Pētniecības pieteikuma vienošanās Nr. ***1.1.1.2/VIAA/1/16/094***

Pusgada populārzinatnisks pārskats (1.09.2017-28.02.2018.) par projekta īstenošanas gaitu "Transponējamo elementu variāciju izpēte parastās priedes (*Pinus sylvestris* L.) gēnu rajonos"

Projekta pirmais posms ir veltīts priedes references genomu bioinformātiskai analīzei ar mērķi identificēt gēnus, kas satur transponējamos elementus (TE) un to potenciāli funkcionālus struktūrelementus. Zināms, ka TE var lokalizēties gēnu sekvencēs un ietekmēt gēnu darbības regulāciju. No lokalizācijas ir atkarīga TE iespējama funkcija un nozīme, priedes genoms satur ap 75% dažādu TE, tomēr tikai nelielai daļai ir potenciāla funkcionāla nozīme. Tāpēc pamatojoties uz citu augu genomu pētījumiem tiks izmēģinātas dažādas pieejas un analīzes metodes nozīmīgo TE identificēšanai. Pētījumā izmantoju pieejamos priedes references genomus *Pinus taeda* un *Pinus lambertiana* sugām. Kailsēkļu genomi raksturojas ar milzīgu izmēru, tādēļ lai veiktu šo genomu analīzi ir nepieciešami augstas kapacitātes datoru resursi, kā arī iemaņas darbam ar specifiskām bioinformātiskām programmām. Šī iemesla dēļ bioinformātiskās analīzes tika veiktas Zviedrijas Lauksaimniecības zinātņu Universitātes Augu bioloģijas laboratorijā, kur ir pieejami resursi no *UPPMAX* (*Uppsala Multidisciplinary Center for Advanced Computational Science*). *UPPMAX* satur vairākas augstas kapacitātes darba klāsterus ar daudzām instalētām un darba kārtībā esošajām bioinformātiskām programmām. Piemēram, viens no klāsteriem satur 6080 kodolus ar 334 mezgliem; 32 lielākie mezgli satur 256 GB atmiņas, bet pārējie vēl 128 GB, neskaitot papildus vietu paredzēto datu glabāšanai. Lai noskaidrot visa genoma gēnus, kas satur šādus pārkartojumus, izolēju gēnu sekvences, kas satur eksonus un intronus, kā arī šo gēnu flankējošās sekvences no 5' un 3' rajona 5kB garumā. Tādā veidā izveidoju četras sekvenču datubāzes. Lai identificētu TE izmantoju skujkoku atkārtojumu datubāzi *PIER* v.2.0., kas satur 19700 ierakstu garumā no 257 līdz 35042 bāžu pāriem. Darba gaitā konstatēju, ka šī datubāze satur daudz neskaidru datu, jo tā tika veidota pamatojoties uz automātisku TE anotēšanas metodi. Tā kā šīs kļūdas var negatīvi iespaidot mūsu pētījuma rezultātus izmantoju *CD-Hit* v.4.6.4 programmu sekvenču klāsterizēšanai un reprezentatīvo sekvenču atlasei. Papildus izolēju funkcionālus TE struktūrelementus, kas ļauj samazināt TE sekvenču garumu un paaugstināt analīzes precizitāti. Taču šie struktūrelementi nav sastopami visos TE tipos. Tādēļ sekvenču salīdzināšanas analīze tika veikta gan ar pilna izmēra TE sekvencēm, gan ar to daļām. Rezultātā ieguva piecas sekvenču bibliotēkas katram analizētajam genomam, katra bibliotēka satur vairākus simtus TE, kas sakrīt ar noteikto gēnu introniem vai gēnu flankējošām sekvencēm. Izplatītāko TE salīdzinājums starp *P.taeda* un *P.lambertiana* genomiem parādīja, ka šīm sugam visizplatītākās TE ģimenes atšķiras. T.n. izplatītākie *P.taeda* gēnos TE nav

visizplatītākie arī *P.lambertiana* gēnos un otrādi. Bet tika identificēti arī tādi TE, kas ir izplatīti abu genomu gēniem. Gēnu kopas, kas satur vienas ģimenes TE, abām sugam ir iesaistīti līdzīgos funkcionālos procesos, tomēr gēni nav homologi. Tika identificēti arī niecīgs skaits (1-3 no 400 gēniem) homologo gēnu, taču tie nav anotēti un tādēļ nav zināma šo gēnu funkcija. Šo atšķirību var skaidrot ņemot vērā, ka *P.taeda* un *P.lambertiana* pārstāv dažādas priežu apakšģintis, kas nodalījās viena no otras aptuveni pirms 17 mlj gadiem, bet TE transpozīcijas laika mediāna šīm sugām ir jaunāka, tātad parsvarā transpozīcijas ir notikušas jau pēc nodalīšanās. No literatūras ir zinams, ka TE iespējams ir iesaistīti sugu specifikācijā, kas tiek pēdējā laikā pieļauts pamatojoties uz citu organismu genomu izpēti. Iepriekš tika salīdzināts egles un priedes ģintis genomi, kas satur vēl vairāk atšķirību nekodējošos gēnu rajonos un arī TE kompozīcijā, kamēr gēnu kodējošās daļas ir konservatīvas. Mūsu iepriekšējais pētījums analizēja TE izplatību visā genomā un jaunie rezultāti apstiprina līdzīgu tendenci arī gēnu rajoniem.

