Vilniaus universitetas
Gyvybės mokslų centras
Biomokslų institutas

Mikrobiologijos ir biotechnologijos magistrantūros studijų programos
II kurso studentė

Justina Kraujūnaitė

CRISPR-Cas sistemų baltymų homologai virusuose: paieška ir analizė bioinformatiniais metodais

Magistro baigiamasis darbas

Darbo vadovas:
dr. Darius Kazlauskas

Vilnius
2020

TURINYS

ĮVADAS

Bakterijos ir archėjos dažnai sąveikauja su egzogeninėmis nukleorūgštimis. Viena vertus, tai joms leidžia į savo genomą inkorporuoti įvairią genetinę medžiagą, tokią kaip plazmidžių koduojamus antibiotikų atsparumo genus ar fagų užkoduotus virulentiškumo faktorius. Kita vertus, tai paverčia ląsteles jautrias parazitiniams invaziniams agentams, tokiems kaip virusai ar plazmidės, kurie kenkia populiacijos tankumui. Siekdamos išvengti pavojų, bakterijos ir archėjos evoliucijos eigoje išvystė kelias apsaugos strategijas, skirtas reguliuoti svetimų nukleorūgščių patekimą į ląsteles (McGinn ir Marraffini, 2019). Tarp šių strategijų yra įvairūs įgimto imuniteto mechanizmai, pavyzdžiui, receptorių maskavimas, restrikcijos-modifikacijos (RM) bei bakteriofagų išskyrimo (BREX) sistemos (Marraffini, 2015; Mohanraju *et al*., 2016). Prie įgimto imuniteto mechanizmų, taip pat, prisideda ir įgytas imunitetas, kurio funkcijas prokariotuose atlieka tik CRISPR-Cas sistemos, teikiančios greitą bei efektyvią adaptaciją prie sparčiai besivystančių judriųjų genomo elementų (JGE) (Marraffini, 2015).

Vis dėlto, sąryšiai tarp CRISPR-Cas sistemų ir JGE yra kur kas sudėtingesni. Kelios JGE klasės ne tik prisidėjo prie CRISPR-Cas kilmės bei evoliucijos, bet ir atvirkščiai, kai kurie JGE pasisavino CRISPR-Cas sistemas bei atskirus jų komponentus funkcijoms, kurios vis dar tebėra menkai ištirtos ir pagrįstos tik ribotais, iš esmės, netiesioginiais įrodymais (Faure et al., 2019a; Faure *et al.*, 2019b). Tad JGE koduojamų CRISPR-Cas sistemų bei jų komponentų tyrimai svarbūs tuo, jog leidžia geriau suprasti CRISPR-Cas funkcinį plastiškumą bei JGE ir šeimininko koevoliucijos kelius (Faure *et al.*, 2019b).

Tarp JGE išskirtiniais laikomi virusai, priklausantys gausiausiai bei mažiausiai ištirtai biologinei formai planetoje (Paez-Espino *et al.*, 2016). CRISPR-Cas sistemų egzistavimas jau ne kartą buvo įrodytas virusų genomuose (Al-Shayeb *et al.*, 2019; Seed *et al.*, 2013), tačiau apie atskirus CRISPR-Cas sistemų komponentus duomenų yra labai mažai (Faure *et al.*, 2019b). Būtent dėl pastarosios priežasties, šiame darbe nuspręsta atlikti plataus masto ir didelio jautrumo CRISPR-Cas sistemų baltymų paiešką virusuose.

Darbo **tikslas** – atlikti CRISPR-Cas sistemų baltymų homologų paiešką bei analizę virusuose, taikant bioinformatinius metodus.

Darbo **uždaviniai**:

1) atlikti CRISPR-Cas sistemų baltymų homologų paiešką virusuose, naudojant HMMER ir HH-suite programinių įrangų paketus;

2) sugrupuoti bei išanalizuoti virusuose rastas CRISPR-Cas sistemų baltymų homologų sekas.

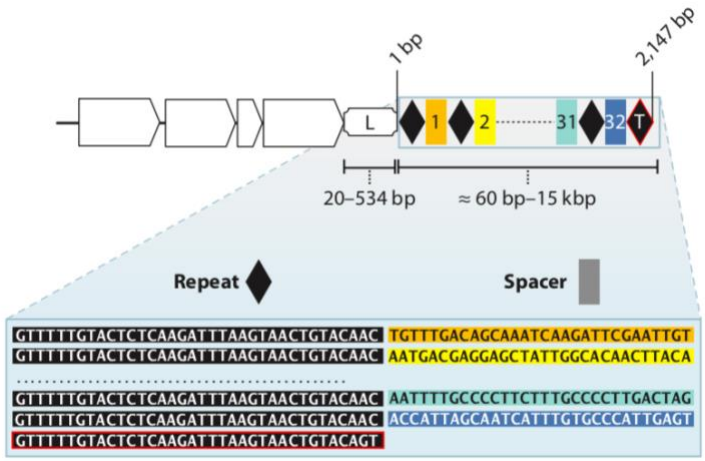
1. LITERATŪROS APŽVALGA

1.1. CRISPR-Cas sistemos prokariotuose

Pirmą kartą CRISPR lokusas buvo nustatytas 1987 m., sekvenuojant *Eschericia coli* genus (Ishino *et al*., 1987), tačiau pats akronimas sugalvotas tik 2002 m., po to, kai panašios struktūros buvo pastebėtos įvairių bakterijų bei archėjų genomuose (Jansen *et al*., 2002). Pastaraisiais metais identifikuota ir su CRISPR sritimi susijusi genų šeima – Cas genai (Jansen *et al*., 2002), o 2005 m. rasta, kad tam tikros CRISPR dalys (skirtukai) atitinka sekas, esančias virusuose ir plazmidėse (Bolotin *et al.*, 2005; Mojica et al., 2005; Pourcel *et al*., 2005), bei tai, jog kuo daugiau skirtukų aptinkama *Streptococcus thermophilus* kamienuose, tuo mažiau fagų geba juos infekuoti (Bolotin *et al.*, 2005). Nuo tada stebimas itin didelis susidomėjimas šiuo prokariotų adaptyvaus imuniteto mechanizmu, kuris dabar geriausiai žinomas kaip naujosios kartos genomo inžinerijos įrankis (Koonin ir Makarova, 2019).

1.1.1. CRISPR-Cas lokuso struktūra

CRISPR-Cas sistemos yra hipervariabilios bei tarp genomų skiriasi paplitimu, genais, sekomis, skaičiumi ir dydžiu (Barrangou ir Marraffini, 2014). CRISPR sritis susideda iš trumpų 21-48 bazių porų (bp) ilgio palindrominių DNR pasikartojimų, pertrauktų unikaliomis 26-72 bp ilgio DNR sekomis, vadinamomis skirtukais (angl. spacers) (1 pav.) (Deveau *et al.*, 2010). Paprastai tariant, CRISPR-Cas sistemos šeimininko DNR vadinama pasikartojimu, o svetima nukleorūgštis – skirtuku (Glemžaitė, 2014). Pasikartojimų bei skirtukų skaičius varijuoja tarp bakterijų ar archėjų tos pačios rūšies skirtingų kamienų. Maždaug 50 % CRISPR turinčių genomų aptinkamas daugiau nei vienas CRISPR lokusas. Viename CRISPR lokuse gali būti nuo 2 iki 375 pasikartojimų. Aukščiau CRISPR masyvo genome yra 20-534 bp ilgio lyderinė seka, turinti promotorių, nuo kurio transkribuojama CRISPR sritis. Šalia CRISPR lokuso visada randami ir Cas (angl. CRISPR-associated) genai (nuo 4 iki 20), kurių koduojami baltymai dažnai turi funkcinius domenus, būdingus nukleazėms, helikazėms, polimerazėms bei kitiems su nukleorūgštimis sąveikaujantiems fermentams (Deveau *et al.*, 2010; Horvath ir Barrangou, 2010).



1 pav. CRISPR sritis. Baltos rodyklės vaizduoja Cas genus. Galimi lyderinės sekos (L) bei CRISPR lokuso ilgiai nurodyti po atitinkamomis genomo dalimis. Juodi rombai žymi pasikartojimų sekas, o spalvoti stačiakampiai – skirtingų skirtukų sekas (Deveau et al., 2010).

Taigi, štai kodėl pilnas CRISPR-Cas sistemos pavadinimas yra taisyklingai pertraukti trumpi susitelkę palindrominiai pasikartojimai bei su jais susiję Cas genai (angl. clustered regularly interspaced short palindromic repeats and CRISPR-associated genes) (Glemžaitė, 2014).

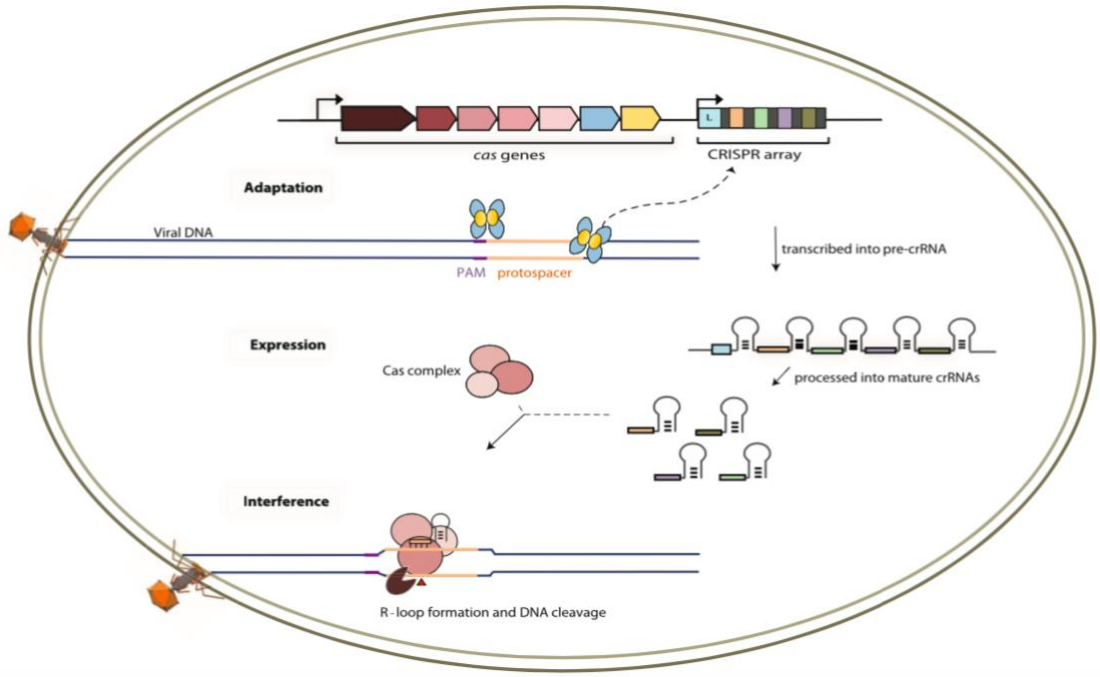
1.1.2. CRISPR-Cas imuninis atsakas

CRISPR-Cas imuninis atsakas susideda iš trijų pagrindinių etapų: adaptacijos (skirtuko įgijimo), ekspresijos (crRNR brendimo) bei interferencijos (2 pav.) (Mohanraju *et al*., 2016). Pirmajame etape veikia adaptacijos (integracijos) kompleksas, kuris formuojamas iš Cas baltymų. Šis kompleksas dažniausiai atpažįsta trumpą (2-4 bp ilgio) DNR taikinio motyvą – PAM (angl. protospacer-adjacent motif), šalia kurio iškerpa fragmentą svetimos nukleorūgšties – protoskirtuką (angl. protospacer) – bei įterpia jį kaip skirtuką CRISPR masyvo 5’ gale, žemiau lyderinės sekos (Faure *et al.*, 2019a). Kai kuriose CRISPR-Cas sistemose naudojamas alternatyvus adaptacijos mechanizmas, t.y. skirtuko įgijimas iš RNR per atvirkštinės transkripcijos kelią, pasitelkiant CRISPR-Cas regione koduojamą atvirkštinę transkriptazę (Makarova *et al.*, 2020). Taigi, skirtuko įgijimo etapas apima tris svarbiausius procesus: substrato „pagavimo“, CRISPR lokuso atpažinimo bei integracijos į masyvą (Jackson *et al*., 2017). Pastarasis etapas ir yra tai, kas paverčia CRISPR-Cas sistemas įgyto imuniteto forma (Faure *et al.*, 2019a).

Sekančiame, ekspresijos etape, CRISPR sritis transkribuojama kaip vienas transkriptas, gaunama ilga viengrandė RNR – prekursorinė CRISPR RNR (pre-crRNR). Ši pre-crRNR procesuojama į mažas, subrendusias CRISPR RNR (crRNR) (Faure *et al.*, 2019a). Kiekviena crRNR susideda iš skirtuko bei gretimų pasikartojimų, kurie palindrominėse vietose formuoja

antrinės struktūros elementus (plaukų segtukus) (Mohanraju *et al.*, 2016). Skirtinguose CRISPR-Cas sistemų variantuose, pre-crRNR brendimas medijuojamas iš keleto Cas baltymų sudarytų kompleksų, pavienio, multidomeninio Cas baltymo arba ne-Cas ribonukleazių, pavyzdžiui, RNazės III (Makarova *et al.*, 2020).

Paskutiniajame, interferencijos etape, crRNR, kuri paprastai lieka susijungusi su apdorojimo (angl. processing) kompleksu (baltymais), tarnauja kaip kelrodė atpažįstant protoskirtuko (ar labai panašios sekos) komplementarumą invaziniame svetimame genome. Pastarasis tuomet yra suskaidomas ir inaktyvuojamas Cas nukleazės (pavyzdžiui, Cas9 ar Cas3), kuri arba jau yra to paties efektorinio komplekso dalis, arba įgijama interferencijos etape (Makarova *et al.*, 2020).



