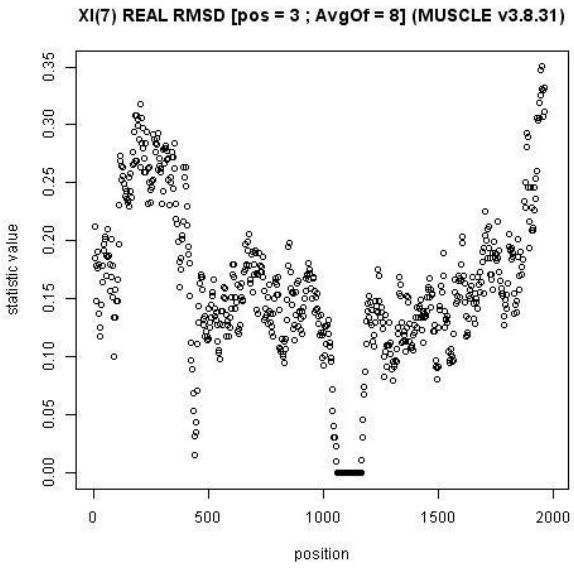

BEETA(47) REAL MaxDif [pos = 3 ; AvgOf = 8] (MUSCLE v3.8.31)
maxDif
position

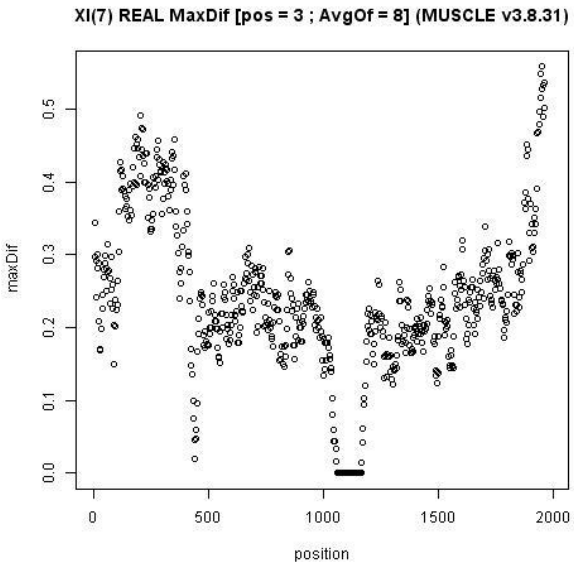

GAMMA(50) REAL RMSD [pos = 3 ; AvgOf = 8] (MAFFT v6.846b)
statistic value
position