Liela nozīme rezultātu interpretācijā ir arī genomu atšķirīgā kvalitāte. Pētījuma gaitā atklājās, ka *P.taeda* genoma versijai 2.0. ir atrodams ievērojami mazāks pilno TE skaits gēnu intronos, nekā *P.taeda* genoma pirmajai versijai. Tomēr šo pašu TE struktūrdaļu izplatība starp abām genomu versijām ir salīdzināma. Līdzīgi rezultāti ir iegūti arī salīdzinot ar *P.lambertiana* v.1.0. Tas skaidri norāda uz to, ka rezultātu atsķirību izcelsme ir *P.taeda* genoma versiju atšķirīgās salikšanas metodes. Tomēr, šīs kļūdas ievērojami samazinās, ja analīzei izmanto nevis pilna izmēra TE, bet to strukturalās daļas- garos terminalos atkartojumus, jeb LTR. Šī pieeja nav iepriekš plaši izmantota, bet ir perspektīva priedes genoma datiem, jo tieši šīs struktūrdaļas satur nozīmīgus funkcionālus signālus (cis-elementus, alternatīvos promoterus, transkripcijas faktorus, transkripcijas terminācjas signālus), turklāt, tie ir īsāki un tādēļ ir vieglāk un precīzāk identificējami esošajos datos. Sākotnējie rezultāti pasvītro šajā projektā plānoto turpmāko uzdevumu lietderīgumu, kaur atrastie polimorfismi tiks pārbaudīti ar precīzākām sekvenēšanas metodēm, jo jaunās paaudzes sekvenēšanas dati pagaidām nevar sniegt pārliecinošu atbildi par pilna vai daļēja TE atrašanos gēna rajonā vienam indivīdam, nerunājot par vairāku indivīdu atšķirībām.

Lai saprast, vai atrastie gēni, kas satur vienādu TE, ir saistīti funkcionāli, izmantoja gēnu tīklu analīzi. Priedēm gēnu funkcionālā analīze pamatojas uz homoloģiju ar vairāk izpētītiem augu gēniem, un vairāk ka puse no priežu gēniem nav anotēta un to funkcija nav zināma. Iespējams, ka tieši šie sugai specifiskie gēni ir iesaistīti pielāgošanās un rezistences procesos. Tāpat jaunākajai *P.taeda* genoma versijai gēnu anotācija pagaidām nav pieejama un nācās optimizēt pieejamās iespējas konkrētā mērķa sasniegšanai. Tomēr arī pieejamās informācijas analīze rezultējās ar

interesanto gēnu tīklu identificēšanu. Sākotnējās vispārējās bioinformātiskās analīzes rezultāti tika apkopoti un prezenēti Latvijas Universitātes 76. Konferencē Molekulārās bioloģijas sekcijā 2018.gada 2.februārī. Referāta nosaukums: "Mobilo ģenētisko elementu izplatība priežu gēnu rajonos" (autori: Angelika Voronova (LVMI Silava), Martha Rendon (Zviedrijas Lauksaimniecības Zinātņu universitāte), Par Ingvarsson (Zviedrijas Lauksaimniecības Zinātņu universitāte) un Dainis Ruņģis (LVMI Silava)). Referāta tēzes tika iesniegtas publicēšanai brīvi pieejamā žurnālā *Environmental and Experimental Biology* (http://eeb.lu.lv/about.shtml). Tika sagatavots un publicēts populārzinatniskais raksts par šo petījumu avīzei "Zinātnes vēstnesis" 2018. gada 26. Marta numuram 6 (548), ISSN 1407-6748: "TRANSPONĒJAMO ELEMENTU VARIĀCIJU IZPĒTE PARASTĀS PRIEDES (*PINUS SYLVESTRIS* L.) GĒNU RAJONOS".

Bioinformātiskā datu analīze turpinās. Lai identificēt, vai ir atrodamas tādas TE ģimenes, kas biežāk ir sastopamas gēnu tuvumā, tika izstrādāta darbplūsma, kā rezultātā sadalīju gēnu flankejošās sekvences posmos un analizēju šos posmus atsevišķi ar TE LTR kopu. *P.taeda* v.2. genomam (36730 gēnu) atrasta būtiska astoņu TE ģimeņu paaugstinātā frekvence gēnu flankējošajās sekvencēs, kas atrodas 1-2 kB attālumā no gēna (atsevišķi tika analizēti arī dažādi posmi- gēna 5' un 3' daļā). Turklāt, tika atrasta viena TE ģimene (*PtRXX_4619*), kas preferenciāli atrodas gēnu 3' daļā. Attālākajos gēnu flankējošās sekvences posmos (3, 4 un 5kB attālumā) vairs nav šādas tendences, t.n. aptuveni vienāds skaits TE atrodas dažādos gēnos. Šie rezultāti pierāda, ka atsevišķu TE ģimeņu izplatība gēnu flankējošās sekvencēs nav neitrāla. Turpinajumā izmantojot iztrādāto metodi, tiks veikta arī *P.taeda* v.1.0. un *P.lambertiana* genomu analīze.

Sagatavoja: Angelika Voronova.