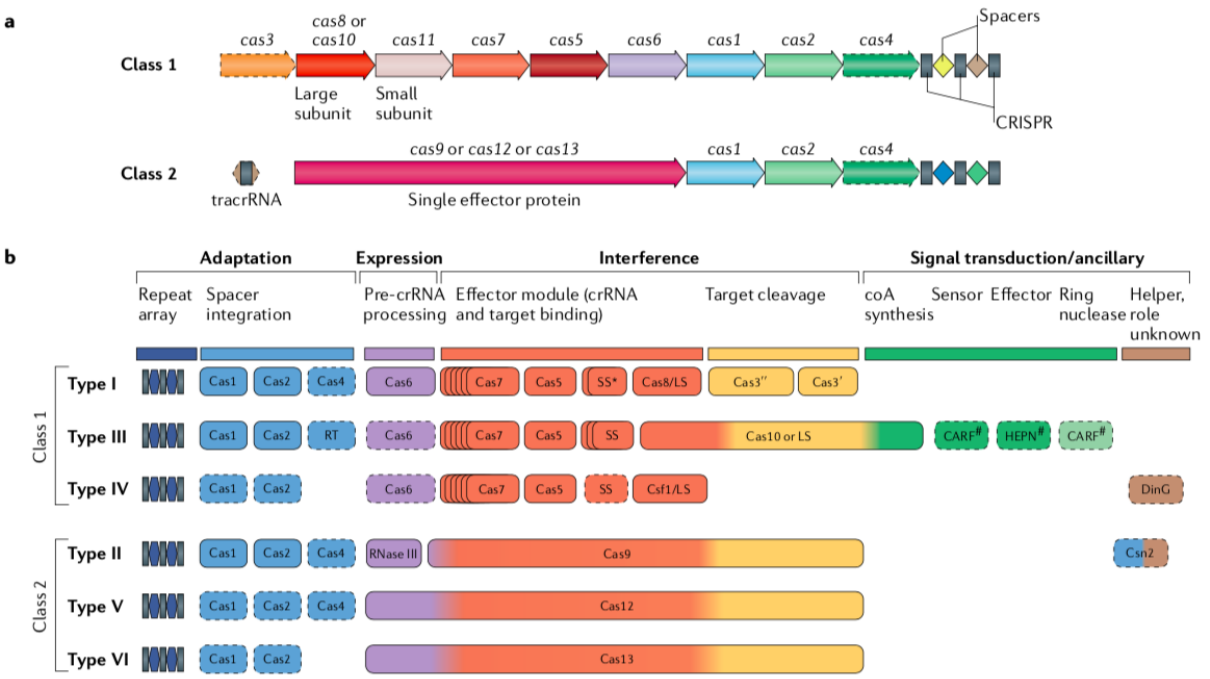
2 pav. CRISPR-Cas sistemos veikimo principas. CRISPR-Cas imuniniame atsake svarbiausią vaidmenį atlieka CRISPR RNR (crRNR) bei Cas baltymai, kartu formuodami daugiakomponenčius CRISPR ribonukleoproteinų (crRNP) kompleksus. Pirmasis etapas – adaptacija – vyksta judriojo genomo elemento (šiuo atveju viruso) invazijos metu. Cas1 (mėlyna) ir Cas2 (geltona) baltymai invaziniame genome atpažįsta PAM seką bei įterpia protoskirtuką (oranžinė), kaip naują skirtuką, į CRISPR masyvą, šalia lyderinės (L) sekos. Sekančiame, ekspresijos etape, CRISPR lokusas transkribuojamas, o pre-crRNR procesuojama į subrendusias crRNR, medijuojant Cas arba ne-Cas baltymams. Paskutiniajame, interferencijos etape, Cas nukleazės degraduoja invazinę DNR, po to, kai crRNR-Cas kompleksas atpažįsta komplementarų taikinį svetimoje nukleorūgštyje (Mohanraju *et al.*, 2016).

Taigi, CRISPR-Cas sistema yra ypatinga tuo, jog sukaupia informaciją apie prokariotą atakavusius invazinius genomo elementus DNR sekoje bei vėliau, tam pačiam egzogeniniam agentui nusitaikius į tą patį mikroorganizmą, ši informacija panaudojama tai svetimai nukleorūgščiai sunaikinti ir mikroorganizmui apsaugoti (Glemžaitė, 2014).

1.1.3. CRISPR-Cas klasifikacija

CRISPR-Cas sistemos pasižymi nepaprastu molekulinės struktūros sudėtingumu ir įvairove (Faure *et al.*, 2019a). Pastovi klasifikavimo schema yra būtina CRISPR-Cas lokusų tiksliam charakterizavimui naujuose genomuose bei tolimesnei CRISPR tyrimų pažangai (Makarova *et al.*, 2015). Naujausia, 2020 m. CRISPR-Cas sistemų klasifikacija apima 1 ir 2 klases, kurios atitinkamai skaidomos į I, III, IV bei II, V, VI tipus, o pastarieji dar smulkiau grupuojami į 33 potipius (Makarova *et al.*, 2020). Vieni mikroorganizmų kamienai savo genome gali turėti kelias CRISPR-Cas sistemas, priklausančias tam pačiam tipui, kiti – priklausančias keliems skirtingiems tipams (Jackson *et al.*, 2017).

Randamos maždaug 45 % bakterijų ir 85 % archėjų genomų, CRISPR-Cas sistemos klasifikuojamos pagal Cas genus (McGinn ir Marraffini, 2019). Visi Cas genai gali būti suskirstyti į keturis skirtingus, dalinai persidengiančius, funkcinius modulius (3 pav.). Adaptacijos modulis apima geną, koduojantį pagrindinį skirtuko įgijimo fermentą (Cas1 integrazę) bei struktūrinį adaptacinio komplekso subvienetą Cas2, o kai kuriuose CRISPR-Cas sistemų variantuose dar ir Cas4 nukleazę, Csn2 baltymą ar atvirkštinę transkriptazę. Ekspresijos modulis atsakingas už pre-crRNR apdorojimą, kurį daugumoje 1 klasės sistemų vykdo Cas6 fermentas. Tuo tarpu, 2 klasės sistemose, apdorojimą katalizuoja ne Cas baltymas arba didelio efektorinio Cas baltymo katalizinis centras. Interferencijos, kitaip tariant, efektorinis modulis, susijęs su taikinio atpažinimu bei jo skaidymu. 1 klasės CRISPR-Cas sistemose, efektorinis modulis susideda iš keleto Cas baltymų (Cas3, Cas5-Cas8, Cas10 ir Cas11), kurie kartu su crRNR formuoja kompleksą, dalyvaujantį interferencijos etape. Šių baltymų išsidėstymas skiriasi, priklausomai nuo tipo (3 pav.). Priešingai 1 klasei, 2 klasės sistemose efektorinį modulį reprezentuoja atskiras, didelis, multidomeninis baltymas (Cas9, Cas12 ar Cas13), kuris yra funkciškai analogiškas visam 1 klasės efektoriniam kompleksui. Signalo perdavimo (pagalbinis) modulis sudarytas iš su CRISPR susijusių genų, kurių daugumos funkcijos, šiuo metu, yra tik preliminarios. Išimtimi laikomos tik III tipo sistemos, kurių signalo perdavimo kelias yra pilnai charakterizuotas (Makarova *et al.*, 2020). Šiame kelyje Csm6/Csx1 HEPN (angl. higher eukaryotes and prokaryotes nucleotide-binding) RNazė aktyvuojama Cas10 polimerazės sintetinamo ciklinio oligoadenilato (coA), pastarajam susirišus su Csm6/Csx1 CARF (angl. CRISPR-associated Rossmann fold) domenu. Ko pasekoje įvyksta invazinės DNR transkripto degradacija (Kazlauskiene *et al.*, 2017).



3 pav. 1 ir 2 klasės CRISPR-Cas sistemos. **(a)** Bendra abiejų klasių organizacija. **(b)** CRISPR-Cas sistemų funkciniai moduliai. Schema iliustruoja genetinius, struktūrinius bei funkcinius ryšius tarp visų šešių CRISPR-Cas sistemų tipų. Baltymų pavadinimai atitinka dabartinę nomenklatūrą. Žvaigždutė nurodo tariamą mažąjį subvienetą, kuris gali būti sulietas su didžiuoju subvienetu, keliuose I tipo potipiuose. Grotelių simbolis (#) žymi, jog kitos nežinomos jutiklių, efektorių ir žiedo nukleazių baltymų šeimos gali būti susijusios su tuo pačiu signalo perdavimo keliu. Punktyriniais kontūrais pažymėti tie komponentai, kurie nėra būtini (ir/arba dingę), kai kuriuose tam tikro tipo potipiuose. Trys Cas9, Cas10, Cas12 ir Cas13 spalvos nurodo, jog šie baltymai prisideda prie skirtingų CRISPR-Cas atsako stadijų. CARF ir HEPN domenus turintys baltymai yra geriausiai charakterizuotos signalo perdavimo (pagalbinių) modulių struktūrinės dalys, tačiau egzistuoja ir jų alternatyvos. TracrRNA, transaktyvuojanti CRISPR RNR; RT, atvirkštinė transkriptazė; LS, didysis subvienetas; SS, mažasis subvienetas (Makarova *et al.*, 2020).

Du pagrindiniai prokariotų įgyto imuniteto sistemų komponentai yra adaptacijos bei efektorinis moduliai. Būtent pastarojo Cas baltymų kompozicijos skirtumais bei sekų divergencija, iš esmės, paremta CRISPR-Cas sistemų klasifikacija. 1 klasės efektoriniai kompleksai turi panašią baltymų organizaciją tarp visų trijų tipų. I tipą iš kitų, išskiria DNR skeliantis Cas3 baltymas, kuris, paprastai, sudarytas iš susijungusių helikazės bei HD (histidino-aspartato) nukleazės domenų. III tipas skiriasi tuo, jog HD nukleazė yra susiliejusi su Cas10, dideliu komplekso subvienetu, susijusiu su svetimos DNR transkripto skaidymu. IV tipo sistemos visai neturi nukleazių, reikalingų interferencijai (Makarova *et al.*, 2020).

Savo ruožtu, 2 klasės tipai skiriasi efektorinių didelių baltymų domenų architektūra, konkrečiai, pre-crRNR apdorojimo mechanizmu. VI tipo bei keliuose V tipo variantuose multidomeniniai baltymai apima ir RNazės aktyvumą, kai tuo tarpu, II tipo (taip pat, ir keliuose V tipo variantuose) apdorojimo aktyvumas perduotas ne Cas fermentui, o RNazei III. Pastaruoju atveju, efektorinis modulis apima ir papildomą RNR molekulę, transaktyvuojančią CRISPR (tracr) RNR, kuri formuoja stabilius dupleksus su iš dalies komplementariais pre-

crRNR pasikartojimais. Po šio RNR duplekso suskaidymo, kurį vykdo RNazė III, subrendusi kelrodė RNR, tiksliau, crRNR-tracrRNR kompleksas, lieka stabiliai susijungęs su efektoriniais baltymais, leisdamas vykti DNR interferencijai. Svarbu paminėti, jog VI tipo sistemų efektoriai unikaliai apima du HEPN RNazės domenus, įgalindami šį tipą būti apibūdintam kaip pirmąja bei iki šiol vienintele grupe, išskirtinai kerpančia tik RNR molekules (Makarova *et al.*, 2020).

1.2. CRISPR-Cas judriuosiuose genomo elementuose

Vis didėjantis įrodymų kiekis leidžia teigti, jog CRISPR-Cas sistemos dalyvauja ir kitose įvairiose ląstelių funkcijose, esančiose už prokariotų adaptyviojo imuniteto ribų. Manoma, jog šie „nekanoniniai“ aspektai apima CRISPR-Cas sistemų poveikį horizontaliosios genų pernašos ir genų ekspresijos reguliacijai bei siejasi su DNR taisymo, užprogramuotos ląstelės mirties ir, jau anksčiau minėtais, signalo perdavimo mechanizmais. Negana to, be būdingo egzistavimo bakterijų bei archėjų chromosomose, CRISPR-Cas sistemos ar jų sudedamosios dalys aptinkamos judriuosiuose genomo elementuose (JGE), įskaitant transpozonus, plazmides bei virusus (Faure *et al.*, 2019a). JGE koduojamų CRISPR-Cas sistemų bei jų komponentų tyrimai leidžia geriau suprasti CRISPR-Cas funkcinį plastiškumą bei JGE ir šeimininko koevoliucijos kelius (Faure *et al.*, 2019b).

1.2.1. CRISPR-Cas transpozonuose ir plazmidėse

Transpozonai – plati JGE klasė, plintanti įterpdama elementą į naujas šeimininko genomo vietas, naudojant tam tikrą fermentą – rekombinazę, dar įvairiai vadinamą transpozaze, integraze ar vietai specifine rekombinaze. CRISPR adaptacijos molekulinis mechanizmas atspindi vienos iš geriausiai ištirtų transpozazių klasių (DD[D/E]-transpozazių) biochemiją ir todėl manoma, jog išsivystė iš atskiros transpozonų grupės – kaspozonų – koduojančios kaspozazę (rekombinazę) homologišką Cas1 baltymui. Be Cas1, kai kurie kaspozonai koduoja ir kitą nukleazę, susijusią su adaptacijos procesu – Cas4. Kitas svarbus transpozonų indėlis į CRISPR-Cas evoliuciją, susijęs su 2 klasės II bei V tipo efektoriais, Cas9 ir Cas12, kurie, tikriausiai, išsivystė iš skirtingų TnpB baltymų, koduojamų IS605-tipo transpozonų (Faure *et al.*, 2019b). RuvC nukleazių domenus turintis TnpB baltymas, kartu su TnpA baltymu, dalyvauja IS605-tipo transpozonų transpozicijoje (Faure *et al.*, 2019b; He *et al.*, 2015). Beje, be transpozonų, manoma, jog CRISPR-Cas sistemų evoliucijai įtakos turėjo ir dar dvi atskiros JGE grupės. Tai II tipo intronai, kurie „perdavė“ atvirkštinę transkriptazę III tipo adaptacijos moduliui, bei mikrobinė toksino-antitoksino (TA) sistema, iš kurios kilo Cas2 baltymas (Koonin ir Makarova, 2019).

Judrieji genomo elementai ne tik prisidėjo prie CRISPR-Cas sistemų atsiradimo bei evoliucijos, bet ir priešingai, „pasisavino“ kai kuriuos sistemų variantus ar jų komponentus iš prokariotų. Geriausiai ištirtas atvejis yra „minimalių“ I bei V tipo CRISPR-Cas sistemų inkorporavimas į Tn7-tipo transpozonus. Šie „minimalūs“ variantai talpina visus komponentus, reikalingus pre-crRNR brendimui bei taikinio atpažinimui, tačiau stokoja adaptacijos modulio bei interferencijoje dalyvaujančios nukleazės. Pagal pasiūlytą hipotetinį veikimo mechanizmą, crRNR susirišimas su komplementaria fago ar plazmidės DNR seka, palengvina transpozono integraciją į atitinkamą vietą. Be to, IV tipo „minimalios“ CRISPR-Cas sistemos, aptinkamos bakterijų ekstrachromosominėse genetinėse struktūrose – plazmidėse, kur, galimai, prisideda prie tarp plazmidžių bei tarp plazmidžių ir virusų egzistuojančios konkurencijos (Faure *et al*., 2019b).