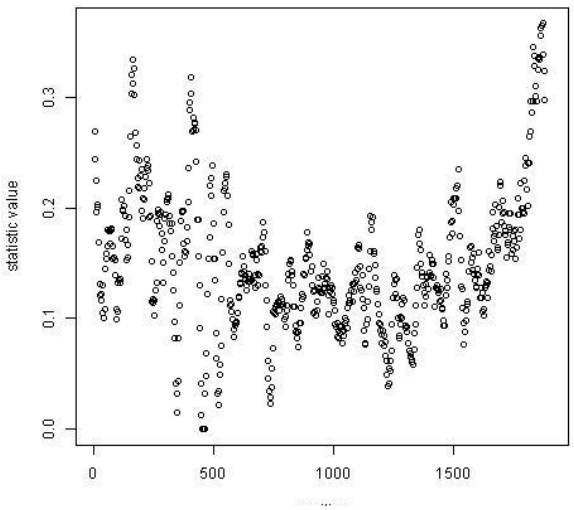

GAMMA(50) REAL MaxDif [pos = 3 ; AvgOf = 8] (MAFFT v6.846b)
maxDif
position

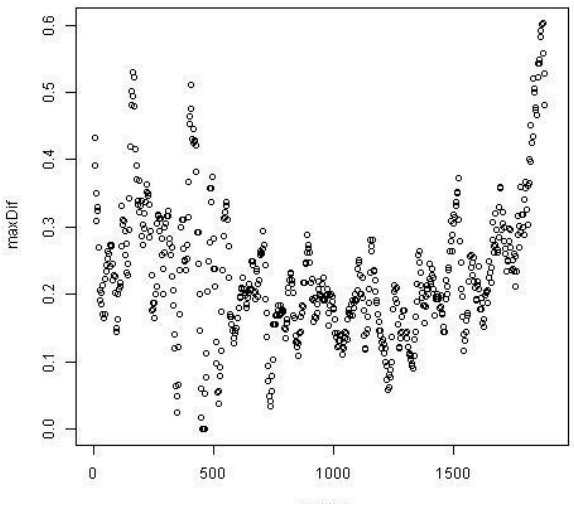

DELTA(12) REAL RMSD [pos = 3 ; AvgOf = 8] (MUSCLE v3.8.31)
statistic value
position