1.2.2. CRISPR-Cas virusuose

Dar viena atskira judriųjų genomo elementų grupė, be „minimalių“ CRISPR-Cas variantų, koduoja ir pilnas CRISPR-Cas sistemas bei tik atskirus jų komponentus. Šiai grupei priklauso submikroskopiniai neląstelinės sandaros obligatiniai viduląsteliniai parazitai, paprasčiau tariant, virusai (Faure *et al*., 2019b). Prokariotų virusuose aptinkamos CRISPR-Cas struktūros, funkcionuoja kaip anti-gynybiniai prietaisai, padedantys apeiti šeimininko apsaugos mechanizmus bei išvengti tarp-virusinių konfliktų (Faure *et al*., 2019a).

Pilnos CRISPR-Cas sistemos egzistavimas virusuose pirmą kartą buvo charakterizuotas 2013 m., kai K. D. Seed su kolegomis (Seed *et al*., 2013) ištyrė, jog *Vibrio cholerae* O1 bakteriofagas ICP1 koduoja I-F potipio CRISPR-Cas sistemos homologą, susidedantį iš 2 CRISPR masyvų bei 6 Cas genų. Šios ICP1 CRISPR-Cas sistemos taikinys yra chromosominės patogeniškumo salos, teikiančios bakterijai apsaugą nuo fagų (Naser *et al*., 2017; Seed *et al*., 2013; Villion ir Moineau, 2013). Tai reiškia, jog ICP1 koduojama CRISPR-Cas sistema yra būtina sėkmingai šio fago reprodukcijai *V. cholerae* bakterijose (Faure *et al*., 2019b).

Ilgą laiką ICP1 buvo vienintelis žinomas virusas, koduojantis pilną, funkciškai apibūdintą CRISPR-Cas sistemą (Faure *et al*., 2019b), tačiau 2019 m. B. Al-Shayeb ir bendraautorių (Al-Shayeb *et al*., 2019) atliktos analizės rezultatai atskleidė visų pagrindinių tipų (išskyrus VI) CRISPR-Cas lokusų egzistavimą dideliuose faguose (tai tokie fagai, kurių genomai ilgesni nei 200 kilobazių (kb)). Dauguma šių fagų koduojamų CRISPR-Cas sistemų stokoja adaptacijos modulio genų, o kai kuriose nėra ir efektorinės nukleazės geno, kas nurodo evoliucinę trajektoriją analogišką anksčiau aprašytoms transpozonų CRISPR-Cas sistemoms (Faure *et al*., 2019b). Nustatyta, jog didelių fagų koduojami CRISPR masyvai trumpesni

(mediana – 6 pasikartojimai viename masyve), nei tipiniai prokariotų masyvai, bei talpina skirtukus, kurių taikiniai yra arba kitų fagų genai, arba bakterijų genai, dalyvaujantys transkripcijoje ir transliacijoje. Tai pagrindžia faktą, jog virusai į savo genomą inkorporuoja CRISPR-Cas sistemas tiek tarp-virusinių konfliktų sprendimui, tiek šeimininko funkcijų manipuliavimui. Be to, svarbu paminėti, jog pastarojo tyrimo metu dideliuose virusuose nustatytas V-I potipio CRISPR-Cas sistemos efektorius Cas12i bei galimai nauji efektoriniai baltymai Casϕ (Cas12j) ir Cas14 (Cas14i, Cas14J, Cas14K) (Al-Shayeb *et al*., 2019).

Atsižvelgiant į retą pilnų ar tik iš dalies pilnų CRISPR-Cas sistemų inkorporavimą į virusų genomus, lyginant su plačiu paplitimu bakterijose bei archėjose, G. Faure ir kolegos (Faure *et al*., 2019b) nusprendė charakterizuoti potencialų atskirų CRISPR-Cas komponentų egzistavimą virusuose. Naudodami sukonstruotus Cas baltymų šeimų profilius, mokslininkai nustatė, jog didžioji dalis identifikuotų Cas genų virusuose priklauso Cas4 nukleazės homologams, iš kurių dauguma egzistuoja kaip solo (atskiri) genai, nesusiję su jokiais kitais Cas ar CRISPR lokusais (Faure *et al*., 2019b). Manoma, jog virusų koduojamas Cas4 veikia kaip anti-gynybinis prietaisas prieš šeimininko CRISPR-Cas sistemą (Hooton *et al*., 2016; Zhang *et al*., 2019). Archėjų virusuose Cas4 dažnai randamas šalia genų, koduojančių anti-CRISPR (Acr) baltymus (tai specialūs virusų baltymai, gebantys tiesiogiai prisijungti prie prokariotų CRISPR-Cas interferencijos mechanizmų bei juos inhibuoti (Koonin ir Makarova, 2018; Pawluk *et al*., 2018)). Tai rodo tikėtiną Cas4 bei Acr funkcinį ryšį (Faure *et al*., 2019b). Taip pat, dar vieno tyrimo rezultatai leidžia teigti, jog Cas4-tipo baltymas yra struktūrinis *Thermoproteus tenax virus 1* (TTV1) nukleokapsidės baltymas (Krupovic *et al*., 2015). Beje, anksčiau minėto darbo metu, G. Faure su kolegomis (Faure *et al*., 2019b), ieškodami atskirų CRISPR-Cas komponentų virusų genomuose, be Cas4 aptiko ir solo Cas2 bei Cas3 genus, kurių funkcijos virusuose dar nenustatytos.

Įdomu tai, jog Cas3 ir Cas4 baltymų homologų egzistavimas patvirtintas ir didelių, vienaląsčius eukariotus (amebas) infekuojančių *Mimiviridae* šeimos virusų CRISPR-Cas-tipo sistemoje – MIMIVIRE (angl. mimivirus virophage resistance element). Teigiama, jog ši sistema, susidedanti iš CRISPR masyvo bei šalia esančių genų, apsaugo kai kuriuos mimivirusų kamienus nuo juose parazituojančių Zamilon virofagų (Dou *et al*., 2018; Levasseur *et al*., 2016).

Taigi, CRISPR-Cas inkorporavimas į judriųjų genomo elementų genetinę struktūrą, tiksliai atitinka „ginklų nuomos” (angl. „guns for hire”) koncepciją, pagal kurią šeimininkų gynybos sistemos yra panaudojamos JGE puolimo veikloms vykdyti ir atvirkščiai. Apskritai, CRISPR-Cas komponentų įjungimas į JGE sudėtį yra tik dalis sudėtingo funkcinių ir

evoliucinių ryšių tinklo (Faure *et al.*, 2019b). Tad net neabejojama, jog ateinantys metai atskleis dar daugiau svarbių CRISPR-Cas bei JGE tarpusavio sąveikos detalių.

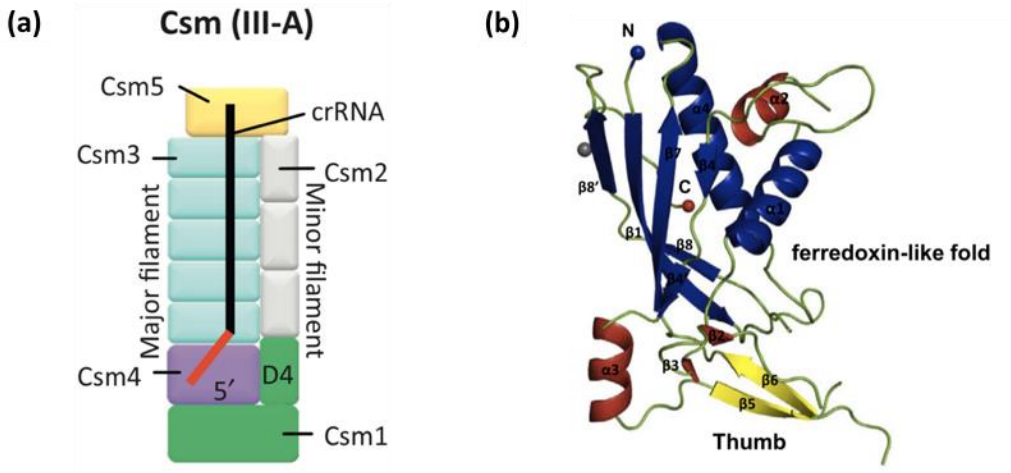
1.3. III tipo CRISPR-Cas sistema ir su ja susijęs Csm3 baltymas

Jei 2 klasės CRISPR-Cas sistemos pastaraisiais metais sulaukė itin didelio dėmesio dėl savo paprastumo, tai 1 klasės sistemos žavi mokslininkus savo sudėtingumu. Išskirtiniu laikomas III tipas, apimantis labai įvairius ir polimorfinius CRISPR-Cas lokusus bei pasižymintis aukšta potipių divergencija. Nors III tipo sistemos nėra gausiausiai paplitusios, tačiau jos laikomos CRISPR-Cas sistemų protėviu. Genomikos tyrimai parodė, jog pirmapradis Cas efektorinis modulis kilo iš Cas10-tipo baltymo – pagrindinio III tipo sistemų komponento (Pyenson ir Marraffini, 2017).

III tipo CRISPR-Cas sistemos smulkiau skirstomos į kelis potipius, iš kurių anksčiausiai aptikti bei geriausiai ištirti yra III-A ir III-B (Pyenson ir Marraffini, 2017). Pirminės analizės parodė, jog III-A efektorinio komplekso (Csm) taikinys yra DNR molekulės, tuo tarpu III-B efektorinio komplekso (Cmr) – RNR. Vis dėlto, vėliau atlikti tyrimai atskleidė, kad abu minėti III tipo kompleksai yra tiek RNazės, tiek nuo transkripcijos priklausomos DNR nukleazės, kas reiškia, jog iš pradžių jie atpažįsta taikinį per specifinę sąveiką tarp crRNR ir susidariusios komplementarios mRNR, o vėliau įvyksta DNR sekos skėlimas, vykdomas Cas10 (Csm1/Cmr2) subvieneto (Pyenson ir Marraffini, 2017; Tamulaitis *et al.*, 2017). Kitais žodžiais tariant, sėkmingam viruso infekcijos ciklui palaikyti, turi būti inicijuojama fago DNR transkripcija, todėl III-A ir III-B potipių vykdomas interferencijos etapas remiasi suderintu transkripto (RNR) bei transkribuojamos DNR skaidymu (Kazlauskiene *et al.*, 2017; Mohanraju *et al.*, 2016). Beje, svarbu pabrėžti, jog anksčiau tekste (1.1.3. poskyryje) minėtas III tipo CRISPR-Cas sistemų signalo perdavimo kelias, kuriame pagrindinį vaidmenį transkripto degradacijoje atlieka Csm6/Csx1 baltymas, tarnauja tik kaip atsarginis variantas, tuo atveju, jei Csm efektorinis kompleksas nesugeba atlikti savo funkcijos kovoje su egzogenine nukleorūgštimi (Kazlauskiene *et al.*, 2017).

III-A potipio CRISPR-Cas sistemose Csm ribonukleoproteinų kompleksas susideda iš penkių efektorinių Cas baltymų bei crRNR. Komplekso pagrindą sudaro dvi susipynusios spiralinių baltymų grandinės, kurių apačioje išsidėstęs didysis subvienetas – Csm1 (Cas10), o viršuje – Csm5 subvienetas. Didžioji grandinė formuojama iš vienos kopijos Csm4 ir kelių kopijų Csm3, tuo tarpu mažoji grandinė – iš kelių Csm2 baltymų bei didžiojo subvieneto C-galo domeno (D4) (4a pav.) (Tamulaitis *et al.*, 2017). Už RNazės aktyvumą šiame komplekse atsakingas Cas7 šeimos baltymas – Csm3 (Takeshita *et al.*, 2019). Zhao ir kolegos (Zhao *et al.*, 2018) nustatė, jog *Staphylococcus epidermidis* Csm3 yra vieno domeno baltymas, sudarytas iš

keturių α-spiralių ir aštuonių β-klosčių (4b pav.). Keturios β-klostės (β1, β4, β7, β8) bei dvi α-spiralės (α1, α4) formuoja kanoninę feredoksino-tipo raukšlę, gebančią susirišti su crRNR. Šios raukšlės antrinės struktūros elementus jungia skirtingo ilgio kilpos. Pagrindinį vaidmenį nukleorūgšties skaidyme vaidina β5 ir β6 klostės, kurios kartu sudaro plaukų segtuko struktūrą (Zhao *et al.*, 2018).



4 pav. (a) Schematinė III-A potipio CRISPR-Cas sistemų Csm komplekso architektūra. Raudona spalva pažymėta crRNR 5'- „rankena", potencialiai atsakinga už Csm1 nukleazės aktyvumą (Tamulaitis *et al.*, 2017). **(b)** Csm3 baltymo struktūra. N-galas pavaizduotas mėlynu rutuliuku, C-galas – raudonu rutuliuku. Baltymo šerdis – feredoksino-tipo raukšlė – pažymėta mėlynai. β-plaukų segtuko struktūra nuspalvinta geltonai. Tarp antrinės struktūros elementų esančios insercijos pavaizduotos raudona spalva. Pilku rutuliuku pažymėtas kalcio jonas. Antrinės struktūros elementai paženklinti skaitmenimis (Zhao *et al.*, 2018).

Csm3 baltymas ypač įdomus tuo, jog su jo galimais homologais eksperimentiškai dirba mūsų centro (Vilniaus universiteto Gyvybės mokslų centro) mokslininkai, kurie *Myoviridae* šeimai priklausančio fago vB_EcoM_VpaE1 genome (Šimoliūnas *et al.*, 2015) aptiko tariamai naują gp87 baltymą. Šiuo metu, šio baltymo tyrimai vyksta, tačiau duomenys apie jį dar nėra publikuoti.