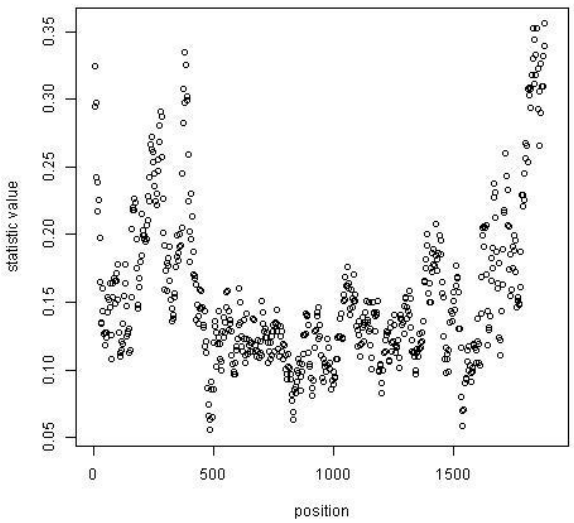

DELTA(12) REAL MaxDif [pos = 3 ; AvgOf = 8] (MUSCLE v3.8.31)
maxDif
position

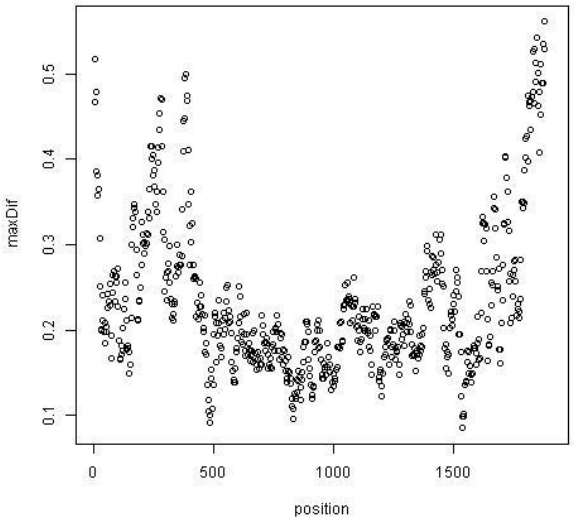

XI(7) REAL RMSD [pos = 3 ; AvgOf = 8] (MUSCLE v3.8.31)
statistic value
position

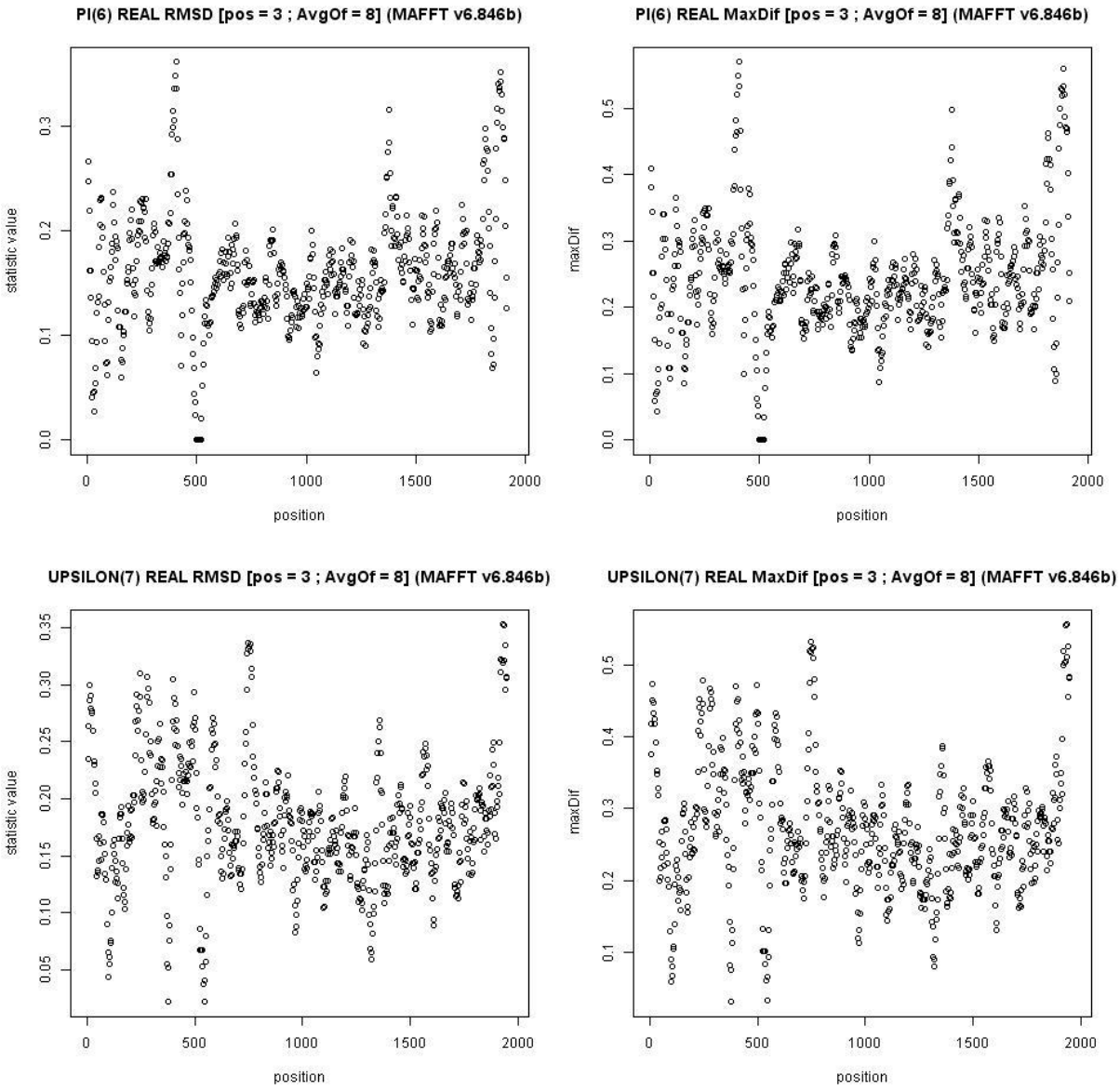

XI(7) REAL MaxDif [pos = 3 ; AvgOf = 8] (MUSCLE v3.8.31)
maxDif
position


LAMBDA(10) REAL RMSD [pos = 3 ; AvgOf = 8] (MUSCLE v3.8.31)
statistic value


LAMBDA(10) REAL MaxDif [pos = 3 ; AvgOf = 8] (MUSCLE v3.8.31)
maxDif


CHI(10) REAL RMSD [pos = 3 ; AvgOf = 8] (MUSCLE v3.8.31)
statistic value
position


CHI(10) REAL MaxDif [pos = 3 ; AvgOf = 8] (MUSCLE v3.8.31)
maxDif
position



Lisa 6. Suuremate papilloomiviiruste perekondade analüüs CDEP meetodiga kasutades RMSD ja MaxDif statistikut kaheksase libiseva aknaga. Graafiku y-telg tähistab 8 statistiku aritmeetilist keskmist ning x-telg positsiooni joondusel, statistikute keskväärtus märgitakse libiseva akna esimesele positsioonile. PV perekonnanime taga sulgudes märgitud analüüsis olnud E1 järjestuste arv. Kõigil graafikutel on näha signaali E1 lõpus, kus asub E1 ja E2 ORF-i ülekate. Üldjuhul on olemas ka signaal E1 alguses, kus asub E1^E4 splaissingu doonorsait. Suuremal osal perekondadest on tuvastatav ka E8 asukoht.