2. METODIKA

2.1. Darbe naudotų kompiuterinių programų ir įrankių komandinės eilutės

Šio darbo metu, atliekant bioinformatinę duomenų analizę, naudotos įvairios kompiuterinės programos bei įrankiai, kurių komandinės eilutės pavaizduotos 5 paveiksle. Komandinė eilutė – tai tekstu paremta sąsaja tarp vartotojo ir kompiuterio.

```
# (a) Hmmbuild ir Hmmsearch įrankių Python rašmuo (angl. script):

import os, subprocess
from optparse import OptionParser
def read_file(filename):
        fr=open(filename, 'r')
        lines=fr.readlines()
        fr.close()
        return(lines)
parser = OptionParser()
parser.add_option("-a", action="store", type="string", dest="aln")
parser.add_option("-d", action="store", type="string", dest="dbli")
(options, args) = parser.parse_args()
alnli=read_file(options.aln)
dbli=read_file(options.dbli)
for x in alnli:
        alnn=x.split('.')[0]
        os.system('hmmbuild '+alnn+'.hmm '+x[:-1])
        for y in dbli:
                if not y.startswith('#'):
                        dbn=y.split('/')[-1][:-1]
                        os.system('hmmsearch --cpu 4 -E 1e-03 -A '+alnn+dbn+'.haln --tblout '+alnn+dbn+'.tout '+x.split('.')[0]+'.hmm '+y[:-1]+' > '+alnn+dbn+'.hout')

# (b) CLANS programos komandinė eilutė:

java -Xmx14G -jar clans.jar -infile file -blastpath 'blastpgp -j 1 -h 1e-03' -formatdbpath 'formatdb ' -eval 1e-03 -pval 1e-03 -cpu 8

# (c) MAFFT programos komandinė eilutė:

mafft --auto input > output

# (d) Reformat.pl, Addss.pl ir HHmake įrankių komandinės eilutės:

reformat.pl fas a3m file.fas file.a3m
addss.pl file.a3m
hhmake -i file.a3m

# (e) HHblits įrankio komandinė eilutė:

hhblits -cpu 80 -n 2 -e 1e-03 -i file.fas -d uniclust30_2018_08 -oa3m file.a3m

# (f) HHsearch įrankio komandinė eilutė:

hhsearch -Z 5000 -B 5000 -cpu 48 -i file.hhm -d /tmp/db/pfam -d /tmp/db/pdb70 -d /tmp/db/NCBI_CD -d /tmp/db/scop70 -d /tmp/db/CRISPR-Cas2018HMM -o file.hhr
```

5 pav. Kompiuterinių programų ir įrankių komandinės eilutės. Grotelių simboliu (#) pažymėtos eilutės nurodo programos ar įrankio pavadinimą.

2.2. Pradinių duomenų parsisiuntimas

Iš NCBI FTP svetainės (ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/CRISPR2018/; Makarova *et al*., 2018) FASTA formatu atsisiųsti CRISPR-Cas sistemų baltymų daugybiniai sekų palyginiai (CRISPR-Cas2018 DB).

Iš NCBI baltymų duomenų bazės (https://www.ncbi.nlm.nih.gov/protein?term=Viruses[Organism]+AND+srcdb_refseq[PROP]+NOT+cellular+organisms[ORGN]) 2019 m. gegužės 30 d. FASTA formatu parsisiųstos virusų koduojamų baltymų sekos (Virus2019 DB).

2.3. HMMER baltymų homologų paieška

Ieškant CRISPR-Cas sistemų baltymų homologų virusuose, pirmiausia, naudotas standartinis profilio-sekos lyginimo metodas. Tam pasitelktas HMMER programinės įrangos paketas (Eddy, 2011), leidžiantis sukurti baltymų tikimybinio modelio – paslėptojo Markovo modelio (angl. hidden Markov model (HMM)) – profilį ir naudoti šį profilį sekų homologų paieškoje lokalioje duomenų bazėje. HMM profilis – tai glausta daugybinio sekų palyginio reprezentacija (Soding *et al*., 2012). Kitaip tariant, tai pozicijai specifinis vertinimo modelis, nusakantis, kurie simboliai (šiuo atveju, aminorūgštys) labiausiai tikėtini bei kokia yra insercijų/delecijų atsiradimo tikimybė kiekvienoje daugybinio sekų palyginio vietoje (stulpelyje) (Eddy *et al*., 2019).

Visų pirma, HMMER Hmmbuild įrankiu iš CRISPR-Cas2018 DB sekų palyginių sukonstruoti HMM profiliai. Tuomet, gauti profiliai, Hmmsearch įrankio pagalba, palyginti su Virus2019 DB sekomis, naudojant 0,001 įtraukimo ribą (angl. E-value). Hmmbuild bei Hmmsearch įrankių paleidimo rašmuo (angl. script), parengtas Python programavimo kalba (Rossum, 1995), pavaizduotas 5a paveiksle.

2.3.1. HMMER paieškos rezultatų eksportavimas

Tam, jog būtų galima toliau analizuoti HMMER paieškos metu gautus rezultatus, jie turi būti eksportuojami. Iš Hmmsearch įrankiu gautų failų, naudojant Python programavimo kalbą, eksportuotos tik tos virusų koduojamų baltymų sekos, kurių panašumas į CRISPR-Cas sistemų baltymų profilius buvo statistiškai reikšmingas, t.y. E-įvertis (angl. E-value) <= 0,001. Kuo žemesnė statistinio įverčio reikšmė, tuo profilio-sekos panašumas yra didesnis. Paieškos

rezultatai eksportuoti CSV formatu. Tai toks failo formatas, kuriame reikšmės yra atskirtos kableliais, todėl duomenys lengvai konvertuojami į lentelę.

2.3.2. HMMER paieškos rezultatų grupavimas

Praeito etapo metu eksportuotos virusų baltymų sekos sugrupuotos pagal jų panašumo statistinius įverčius (E-įverčius) CLANS (angl. Cluster Analysis of Sequences) (Frickey ir Lupas, 2004) kompiuterine programa. CLANS sekų palyginimui naudoja BLAST (angl. Basic Local Alignment Search Tool) paiešką (Altschul *et al.*, 1990), o sekų grupių vizualizacijai – Fruchtermano-Reingoldo grafo išdėstymo algoritmą. Pastarasis sekų E-įverčius perskaičiuoja į P-įverčius (angl. P-value), pagal kuriuos sekas išdėsto erdvėje (2D arba 3D projekcijoje). Kitaip tariant, grupės vizualizuojamos pagal jėgos nukreipimą. Fruchtermano-Reingoldo algoritmas palygina jėgas, susidariusias tarp taškų (angl. points). Tos jėgos gali būti atstumiančios arba pritraukiančios. Kuo taškai panašesni, tuo juos veiks stipresnės traukos jėgos bei tokiu būdu jie sudarys grupę, ir atvirkščiai, kuo taškai mažiau panašūs, tuo labiau juos veiks atstumiančios jėgos – taip susidarys dvi atskiros grupės. Kuo ryšys tarp taškų ilgesnis, tuo taškai mažiau panašūs (Fruchterman ir Reingold, 1991). CLANS programos komandinė eilutė pavaizduota 5b paveiksle.

2.3.2.1. CLANS grupių anotavimas

Naudojant CLANS kompiuterinės programos Convex klasterizavimo metodą, skirtingomis spalvomis išskirtos grupės, turinčios ne mažiau nei 5 taškus (sekas). Tuomet, šios grupės anotuotos pasitelkiant Batch CD-Search įrankį (Marchler-Bauer ir Bryant, 2004; Marchler-Bauer *et al.*, 2011), kuris, naudodamas BLAST paiešką, leidžia nustatyti baltymų konservatyviuosius domenus iš pateikto sekų sąrašo. Konservatyviųjų domenų paieška vykdyta per CDD (angl. Conserved Domain Database) duomenų bazę (Lu *et al.*, 2020), nustačius 0,01 įtraukimo ribą. CLANS grupės anotuotos pagal geriausią reikšmingumo įvertį (E-įvertį) turintį domeną.

Tais atvejais, kai Batch CD-Search įrankiu rezultatai nebuvo gauti, CLANS grupių anotacijai pasitelktas HHpred serveris (Soding *et al.*, 2005; Zimmermann *et al.*, 2017). Tai atvirojo kodo paieškos sistema, naudojanti HH-suite programinės įrangos paketą (detaliau apie jį – 2.4. poskyryje). Šio darbo metu, HHpred užklausai pateikta viena seka iš anotuojamos CLANS grupės. Homologų paieška vykdyta per PDB (angl. Protein Data Bank) (Berman *et al.*, 2000), Pfam (angl. Protein families database) 32.0 (El-Gebali *et al.*, 2019) ir CDD duomenų

bazes, pakeitus daugybinių sekų palyginių generavimo metodą į PSI-BLAST (angl. Position-Specific Iterated BLAST). CLANS grupės anotuotos pagal geriausią įvertį turintį sekos homologą.

2.3.2.2. CLANS grupių analizė

Anotuotos CLANS grupės detaliau analizuotos nustatant grupę reprezentuojančio baltymo funkciją virusuose. Reprezentatyvus baltymas pasirinktas sudarius grupės daugybinį sekų palyginį bei identifikavus vidutinio ilgio seką, atitinkamai MAFFT (Katoh *et al.*, 2002) (komandinė eilutė pavaizduota 5c paveiksle) ir Jalview (Waterhouse *et al.*, 2009) kompiuterinėmis programomis. Funkcijos nustatymui atlikta baltymo sekos HHpred paieška (tuo pačiu principu, kaip ir 2.3.2.1. poskyryje), tarp kurios rezultatų pasirinktas geriausią panašumo tikimybę (angl. probability) turintis viruso baltymas.

Taip pat, buvo nuspręsta CLANS grupes palyginti su G. Faure bei kolegų (Faure *et al.*, 2019b) virusuose rastais Cas baltymų homologais. Šiam tikslui pasiekti, prie CLANS sugrupuotų sekų buvo pridėtos G. Faure ir bendraautorių (Faure *et al.*, 2019b) identifikuotos virusų baltymų sekos.

2.3.3. Sekų atrinkimas tolimesnei analizei

Tolimesnei analizei pasirinktos tos virusų baltymų sekos, kurių statistiniai įverčiai, atlikus HMMER paiešką, buvo geriausi. Pirmiausia, Python programavimo kalbos pagalba eksportuotos virusų baltymų sekų E-įverčių laipsnio reikšmės (pavyzdžiui, jei E-įvertis yra 1e-40, „ištrauktas“ tik po e- einantis skaičius, šiuo atveju, tai 40). Tuomet, iš šių įverčių laipsnio reikšmių sukurta histograma bei nustatyta intervalų riba toje vietoje, kurioje pasireiškia staigus sekų, su mažesne E-įverčių reikšme, skaičiaus didėjimas.

Kiekvienai atrinktai virusų baltymų sekai, naudojant NCBI duomenų bazes (NCBI, 2016), nustatytas atitinkamo viruso genomo dydis, baltymą koduojančio geno lokuso žymuo (angl. locus tag) bei atliktos HHpred ir BLASTp (angl. protein-protein BLAST) paieškos. Taip pat, patikrinta, kuriai CLANS grupei bei kuriam CRISPR-Cas2018 DB profiliui priklauso kiekviena atrinkta seka. Be to, atliktas atrinktus baltymus koduojančių virusų palyginimas su G. Faure ir kolegų (Faure *et al.*, 2019b) nustatytais Cas baltymų homologus koduojančiais virusais.

2.3.4. Cas12 baltymų homologų analizė

Nustačius, jog tarp praeitame etape atrinktų virusų baltymų sekų, didžiausią dalį sudaro Cas12 baltymų homologai, nuspręsta šiuos homologus palyginti su jau charakterizuotais panašiais virusų baltymais. Tam, naudojant MAFFT kompiuterinę programą, sudarytas daugybinis Cas12 baltymų homologų sekų palyginys, prie kurio pridėtos B. Al-Shayeb bei kolegų (Al-Shayeb *et al.*, 2019) virusuose rastos Cas12j ir Cas14 baltymų sekos.

Taip pat, buvo patikrinta šio darbo metu rastų Cas12 baltymų homologų sąsaja su IS605-tipo transpozonų koduojamais TnpA bei TnpB baltymais. Naudojant NCBI nukleotidų duomenų bazę, pažiūrėta ar virusų genomuose šalia Cas12 homologų yra TnpA genas. Tuo atveju, kai 1 geno atstumu pasitaikydavo hipotetinis baltymas, papildomai atliktos nežinomo baltymo Batch CD-Search bei HHpred paieškos. Jalview kompiuterine programa prie prieš tai sukonstruoto Cas12-Cas14 baltymų palyginio pridėta TnpB baltymo seka (NP_052299.1), išskirti konservatyvūs domenai bei sudarytos konservatyvių vietų diagramos (konsensuso sekos).

2.4. HH-suite baltymų homologų paieška

Atsižvelgiant į tai, jog anksčiau šiame darbe naudotas profilio-sekos lyginimo metodas (HMMER paieška) yra vidutiniškai jautrus ir labiau tinka tik „bendram vaizdui" apie CRISPR-Cas sistemų baltymų homologus virusuose susidaryti, toliau nuspręsta panaudoti lėtesnį, tačiau jautresnį profilio-profilio lyginimo metodą. Tam pasitelktas HH-suite programinės įrangos paketas (Steinegger *et al.*, 2019), baltymų homologų nustatymui naudojantis HMM-HMM profilių palyginį bei paiešką vykdantis per daugybę duomenų bazių.

Visų pirma, iš CRISPR-Cas2018 DB daugybinių sekų palyginių sukonstruota lokali HMM profilių duomenų bazė (CRISPR-Cas2018HMM DB), naudojant HH-suite paketo Reformat.pl, Addss.pl ir HHmake įrankius, kurie atitinkamai atlieka palyginio failo performatavimą iš FASTA į A3M, antrinės struktūros pridėjimą ir palyginio pavertimą į HMM profilį. HMM profilių generavimo komandinės eilutės pavaizduotos 5d paveiksle.

Tuomet, HH-suite HHblits įrankiu, kiekvienai šio darbo pradžioje parsisiųstai Virus2019 DB sekai sukurti HMM profiliai, atliekant dviejų iteracijų paiešką Uniclust30_2018_08 duomenų bazėje (Mirdita *et al.*, 2016), nustačius 0,001 įtraukimo ribą (komandinė eilutė pavaizduota 5e paveiksle). Šie HMM profiliai konstruojami iš virusų DB sekos ir į ją panašių sekų.

Sukurti virusų sekų HMM profiliai, HH-suite HHsearch įrankio pagalba (komandinė eilutė pavaizduota 5f paveiksle), palyginti su HMM profiliais iš lokalios CRISPR-Cas2018HMM duomenų bazės bei iš HH-suite autorių nurodytos nuorodos (http://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/) parsisiųstų Pfam 32.0, PDB, SCOP (angl. Structural Classification of Proteins) (Andreeva *et al*., 2014; Andreeva *et al*., 2020) ir CDD duomenų bazių.

2.4.1. HH-suite paieškos rezultatų eksportavimas

Tolimesnei analizei iš HHsearch įrankiu gautų failų, naudojant Python programavimo kalbą, eksportuoti tik tie paieškos rezultatai, kurių panašumo į CRISPR-Cas2018HMM DB profilius tikimybė buvo didesnė nei 70 % ir profilio padengimas didesnis nei 50 %.

2.4.2. HH-suite paieškos rezultatų grupavimas

Atsižvelgiant į tai, jog atlikus HHsearch paiešką, gautas didelis kiekis rezultatų, kurių grupavimas yra sudėtingas, pasirinkta strategija pirmiausia suklasterizuoti profilius, rastus su CRISPR-Cas2018HMM DB profiliais, o po to anotuotoms profilių grupėms priskirti HHsearch paieškos rezultatus.