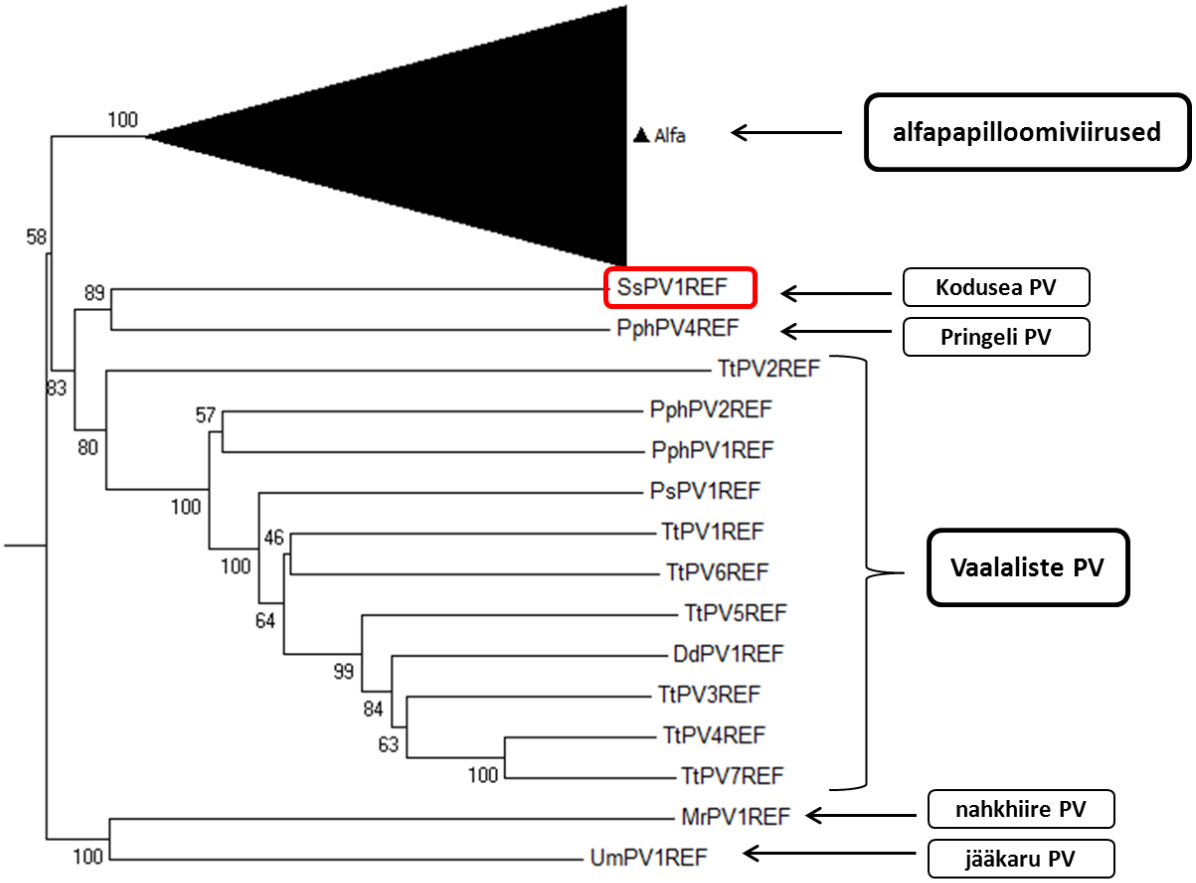
Lisa 7. Antud töö raames ennustatud E8 CDS-i ATG ja splaissaidi asukohtade erinevused võrreldes PaVe andmebaasis annoteeritud E8-ga.