Visiems CRISPR-Cas2018HMM DB profiliams atliktos paieškos anksčiau minėtose (2.4. poskyryje) duomenų bazėse, parsisiųstoms iš HH-suite autorių nurodytos nuorodos. Tam, taip pat, naudotas HHsearch įrankis. Profiliai su panašumo tikimybe didesne nei 95 %, Python rašmens pagalba, buvo eksportuoti ir pridėti prie CRISPR-Cas2018HMM DB profilių.

Visi profiliai sugrupuoti pagal jų panašumo statistinius įverčius (E-įverčius) CLANS programa. Naudojant CLANS Convex klasterizavimo metodą, skirtingomis spalvomis išskirtos grupės, turinčios ne mažiau nei 20 taškų (profilių). Šios grupės anotuotos anksčiau (2.3.2.1. poskyryje) aprašytu būdu. Anotacijas palyginus su CRISPR-Cas2018 DB daugybinių sekų palyginių anotacijomis, kai kurios profilių grupės buvo pervadintos, jog geriau atspindėtų gautus rezultatus.

HH-suite HHsearch įrankiu gauti paieškos rezultatai, t.y. virusų koduojamos baltymų sekos, priskirtos CLANS profilių grupėms, naudojant Python programavimo kalbą.

2.4.3. Csm3 baltymų homologų analizė

Detaliau analizuota tik su CRISPR susijusio Csm3 baltymo grupė, į kurią patenka vB_EcoM_VpaE1 fago koduojamas gp87 baltymas. Robetta serverio (https://robetta.bakerlab.org) pagalba sudarytas gp87 baltymo 3D struktūros modelis, kuris palygintas su PDB duomenų bazėje esančiomis kitų baltymų struktūromis, naudojant Dali serverį (Holm, 2019). PyMOL molekulinės vizualizacijos kompiuterine programa (PyMOL) gp87 baltymo struktūra užklota ant kelių didžiausią panašumą turinčių baltymų struktūrų bei iškelta gp87 baltymo funkcijos hipotezė.

2.5. HMMER ir HH-suite paieškų rezultatų palyginimas

Šio darbo metu atliktų profilio-sekos lyginimo (HMMER) bei profilio-profilio lyginimo (HH-suite) metodų rezultatai tarpusavyje palyginti pagal abiejų paieškų metu rastų virusų koduojamų baltymų identifikacinį kodą (angl. accession number). Palyginimas atliktas pasitelkus Python programavimo kalbą.

3. REZULTATAI

Šiame darbe ištirtos 8 814 virusų genomuose koduojamos 362 446 baltymų sekos. Dėl didelio duomenų kiekio, pradinė sekų analizė atlikta greičiau veikiančiu, tačiau mažiau jautriu HMMER programinės įrangos paketu. Tuomet, kol buvo tiriami pastarosios analizės rezultatai, papildomai vykdyta lėtesnė, bet didesnį jautrumą turinti CRISPR-Cas sistemos baltymų homologų paieška, naudojant HH-suite programinės įrangos paketą.

Dėl didelio gautų rezultatų kiekio, detaliau analizuoti nuspręsta po vieną Cas baltymų homologų grupę iš abiejų vykdytų paieškų:

1) profilio-sekos lyginimo metodo (HMMER) metu rastą Cas12 baltymų homologų grupę, kurios nariai sudarė didžiausią dalį tarp atrinktų, aukštus statistinius įverčius turinčių virusų sekų;
2) profilio-profilio lyginimo metodo (HH-suite) pagalba pirmą kartą nustatytą Csm3 baltymų homologų grupę, į kurią patenka mūsų centro (Vilniaus universiteto Gyvybės mokslų centro) mokslininkų sekvenuoto vB_EcoM_VpaE1 fago koduojamas gp87 baltymas.

3.1. HMMER baltymų homologų paieškos rezultatai

Iš šio darbo pradžioje parsisiųstų 530 CRISPR-Cas sistemų baltymų daugybinių sekų palyginių (CRISPR-Cas2018 DB) bei 362 446 baltymų sekų (Virus2019 DB) HMMER Hmmbuild ir Hmmsearch įrankių pagalba, atrinkta 70 CRISPR-Cas baltymų profilių bei 2 117 baltymų sekų, koduojamų 1 330 virusų. Eksportuoti Hmmsearch rezultatai pateikti prieduose esančioje 1 lentelėje, kurios fragmentas pavaizduotas 6 paveiksle.

	CRISPR-Cas profilis	
Viruso pavadinimas	cd06127	cd09639
Arthrobacter phage KellEzio	YP_009301318.1~1.5e-28	
Elephantid betaherpesvirus 1		YP_007969774.2~0.0003
Gordonia phage Phinally	YP_009291409.1~1.3e-25	

6 pav. HMMER paieškos rezultatų lentelės fragmentas. Lentelėje pateikti CRISPR-Cas baltymų profiliai bei atitinkamo viruso genome koduojamo baltymo identifikacinis kodas, šalia kurio (atskirta „~“ simboliu) nurodytas E-įvertis.

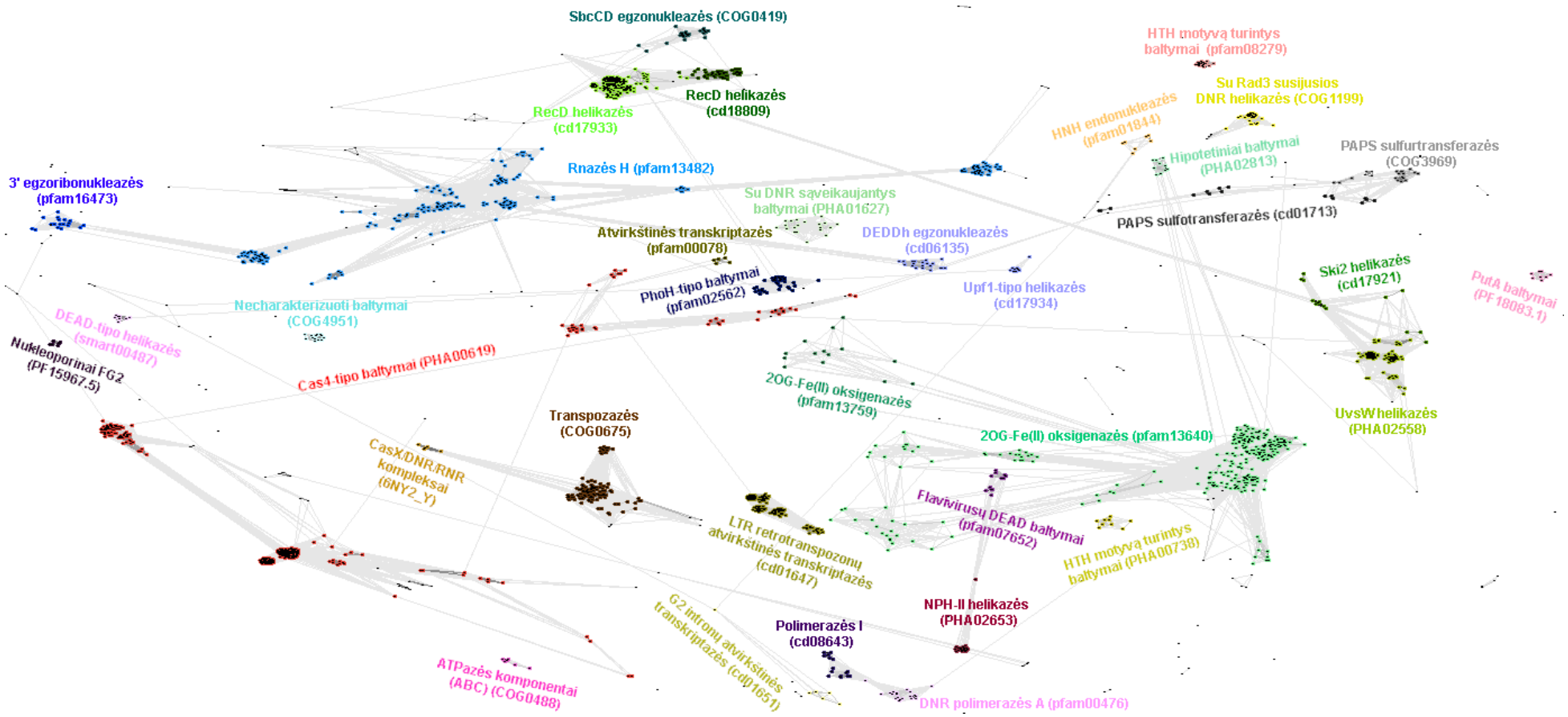
3.1.1. HMMER paieškos rezultatų grupės

Sugrupavus HMMER Hmmsearch įrankiu gautus rezultatus CLANS programa, išskirtos 35 baltymų grupės (7 pav.). Didžiausią grupę, turinčią 348 sekas, sudaro Cas4-tipo baltymai (PHA00619). Savo dydžiu, taip pat, išsiskiria 2OG-Fe(II) oksigenazių (pfam13640),

RecD helikazių (cd17933) bei ribonukleazių H (pfam13482) grupės, atitinkamai turinčios po 287, 232 ir 218 narių. Mažiausias grupes, susidedančias tik iš 5 narių, formuoja atvirkštinių transkriptazių (pfam00078) bei CasX/DNR/RNR kompleksų baltymai (6NY2_Y). Informacija apie kiekvieną CLANS sekų grupę pateikta prieduose esančioje 2 lentelėje.

Atlikus kiekvienos CLANS sekų grupės reprezentatyvaus baltymo HHpred paieškas, nustatyta, jog didžioji dalis baltymų grupių atsakingos už virusų replikaciją (priedai; 2 lentelė). Kelių grupių funkcija virusuose yra struktūrinė, gynybinė bei anti-CRISPR.

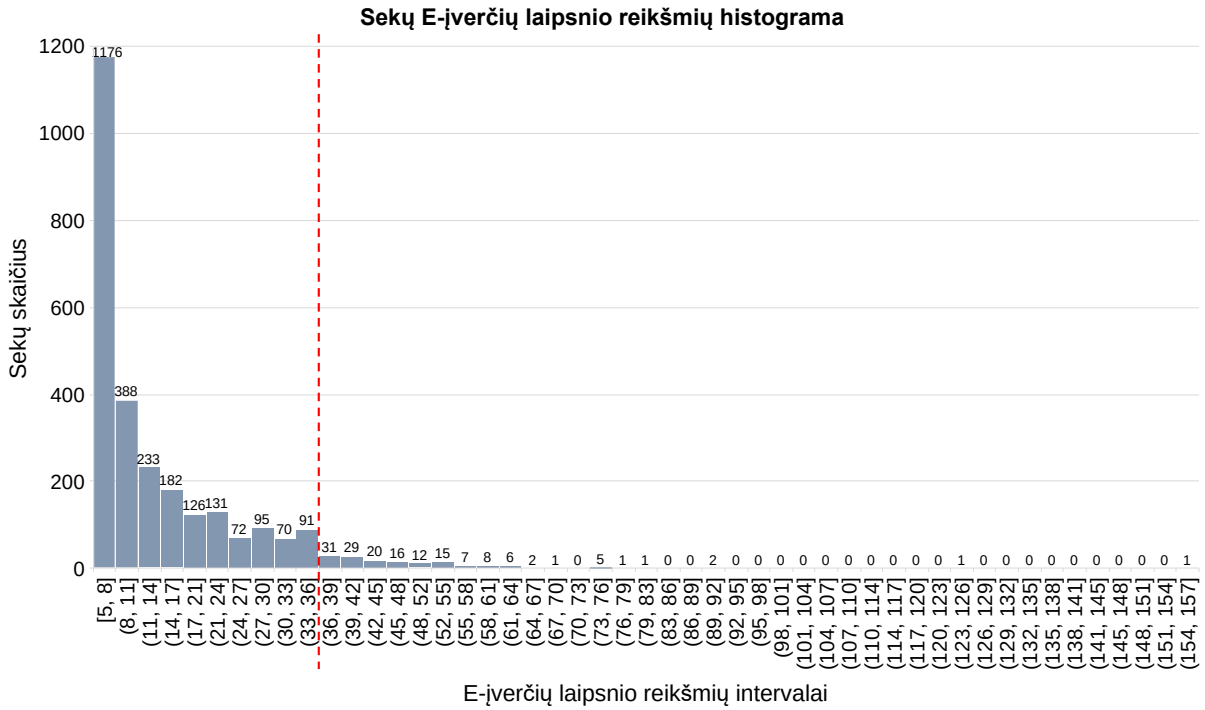
Prie CLANS sugrupuotų sekų pridėjus G. Faure ir kolegų (Faure *et al.*, 2019b) identifikuotas virusų baltymų sekas, kai kurios iš jų pateko į Cas4-tipo baltymų (PHA00619) ir Ski2 helikazių (cd17921) grupes. Šias sekas G. Faure ir bendraautoriai (Faure *et al.*, 2019b) atitinkamai charakterizavo kaip Cas4 ir Cas3 baltymų homologus. Kitos sekos suformavo 3 atskiras naujas grupes arba išsidėstė kaip pavieniai taškai. Šį nesutapimą galima būtų paaiškinti skirtingais G. Faure ir kolegų (Faure *et al.*, 2019b) bei šio darbo metu tirtais sekų rinkiniais.



7 pav. HMMER paieškos rezultatai, sugrupuoti pagal sekų tarpusavio panašumą. Taškai žymi sekas. Linijomis sujungtos sekos, kurių CLANS panašumo įvertis (P-įvertis) ≤ 0,001. Kuo sekos panašesnės, tuo linijos trumpesnės bei ryškesnės. Grupių anotacijų spalvos sutampa su grupės narių spalvomis. Šalia grupių anotacijų skliausteliuose nurodyti charakterizuotų baltymų profilių identifikaciniai kodai.

3.1.2. Tolimesnė HMMER rezultatų analizė

Kaip jau buvo minėta 2.3.3. poskyryje, toliau analizuoti nuspręsta tik tas HMMER paieškos metu atrinktas virusų baltymų sekas, kurių statistiniai įverčiai buvo geriausi. Tam iš sekų E-įverčių laipsnio reikšmių sukurta histograma (8 pav.). Toliau darbe naudotos tos sekos, kurios patenka į [1e-36, 1e-157] E-įverčių laipsnio reikšmių intervalą.



8 pav. HMMER paieškos metu atrinktų virusų baltymų sekų E-įverčių laipsnio reikšmių histograma. Horizontalioje (X) ašyje pažymėti E-įverčių laipsnio reikšmių intervalai, o vertikalioje (Y) ašyje – sekų skaičius. Virš kiekvieno histogramos stulpelio nurodytas tikslus virusų baltymų sekų skaičius, patenkantis į tam tikrą E-įverčių laipsnio reikšmių intervalą. Raudona punktyrinė linija nurodo toliau darbe analizuotų sekų E-įverčių laipsnio reikšmių intervalų ribą.