PV	E8 ATG	Pave E8 ATG	E8 splaissait	Pave E8 splaissait	Leitud E8 CDS
BPV10REF	+	+	1053	1044	MTKIFLERWF
BPV11REF	+	+	1346	1337	MKIFLQRWR
BPV12REF	+	+	1154	1145	MLKRLLKRWY
BPV1REF	1204	1270	1235	1451	MKLTVFLRPS
BPV3REF	+	+	1090	1081	MKIFLQRWL
BPV4REF	+	+	1346	1340	MKIFLQKWY
BPV9REF	+	+	1426	1414	MKIFLERWLL
CgPV2REF	+	+	1131	1125	MKLKIFLLRW
EdPV1REF	-	1265	-	1347	-
FcaPV3REF	+	+	1140	1128	MKLKILLYRGYKQ
HPV100REF	+	+	1333	1327	MKLKILLKRW
HPV101REF	+	+	685	676	MKLKILLRRWQ
HPV102REF	+	+	1155	1128	MAIRRWKLEHWKA
HPV103REF	+	+	694	685	MSLKILMLRWK
HPV104REF	+	+	1327	1321	MKLKIFLKRW
HPV105REF	+	+	1331	1325	MKLKMFLLRW
HPV107REF	+	+	1401	1395	MKLKILLKRW
HPV10REF	+	+	1243	1216	MAKHRWIRNRDQN
HPV110REF	+	+	1278	1269	MKLKIILRHWRWR
HPV111REF	+	+	1253	1244	MKLKIILKHRWK
HPV113REF	+	+	1223	1214	MKLKIFLKHRWK
HPV114REF	1321	1255	1361	1334	MAIRRWRLEHPKD
HPV117REF	+	+	1241	1214	MAKHRWIRNRDQH
HPV122REF	+	+	1101	1092	MKLKIILKHRWK
HPV125REF	+	+	1145	1118	MAKHRWIRNRDHN
HPV12REF	+	+	1353	1347	MKLKMFLLRW
HPV131REF	+	+	1064	1058	MKLKILMY
HPV132REF	+	+	1059	1053	MKLQILMY

HPV143REF	+	+	1157	1151	MRLKMLLLRW
HPV14REF	+	+	1400	1394	MKLKMFLLRW
HPV150REF	+	+	1445	1439	MSLKILLQRW
HPV151REF	+	+	1336	1330	MKLKILLKRW
HPV153REF	+	+	1106	1097	MKLKVLTKYF
HPV160REF	+	+	1254	1227	MAKHRWIRNRDQA
HPV163REF	+	+	1295	1259	MIVACVKMKLKILLR
HPV17REF	+	+	1315	1306	MKLKILLSRLWRCR
HPV19REF	+	+	1377	1371	MKLKMFLLRW
HPV21REF	+	+	1387	1381	MKKMFLLRW
HPV22REF	+	+	1342	1336	MKLKILLKRW
HPV25REF	+	+	1362	1356	MKLKMFLLRW
HPV28REF	+	+	1240	1213	MAKHRWIRNRDHC
HPV29REF	+	+	1249	1222	MAKHRWIRTRDQA
HPV31REF	+	+	1296	1290	MAILKWKRSRWY
HPV33REF	+	+	1316	1307	MAILKWKLSRWYN
HPV36REF	+	+	1350	1344	MKLKMLLLRW
HPV37REF	+	+	1312	1303	MKLKILLSRLWRCR
HPV39REF	1331	1223	1368	1233	MAIWKWKQLKWR
HPV3REF	+	+	1249	1222	MAKHRWIRNRDQN
HPV40REF	1283	1208	+	+	MAILKWKQQRD
HPV42REF	1238	1100	1272	1110	MAILKWKYSRH
HPV47REF	+	+	1358	1352	MKLKMLLLRW
HPV59REF	1272	1218	1306	1234	MAILKWKCSRL
HPV5REF	+	+	1359	1353	MKLKMLLLRW
HPV62REF	+	+	1165	1138	MAIRRWIRDHPKD
HPV67REF	+	+	1279	1273	MAILKWKLKRQWC
HPV72REF	+	+	1275	1248	MAIRRWILEHQKA
HPV77REF	+	+	1252	1225	MAKHRWIRTRDQA
HPV78REF	+	+	1245	1218	MAKHRWIRNRDQC
HPV81REF	+	+	1287	1260	MAIRRWILEHPKD
HPV83REF	+	+	1161	1134	MAIRRWKLEHQKV
HPV84REF	1109	1043	1149	1122	MAIRRWISEHPRV