Atlikus detalesnę atrinktų, aukštus statistinius įverčius turinčių, virusų koduojamų skirtingų baltymų sekų analizę, nustatyta, jog šiek tiek daugiau nei puse (49 sekos iš 95) yra Cas12 baltymų homologai (priedai; 3 lentelė). Likusios sekos pateko į 2-oksoglutarato (2OG) oksigenazių, ATPazių, DEDDh egzonukleazių, DinG helikazių bei atvirkštinių transkriptazių CRISPR-Cas baltymų profilių šeimas. Be to, palyginus atrinktų baltymų virusų pavadinimus su G. Faure ir kolegų (Faure *et al.*, 2019b) charakterizuotų baltymų virusais, nustatyti tik 2 bendri virusai, patenkantys į DinG helikazių šeimą (priedai; 3 lentelė).

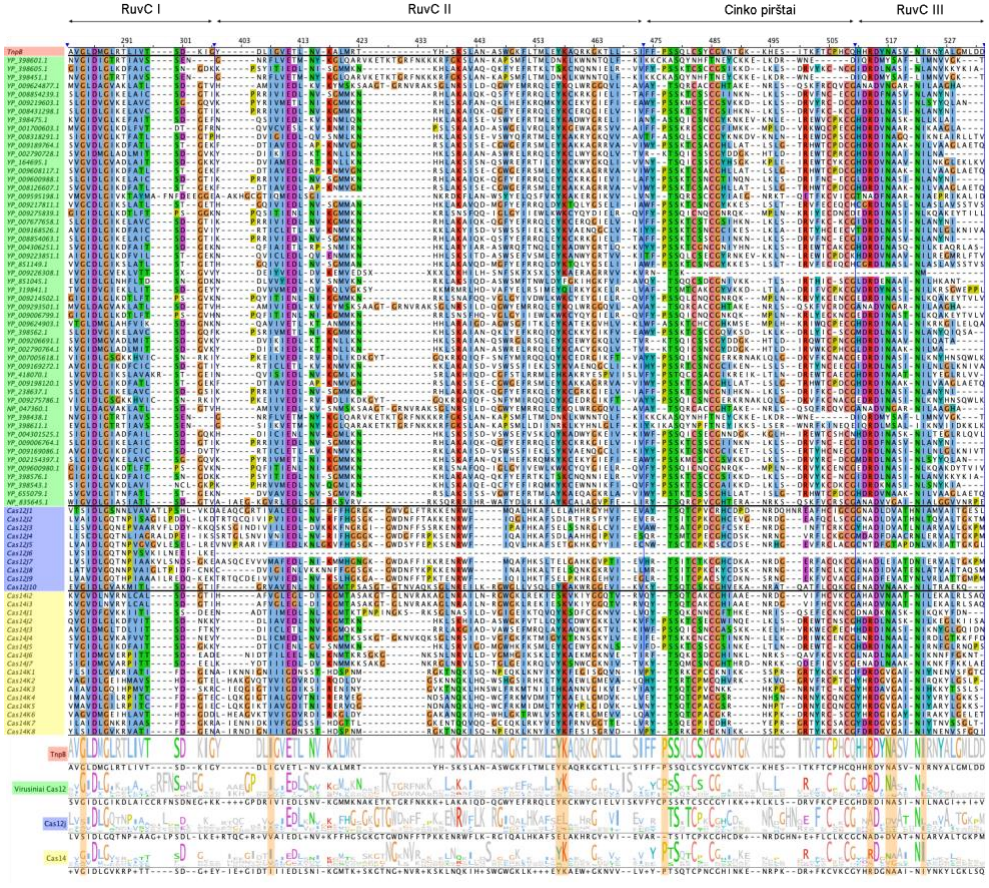
3.1.3. Cas12 baltymų homologai

HHpred paieškos rezultatai patvirtino faktą, jog šio darbo metu nustatyti Cas12 baltymų homologai (toliau – virusiniai Cas12 baltymai) turi panašumą su CRISPR-Cas baltymais, o

BLAST paieška bei priskyrimas CLANS grupėms parodė sekų giminingumą su transpozazėmis (9 pav.).

Patikrinus virusinių Cas12 baltymų sąsajas su IS605-tipo transpozonų transpozicijos baltymu TnpA, HHpred paieškos pagalba nustatyta, jog tik 3 virusų genomuose (Sulfolobus monocaudaviruso SMV3, Lactobacillus fago CL1 bei Microcystis viruso Ma-LMM01) šalia Cas12 homologo (1 geno atstumu) galimai išsidėstęs TnpA homologo TnpAREP genas (Messing *et al.*, 2012) (priedai; 3 lentelė). Taip pat, pastebėta, jog rastus virusinius Cas12 baltymus koduoja vidutinio dydžio virusai (genomo ilgio vidurkis ~120 kb).

Pagal IS605-tipo transpozonų koduojamo TnpB baltymo konservatyvius domenus palyginus virusinių Cas12 baltymų sekas su B. Al-Shayeb ir kolegų (Al-Shayeb *et al.*, 2019) charakterizuotomis didelių virusų Cas12j ir Cas14 baltymų sekomis, nustatyta, jog šio darbo metu rasti Cas12 baltymų homologai panašesni į Cas14 baltymus. Tą galima pamatyti iš 9 paveiksle pateikto palyginio konservatyvių vietų diagramų (konsensuso sekų) 284 (RuvC I domenas), 408, 457-458 (RuvC II domenas), 477 (cinko pirštų domenas), 514, 517-518 ir 523 (RuvC III domenas) pozicijų, kuriose aminorūgščių liekanos sutampa tarp TnpB, virusinių Cas12 bei Cas14 baltymų, tačiau skiriasi Cas12j sekose. Minėtose pozicijose konservatyvios yra, atitinkamai glicino (G), izoleucino (I), tirozino (Y) bei lizino (K), prolino (P), arginino (R), asparagino (N) bei alanino (A) ir izoleucino (I) liekanos.



9 pav. TnpB, virusinių Cas12, Cas12j bei Cas14 baltymų palyginys. Paveikslo viršuje nurodyti TnpB baltymo konservatyvūs domenai. TnpB, virusinių Cas12 bei B. Al-Shayeb ir kolegų (Al-Shayeb *et al*., 2019) charakterizuotų Cas12j ir Cas14 baltymų sekos paryškintos, atitinkamai raudona, žalia, mėlyna ir geltona spalvomis. Paveikslo apačioje pavaizduotos kiekvienos baltymų grupės konservatyvių vietų diagramos (konsensuso sekos). Diagramose oranžine spalva paryškintos pozicijos, kuriose aminorūgščių liekanos sutampa tarp TnpB, virusinių Cas12 ir Cas14, tačiau skiriasi nuo Cas12j baltymų sekų.

3.2. HH-suite baltymų homologų paieškos rezultatai

Atlikus jautresnę CRISPR-Cas sistemų baltymų homologų paiešką HH-suite HHsearch įrankiu, atrinktos 8 288 virusų koduojamų baltymų sekos. 10 paveiksle pateiktas eksportuotų HH-suite rezultatų lentelės fragmentas. Prieduose esančios 4 lentelės duomenys nurodo sukurtų virusų baltymų HMM profilių reprezentatyvių sekų identifikacinius kodus, šalia kurių pateikta informacija apie lyginimo metu rastą geriausią E-įvertį turintį baltymų profilį bei panašiausią CRISPR-Cas baltymų profilį.

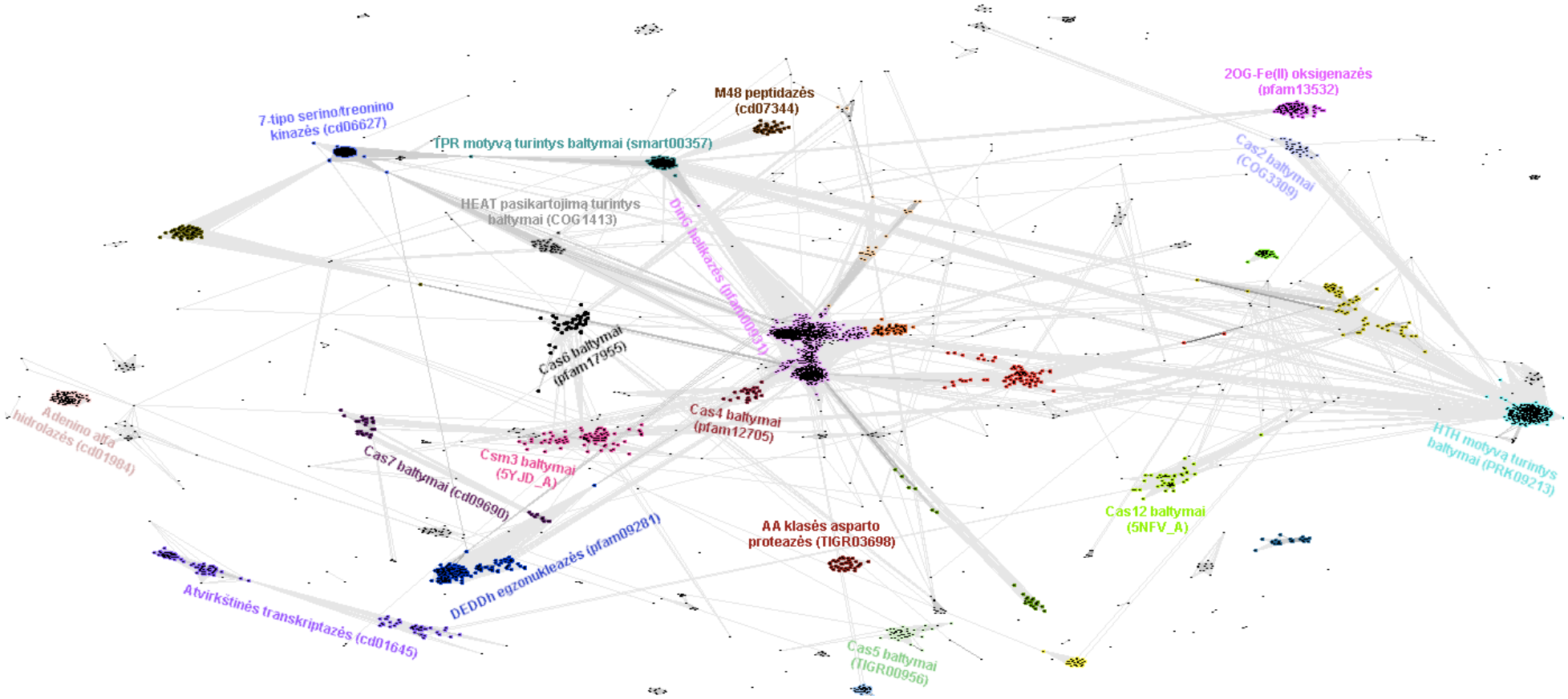
Viruso profilis	Panašiausias profilis					CRISPR-Cas profilis				
Reprezentatyvios sekos identifikacinis kodas	Pavadinimas	Regionas	Panašumo tikimybė (%)	Ilgis (stulpeliai)	Padengimas (%)	Pavadinimas	Regionas	Panašumo tikimybė (%)	Ilgis (stulpeliai)	Padengimas (%)
YP_009213801	6M9K_C	4-205	100,0	200	88,50	Kooncd09637	63-203	96,2	125	70,22
YP_009509001	1CU1_A	6-625	100,0	608	94,26	Kooncd09639	200-474	99,8	251	71,10
YP_009509004	3I5K_B	1-526	100,0	519	91,70	Koonicity0040	104-356	95,8	232	78,91

10 pav. HH-suite paieškos rezultatų lentelės fragmentas.

3.2.1. HH-suite paieškos rezultatų grupės

Kaip jau buvo minėta 2.4.2. poskyryje, grupuoti nuspręsta ne HH-suite paieškos metu rastų virusų profilių reprezentatyvias sekas, bet su CRISPR-Cas profiliais rastus profilius ir tuomet virusų sekas priskirti anotuotoms profilių grupėms.

HHsearch įrankiu atlikus CRISPR-Cas profilių homologų paiešką, rasti 4 049 skirtingi, aukšta panašumo tikimybe (> 95 %) pasižymintys profiliai (priedai; 5 lentelė), kurie CLANS programa suskirstyti į 28 grupes. HH-suite paieškos metu rastos virusų sekos (6 688) pateko ne į visas, o į 18 CLANS profilių grupių (11 pav.). Didžiausias skaičius sekų (2165) priskirta atvirkštinių transkriptazių (cd01645) grupei. Mažiausiai sekų (po 1) pateko į Cas7 (cd09690), Cas6 (pfam17955), Cas5 (TIGR00956), M48 peptidazių (cd07344) bei HEAT pasikartojimą turinčių (COG1413) baltymų grupes.



11 pav. HHsearch profilių paieškos rezultatai, sugrupuoti pagal profilių reprezentatyvių sekų tarpusavio panašumą. Taškai žymi profilius. Linijomis sujungti profiliai, kurių CLANS panašumo įvertis (P-įvertis) ≤ 1e-6. Paveiksle anotuotos tik tos grupės, į kurias patenka HH-suite paieškos metu rastos virusų sekos. Grupių anotacijų spalvos sutampa su grupės narių spalvomis. Šalia grupių anotacijų skliausteliuose nurodyti charakterizuotų baltymų profilių identifikaciniai kodai.

Palyginus abiejų paieškų (HMMER bei HH-suite) metu rastas virusų koduojamų baltymų sekas, nustatyta, jog iš 6 688 CLANS profilių grupėms priskirtų sekų, profilio-sekos lyginimo (HMMER) metodo pagalba rastos 1 549 sekos.

12 paveiksle pateiktas prieduose esančios 6 lentelės fragmentas, kuriame atsispindi HMMER ir HH-suite paieškų rezultatų tarpusavio palyginimas. 5 CLANS profilių grupėms (Csm3 baltymų (5YJD_A), AA klasės asparto proteazių (TIGR03698), 7-tipo serino/treonino kinazių (cd06627), M48 peptidazių (cd07344) ir HEAT pasikartojimą turinčių baltymų (COG1413)) priklausančios sekos rastos tik HH-suite paieška.

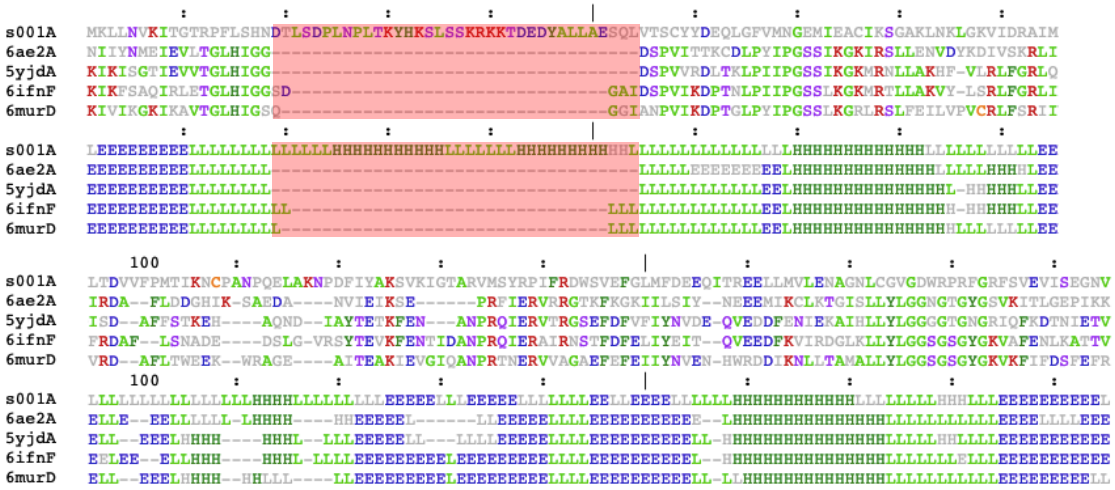
Profilių CLANS grupė	Sekų kiekis (kiek iš jų rasta su HMMER paieška)
Atvirkštinės transkriptazės (cd01645)	2165 (180)
DEDDh egzonukleazės (pfam09281)	1418 (287)
DinG helikazės (pfam00931)	1103 (214)
Cas4 baltymai (pfam12705)	846 (375)
2OG-Fe(II) oksigenazės (pfam13532)	722 (326)
Cas12 baltymai (5NFV_A)	112 (111)
HTH motyvą turintys baltymai (PRK09213)	104 (16)
Adenino alfa hidrolazės (cd01984)	90 (35)
TPR motyvą turintys baltymai (smart00357)	65 (1)
Cas2 baltymai (COG3309)	20 (1)
Csm3 baltymai (5YJD_A)	15 (0)
AA klasės asparto proteazės (TIGR03698)	14 (0)
7-tipo serino/treonino kinazės (cd06627)	9 (0)
Cas7 baltymai (cd09690)	1 (1)
Cas6 baltymai (pfam17955)	1 (1)
Cas5 baltymai (TIGR00956)	1 (1)
M48 peptidazės (cd07344)	1 (0)
HEAT pasikartojimą turintys baltymai (COG1413)	1 (0)

12 pav. HMMER ir HH-suite paieškų metu rastų sekų palyginimas.

Prieduose esančioje 6 lentelėje, taip pat, pateikta detalesnė informacija apie kiekvieną CLANS profilių grupę, t.y. apie kiekvieną grupę sudarančių baltymų profilių vidutinę panašumo tikimybę į virusų baltymų profilius, vidutinį CRISPR-Cas profilių padengimą bei profilių priklausymą tam tikrai CRISPR-Cas profilių kategorijai ir šeimai.

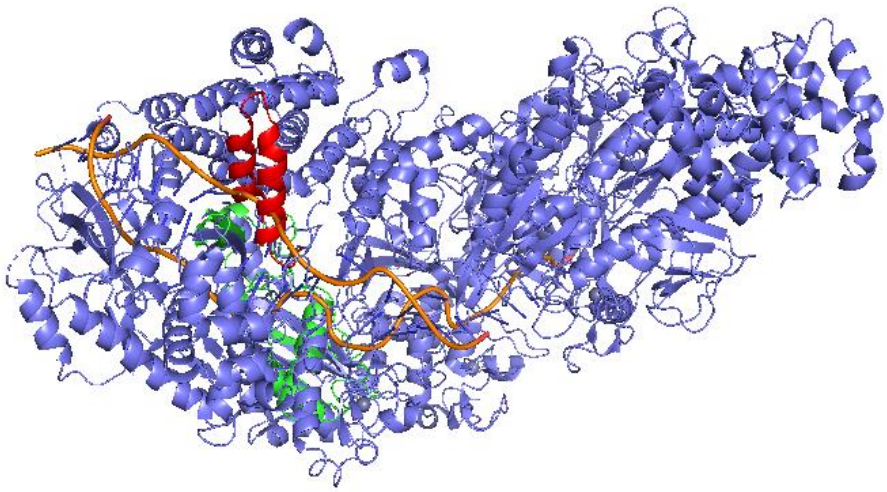
3.2.2. Csm3 baltymų homologai

Dali serverio pagalba palyginus į Csm3 baltymų (5YJD_A) profilių grupę patenkančio vB_EcoM_VpaE1 fago koduojamo gp87 baltymo seką su kitomis jau prokariotuose charakterizuotomis Csm komplekso sekomis, nustatyta, jog gp87 baltymas savo struktūroje turi intarpą, kurio neturi kiti į jį panašūs baltymai (13 pav.).



13 pav. Gp87 ir Csm komplekso baltymų palyginiai. Viršuje esanti dalis nurodo aminorūgščių palyginį, apačioje – antrinės struktūros palyginį (H – spiralė, E – klostė, L – ritė). Raudona spalva pažymėtas gp87 baltymo intarpas, kurio nėra kituose baltymuose. S001A, vB_EcoM_VpaE1 fago gp87 baltymas; 6ae2A, *Thermoplasma volcanium* Csm3 baltymas; 5yjdA, *Staphylococcus epidermidis* Csm3 baltymas; 6ifnF, *Streptococcus thermophilus* Csm kompleksas; 6murD, *Thermococcus onnurineus* Csm kompleksas.

Užklojus gp87 baltymo 3D struktūros modelį ant prieš tai minėtų Csm3 baltymų bei Csm kompleksų struktūrų modelių, pastebėta, kad nustatytas gp87 intarpas atsiduria sąveikos su crRNR pusėje (14 pav.).

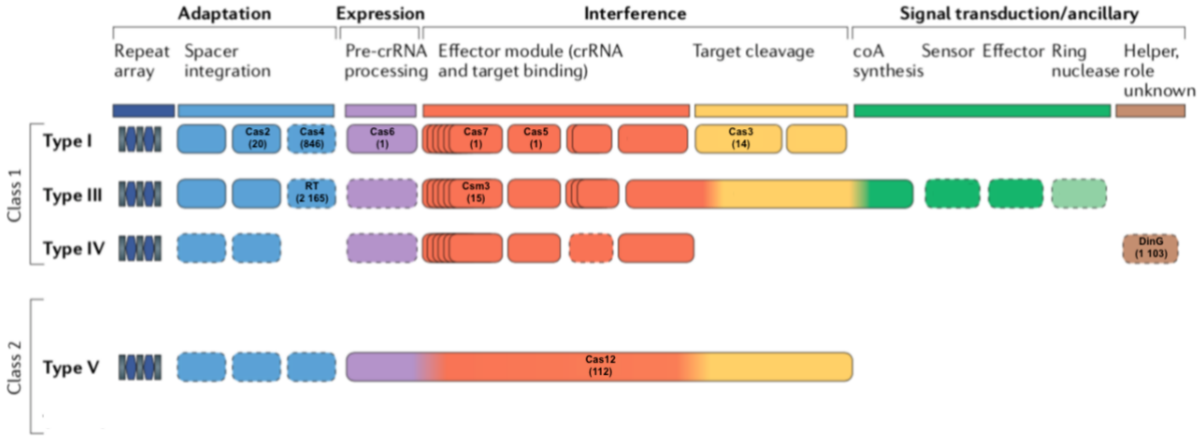


14 pav. Gp87 baltymo modelis palygintas su Csm komplekso modeliu (6murD). Mėlyna spalva žymi Csm komplekso struktūrą, oranžinė – crRNR, žalia – gp87 baltymą, raudona – gp87 intarpą.

4. REZULTATŲ APTARIMAS

Šio darbo metu atlikta paieška – tai pirma tokios plačios apimties bei tokio aukšto jautrumo CRISPR-Cas sistemų baltymų homologų paieška virusuose. Tyrime vykdyta analizė ne tik patvirtino anksčiau virusuose identifikuotus solo Cas baltymų homologus, bet ir padėjo rasti naujų Cas baltymų homologų (pvz., Csm3).

Iš 15 paveiksle pateiktos, darbo rezultatus apibendrinančios iliustracijos, galima matyti, jog virusai į savo genomus, pagrinde, inkorporuoja Cas baltymus, susijusius su adaptacijos bei interferencijos CRISPR-Cas sistemų funkciniais moduliais. Be to, kaip ir buvo tikėtasi, didžioji dalis kitų rastų baltymų priklauso skirtingoms, plačiai paplitusioms baltymų grupėms, susijusioms su virusų replikacija ar gynyba. Tai įvairios helikazės, nukleazės, polimerazės, transpozazės, hidrolazės, proteazės, kinazės ir pan.



15 pav. Tyrimo metu virusuose identifikuoti Cas baltymų homologai. Skliausteliuose nurodyti rastų baltymų kiekiai, remiantis jautresnės paieškos (HH-suite) rezultatais. Tuo atveju, kai baltymai nebuvo aptikti HH-suite paieška (Cas3), jų kiekis nurodytas pagal HMMER paieškos rezultatus. Skirtinguose tipuose pasikartojantys sistemų komponentai, pažymėti tik I tipe. RT, atvirkštinės transkriptazės. Iliustracija šiam darbui pritaikyta pagal K. S. Makarova ir bendraautorių (Makarova *et al.*, 2020) sukurtą paveikslą.

Profilio-sekos lyginimo metodo (HMMER paieškos) pagalba rasti jau prieš tai G. Faure ir kolegų (Faure *et al.*, 2019b) virusuose identifikuoti Cas4 bei Cas3 baltymų homologai, tikėtina, jog buvo pritaikyti skirtingoms virusų funkcijoms atlikti, pavyzdžiui, kovoti su šeimininko CRISPR-Cas sistema ar spręsti tarp-virusinius konfliktus. Greičiausiai, panašiai veikia ir šiame tyrime identifikuoti aukštu panašumu į CRISPR-Cas baltymų profilius pasižymintys Cas12 baltymų homologai. Kaip parodė šių baltymų detali analizė, tik 3 nustatyti Cas12 homologai turi artimą ryšį su TnpA transpozazėmis, kas leidžia teigti, jog visi kiti rasti Cas12 homologai yra ne inserciniai elementai, o į virusų genomus įsiterpę prokariotų Cas12 baltymai. Be to, virusuose rastus Cas12 baltymų homologus teisingiau būtų vadinti virusiniais

Cas14 baltymais, kaip tą atliktos didelių fagų genomų analizės rezultatuose nurodo B. Al-Shayeb ir bendraautoriai (Al-Shayeb *et al*., 2019).

Pasitelkus profilio-profilio lyginimo metodą (HH-suite paiešką), virusų genomuose aptikti jau prieš tai šiame darbe identifikuoti Cas4 bei Cas12 baltymų homologai. Papildomai, jautresnės paieškos metu nustatyti ir Cas2 baltymų homologai, kurių egzistavimą jau anksčiau nurodė G. Faure ir kolegos (Faure *et al*., 2019b), tačiau jų funkcija virusuose ir toliau lieka neaiški. Tuo tarpu, Cas5, Cas6 bei Cas7 baltymų grupėms priskirtas sekas reiktų tirti detaliau, kadangi į šias grupes pateko tik po 1 virusų koduojamų baltymų seką.

Šio tyrimo svarbiausiu rezultatu galima įvardinti HH-suite paieškos metu virusų genomuose identifikuotus Csm3 baltymų homologus, kurie iki šiol dar niekur nebuvo aprašyti. Manome, jog pastarieji baltymai sąveikauja su šeimininko CRISPR-Cas sistemų baltymais bei galimai blokuoja jų susirišimą su crRNR, taip apsaugant virusą nuo prokariotų įgyto imuniteto mechanizmo.

IŠVADOS

1. Atlikus profilio-sekos bei profilio-profilio lyginimo CRISPR-Cas baltymų homologų paieškas, tarp 362 446 ištirtų virusų koduojamų baltymų sekų, rasta 1 010 CRISPR-Cas sistemų baltymų homologų. Virusai į savo genomus, pagrinde, inkorporuoja Cas baltymus, susijusius su adaptacijos bei interferencijos CRISPR-Cas sistemų funkciniais moduliais.
2. Sugrupavus bei išanalizavus rastas CRISPR-Cas baltymų homologų sekas, nustatyta:
    a) pirmą kartą virusuose charakterizuota Csm3 baltymų grupė, galimai apsauganti virusą nuo šeimininko CRISPR-Cas sistemų mechanizmo;
    b) naujai virusuose rastos Cas5, Cas6 ir Cas7 baltymų grupės, kurių apibūdinimui reikalinga detalesnė analizė;
    c) jau anksčiau virusuose identifikuotos Cas2, Cas3 bei Cas4 baltymų grupės, kurios, tikriausiai, pritaikytos skirtingoms virusų funkcijoms atlikti;
    d) prieš tai tik dideliuose faguose aptikta Cas12/14 baltymų grupė, kuri, tikėtina, jog veikia arba kaip transpozazė, arba įeina į virusų CRISPR-Cas komplekso sudėtį (tik tuo atveju, kai šios grupės baltymų homologai artimai giminingi šeimininko Cas12/14 baltymams).

PADĖKA

Dėkoju savo darbo vadovui dr. Dariui Kazlauskui už kantrybę, perduotas žinias bei suteiktą galimybę rengti magistro baigiamąjį darbą Vilniaus universiteto Gyvybės mokslų centro Biotechnologijos instituto Bioinformatikos skyriuje. Taip pat, noriu padėkoti dr. Albertui Timinskui už pagalbą atliekant HH-suite paiešką.

LITERATŪROS SĄRAŠAS

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic Local Alignment Search Tool. J. Mol. Biol. 215, 403-410. https://doi.org/10.1016/S0022-2836(05)80360-2
2. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG, 2014. SCOP2 prototype: a new approach to protein structure mining. Nucl Acids Res. 42(D1), D310-D314. https://doi.org/10.1093/nar/gkt1242
3. Andreeva A, Kulesha E, Gough J, Murzin AG, 2020. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Research. 48(D1), D376-D382. https://doi.org/10.1093/nar/gkz1064
4. Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, et al. Clades of Huge Phage from across Earth's Ecosystems. Priimta spaudai, 2019. DOI:10.1101/572362
5. Barrangou R, Marraffini LA, 2014. CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity. Molecular Cell. 54(2), 234-244. https://doi.org/10.1016/j.molcel.2014.03.011
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al., 2000. The Protein Data Bank. Nucleic Acids Research. 28 (1), 235-42. https://doi.org/10.1093/nar/28.1.235
7. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD, 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology. 151(8), 2551-2561. https://doi.org/10.1099/mic.0.28048-0
8. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016;44(D1):D7-D19.
9. Deveau H, Garneau JE, Moineau S., 2010. CRISPR/Cas System and Its Role in Phage-Bacteria Interactions. Annu Rev Microbiol. 64(1), 475-493. https://doi.org/10.1146/annurev.micro.112408.134123
10. Dou C, Yu M, Gu Y, Wang J, Yin K, et al., 2018. Structural and Mechanistic Analyses Reveal a Unique Cas4-like Protein in the Mimivirus Virophage Resistance Element System. iScience. 3, 1-10. https://doi.org/:10.1016/j.isci.2018.04.001
11. Eddy SR, 2011. Accelerated Profile HMM Searches. Pearson WR, ed. PLoS Comput Biol. 7(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195
12. Eddy SR, HMMER development team, 2019. HMMER User's Guide. Howard Hughes Medical Institute :229.

13. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, et al.,2019. The Pfam protein families database in 2019. Nucleic Acids Research. 47(D1), D427-D432. https://doi.org/10.1093/nar/gky995
14. Faure G, Makarova KS, Koonin EV, 2019a. CRISPR–Cas: Complex Functional Networks and Multiple Roles beyond Adaptive Immunity. Journal of Molecular Biology. 431(1), 3-20. doi:10.1016/j.jmb.2018.08.030
15. Faure G, Shmakov SA, Yan WX, Cheng DR, Scott DA, et al., 2019b. CRISPR–Cas in mobile genetic elements: counter-defence and beyond. Nat Rev Microbiol. 17(8), 513-525. https://doi.org/10.1038/s41579-019-0204-7
16. Frickey T, Lupas A, 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics. 20(18), 3702-3704. https://doi.org/10.1093/bioinformatics/bth444
17. Fruchterman TMJ, Reingold EM, 1991. Graph drawing by force-directed placement. Softw: Pract Exper. 21(11), 1129-1164. https://doi.org/10.1002/spe.4380211102
18. Glemžaitė M. Thermo Fisher Scientific Baltics. RNA-directed DNA cleavage by the Cas9-crRNA complex. WO2013/142578. 2014.
19. He S, Corneloup A, Guynet C, Lavatine L, Caumont-Sarcos A, et al., 2015. The IS200/IS605 Family and "Peel and Paste" Single-strand Transposition Mechanism. Microbiol Spectrum. 3(4), MDNA3-0039-2014. https://doi.org/10.1128/microbiolspec.MDNA3-0039-2014
20. Holm L, 2019. Benchmarking fold detection by DaliLite v.5. Bioinformatics. 35(24), 5326-5327. https://doi.org/10.1093/bioinformatics/btz536
21. Hooton SPT, Brathwaite KJ, Connerton IF, 2016. The Bacteriophage Carrier State of Campylobacter jejuni Features Changes in Host Non-coding RNAs and the Acquisition of New Host-derived CRISPR Spacer Sequences. Front Microbiol. 7. https://doi.org/10.3389/fmicb.2016.00355
22. Horvath P, Barrangou R, 2010. CRISPR/Cas, the Immune System of Bacteria and Archaea. Science. 327(5962), 167-170. https://doi.org/10.1126/science.1179555
23. Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A, 1987. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. Journal of Bacteriology. 169(12), 5429-5433. https://doi.org/10.1128/JB.169.12.5429-5433.1987
24. Jackson SA, McKenzie RE, Fagerlund RD, Kieper SN, Fineran PC, et al., 2017. CRISPR-Cas: Adapting to change. Science. 356(6333), eaal5056. https://doi.org/10.1126/science.aal5056

25. Jansen R, Embden JDA van, Gaastra W, Schouls LM, 2002. Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol. 43(6), 1565-1575. https://doi.org/10.1046/j.1365-2958.2002.02839.x
26. Katoh K, 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research. 30(14), 3059-3066. https://doi.org/10.1093/nar/gkf436
27. Kazlauskiene M, Kostiuk G, Venclovas Č, Tamulaitis G, Siksnys V, 2017. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. Science. 357(6351), 605-609. https://doi.org/10.1126/science.aao0100
28. Koonin EV, Makarova KS, 2018. Anti-CRISPRs on the march. Science. 362(6411), 156-157. https://doi.org/10.1126/science.aav2440
29. Koonin EV, Makarova KS, 2019. Origins and evolution of CRISPR-Cas systems. Phil Trans R Soc B. 374(1772), 20180087. https://doi.org/10.1098/rstb.2018.0087
30. Krupovic M, Cvirkaite-Krupovic V, Prangishvili D, Koonin EV, 2015. Evolution of an archaeal virus nucleocapsid protein from the CRISPR-associated Cas4 nuclease. Biol Direct. 10(1), 65. https://doi.org/10.1186/s13062-015-0093-2
31. Levasseur A, Bekliz M, Chabrière E, Pontarotti P, La Scola B, et al., 2016. MIMIVIRE is a defence system in mimivirus that confers resistance to virophage. Nature. 531(7593):249-252. https://doi.org/10.1038/nature17146
32. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, et al., 2020. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Research. 48(D1), D265-D268. https://doi.org/10.1093/nar/gkz991
33. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, et al., 2015 An updated evolutionary classification of CRISPR–Cas systems. Nat Rev Microbiol. 13(11), 722-736. https://doi.org/10.1038/nrmicro3569
34. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, et al., 2020. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol. 18(2), 67-83. https://doi.org/10.1038/s41579-019-0299-x
35. Makarova KS, Wolf YI, Koonin EV, 2018. Classification and Nomenclature of CRISPR-Cas Systems: Where from Here? The CRISPR Journal. 1(5), 325-336. https://doi.org/10.1089/crispr.2018.0033
36. Marchler-Bauer A, Bryant SH, 2004. CD-Search: protein domain annotations on the fly. Nucleic Acids Research. 32(Web Server):W327-W331. https://doi.org/10.1093/nar/gkh454

37. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, et al., 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research. 39(Database), D225-D229. https://doi.org/10.1093/nar/gkq1189
38. Marraffini LA, 2015. CRISPR-Cas immunity in prokaryotes. Nature. 526(7571), 55-61. https://doi.org/10.1038/nature15386
39. McGinn J, Marraffini LA, 2019. Molecular mechanisms of CRISPR–Cas spacer acquisition. Nat Rev Microbiol. 17(1), 7-12. https://doi.org/10.1038/s41579-018-0071-7
40. Messing SAJ, Ton-Hoang B, Hickman AB, McCubbin AJ, Peaslee GF, et al., 2012. The processing of repetitive extragenic palindromes: the structure of a repetitive extragenic palindrome bound to its associated nuclease. Nucleic Acids Research. 40(19), 9964-9979. https://doi.org/10.1093/nar/gks741
41. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, et al., 2017. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 45(D1), D170-D176. https://doi.org/10.1093/nar/gkw1081
42. Mohanraju P, Makarova KS, Zetsche B, Zhang F, Koonin EV, et al., 2016. Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science. 353(6299), aad5147. https://doi.org/10.1126/science.aad5147
43. Mojica FJM, Diez-Villasenor C, Garcia-Martinez J, Soria E, 2005. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. J Mol Evol. 60(2), 174-182. https://doi.org/10.1007/s00239-004-0046-3
44. Naser IB, Hoque MM, Nahid MA, Tareq TM, Rocky MK, et al., 2017. Analysis of the CRISPR-Cas system in bacteriophages active on epidemic strains of Vibrio cholerae in Bangladesh. Sci Rep. 7(1), 14880. https://doi.org/10.1038/s41598-017-14839-2
45. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, et al., 2016. Uncovering Earth's virome. Nature. 536(7617), 425-430. https://doi.org/10.1038/nature19094
46. Pawluk A, Davidson AR, Maxwell KL, 2018. Anti-CRISPR: discovery, mechanism and function. Nat Rev Microbiol. 16(1), 12-17. https://doi.org/10.1038/nrmicro.2017.120
47. Pourcel C, Salvignol G, Vergnaud G, 2005. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology. 151(3), 653-663. https://doi.org/10.1099/mic.0.27437-0
48. Pyenson NC, Marraffini LA, 2017. Type III CRISPR-Cas systems: when DNA cleavage just isn't enough. Current Opinion in Microbiology. 37, 150-154. https://doi.org/10.1016/j.mib.2017.08.003

49. Rossum G, 1995. Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam.
50. Seed KD, Lazinski DW, Calderwood SB, Camilli A, 2013. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature. 494(7438), 489-491. https://doi.org/10.1038/nature11927
51. Šimoliūnas E, Vilkaitytė M, Kaliniene L, Zajančkauskaitė A, Kaupinis A, et al., 2015. Incomplete LPS Core-Specific Felix01-Like Virus vB_EcoM_VpaE1. Viruses. 7(12):6163-6181. https://doi.org/10.3390/v7122932
52. Soding J, Biegert A, Lupas AN, 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Research. 33(Web Server), W244-W248. https://doi.org/10.1093/nar/gki408
53. Soding J, Remmert M, Hauser A, 2012. HH-suite for sensitive protein sequence searching based on HMM-HMM alignment. Ludwig-Maximilians-Universitat Munchen. :46.
54. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, et al., 2019. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 20(1), 473. https://doi.org/10.1186/s12859-019-3019-7
55. Takeshita D, Sato M, Inanaga H, Numata T, 2019. Crystal Structures of Csm2 and Csm3 in the Type III-A CRISPR–Cas Effector Complex. Journal of Molecular Biology. 431(4), 748-763. https://doi.org/10.1016/j.jmb.2019.01.009
56. Tamulaitis G, Venclovas Č, Siksnys V, 2017. Type III CRISPR-Cas Immunity: Major Differences Brushed Aside. Trends in Microbiology. 25(1), 49-61. https://doi.org/10.1016/j.tim.2016.09.012
57. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
58. Villion M, Moineau S, 2013. Phages hijack a host's defence. Nature. 494(7438), 433-434. https://doi.org/10.1038/494433a
59. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ, 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 25(9), 1189-1191. https://doi.org/10.1093/bioinformatics/btp033
60. Zhang Z, Pan S, Liu T, Li Y, Peng N, 2019. Cas4 Nucleases Can Effect Specific Integration of CRISPR Spacers. Journal of Bacteriology. 201(12), 16. https://doi.org/10.1128/JB.00747-18
61. Zhao Y, Wang J, Sun Q, Dou C, Gy Y, et al., 2018. Structural insights into the CRISPR-Cas-associated ribonuclease activity of Staphylococcus epidermidis Csm3 and Csm6. Science Bulletin. 63(11), 691-699. https://doi.org/10.1016/j.scib.2018.03.017

62. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kubler J, et al., 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Journal of Molecular Biology. 430(15), 2237-2243. https://doi.org/10.1016/j.jmb.2017.12.007

Vilniaus universitetas
Gyvybės mokslų centras
Biomokslų institutas

Justina Kraujūnaitė
Magistro baigiamasis darbas

CRISPR-Cas sistemų baltymų homologai virusuose: paieška ir analizė bioinformatiniais metodais

SANTRAUKA

CRISPR-Cas sistemos – prokariotų įgytas imunitetas, teikiantis greitą bei efektyvią adaptaciją prie sparčiai besivystančių judriųjų genomo elementų (JGE). Kelios JGE klasės ne tik prisidėjo prie CRISPR-Cas kilmės bei evoliucijos, bet ir atvirkščiai, kai kurie JGE pasisavino CRISPR-Cas sistemas bei atskirus jų komponentus. Tarp JGE išskirtiniais laikomi virusai, kuriuose CRISPR-Cas sistemos bei jų komponentai tebėra menkai ištirti.

Šio darbo metu, taikant bioinformatinius metodus, atlikta CRISPR-Cas sistemų baltymų homologų paieška bei analizė virusuose. Naudojant profilio-sekos (HMMER) bei profilio-profilio (HH-suite) lyginimo paieškas, ištirtos 362 446 virusų koduojamos baltymų sekos, tarp kurių rasta 1 010 CRISPR-Cas sistemų baltymų homologų. Rezultatai parodė, jog virusai į savo genomus pagrinde, inkorporuoja Cas baltymus, susijusius su adaptacijos bei interferencijos CRISPR-Cas sistemų funkciniais moduliais. Sugrupavus bei išanalizavus rastas CRISPR-Cas baltymų homologų sekas, šiame darbe pirmą kartą virusuose charakterizuota Csm3 baltymų grupė, galimai apsauganti virusą nuo šeimininko CRISPR-Cas sistemų mechanizmo. Naujai rastos ir Cas5, Cas6 bei Cas7 baltymų grupės, kurių apibūdinimui reikalinga detalesnė analizė. Taip pat, nustatytos jau anksčiau virusuose identifikuotos Cas2, Cas3, Cas4 bei Cas12/14 baltymų grupės, kurios, tikriausiai, pritaikytos skirtingoms virusų funkcijoms atlikti, o pastaroji grupė, tikėtina, jog veikia arba kaip transpozazė, arba įeina į virusų CRISPR-Cas komplekso sudėtį.

Vilnius University
Life Sciences Center
Institute of Biosciences

Justina Kraujūnaitė
Master thesis

Bioinformatics Analysis of the CRISPR-Cas Systems Protein Homologs in Viruses

SUMMARY

CRISPR-Cas systems – prokaryotic acquired immunity that provides rapid and robust adaptation to the rapidly evolving mobile genetic elements (MGEs). Several classes of MGE not only contributed to the origin and evolution of CRISPR–Cas, but also, conversely, CRISPR–Cas systems and their components were recruited by various MGEs. Viruses are considered to be unique among other MGEs, with CRISPR-Cas systems and their components that remain largely uncharacterized.

In this study, bioinformatics analysis was performed in order to find and characterize CRISPR-Cas systems proteins homologs in viruses. Using profile-sequence (HMMER) and profile-profile (HH-suite) comparison searches, protein sequences encoded by 362 446 viruses were examined, of which 1 010 of the CRISPR-Cas systems proteins homologs were found. Viruses have been observed to incorporate Cas proteins into their genomes, mainly related to the adaptation and interference functional modules of the CRISPR-Cas systems. After grouping and analyzing the found sequences of CRISPR-Cas protein homologs, for the first time in this work the group of Csm3 proteins was characterized in viruses, possibly protecting the virus from the mechanism of host CRISPR-Cas systems. New groups of Cas5, Cas6 and Cas7 proteins have also been found, which require more detailed analysis to be described. Also, previously identified groups of Cas2, Cas3, Cas4, and Cas12/14 proteins have been identified in viruses, that are likely to be adapted to perform different viral functions, and the latter group is likely to act either as transposase or as part of the viral CRISPR-Cas complex.

PRIEDAI

Šio darbo priedus galima rasti interneto adresu https://bit.ly/CRISPR-Cas_virus.