HPV86REF	1106	1040	1146	1119	MAIRRWILERPKA
HPV87REF	1216	1150	1256	1229	MAIRRWILEHPKA
HPV89REF	+	+	1173	1146	MAIRRWIREHPKV
HPV8REF	+	+	1343	1337	MKLKMFLMRW
HPV93REF	+	+	1235	1217	MKLKILLHRWNKRR
HPV94REF	+	+	1237	1210	MAKHRWIRNRDQN
HPV96REF	+	+	1459	1453	MSVKILLRRW
HPV99REF	+	+	1327	1321	MKLRMFLQRW
HPV9REF	+	+	1321	1312	MKLKIFLKHRWK
MfPV1REF	+	+	1101	1095	MKKMFLLRW
MfPV2REF	+	+	1125	1119	MKLKIFLKRRW
OaPV1REF	+	+	1154	1142	MKLIVLLKSAS
OaPV2REF	+	+	1152	1140	MKLIVLLKSAS
PphPV1REF	+	+	1729	1711	MGIRKRAVQRRWWSQH
RnPV1REF	1628	1559	+	+	MKLKILLRRR
TtPV1REF	+	+	1264	1246	MGIHWKRQVEKRWWEQMS
TtPV3REF	+	+	1256	1238	MGIHWKLKVRNRWCAETY
TtPV5REF	+	+	1219	1201	MGIHWKLEVRRRWCAETH
TtPV6REF	+	+	1222	1201	MGIHWKTQVRNRYWTEVC

+ tähistab, et PaVe andmebaasis annoteeritud ja antud töö raames tuvastatud E8 ATG või splaissaidi asukohad on samad.



Lisa 8. E1 valkude joonduse põhjal loodud fülogeneetilise puu üks haru. Antud harus on näidatud, kus asuvad vaalaliste papilloomiviirused alfapapilloomiviiruste suhtes. Lisaks on ära märgitud ka nahkhiire ja jääkaru papilloomiviirus. Antud puult on näha, et kodusealt leitud PV klassifitseerus kokku kõikide teiste hetkel teadaolevate vaalaliste papilloomiviirustega. Nahkhiire ja jääkaru PV-d moodustavad eraldi väikese rühma. E1 valgujärjestused saadud PaVe andmebaasist, joondatud MUSCLE v3.8.31 ning joondusest kustutatud positsioonid, kus oli tühimikke üle 80%, kasutades jalview-d. Fülogeneetiline puu loodud CLUSTALW2-ga vaikesätetega (*Neighbour Joining* ja *bootstrap* = 100). Puu visualiseerimiseks kasutati programmi MEGA 6.06.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Mikk Puustusmaa
(*autori nimi*)
(sünnikuupäev: 17.03.1989)

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

"Papilloomiviirustes E8 valku kodeeriva ala tuvastamine *in silico*",
(*lõputöö pealkiri*)

mille juhendaja on vanemteadur Aare Abroi ning kaasjuhendaja professor Maido Remm,
(*juhendaja nimi*)

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

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu alates **01.07.2015** kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

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

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

Tartus, 26.05.2014 (*kuupäev*)