Männyn (*Pinus sylvestris*) LTR-retrotransposonit ja niiden ekspressio eri solukoissa

Iida Ronkainen

Pro gradu -tutkielma
Oulun yliopisto
Biologian tutkinto-ohjelma
Marraskuu 2021

Sisällys

Tiivistelmä

LTR-retrotransposonit (long terminal repeat retrotransposons) kuuluvat transposoituvien elementtien luokkaan I. LTR-retrotransposonit siirtyvät genomissa paikasta toiseen RNA-välivaiheen kautta. Siirtymisellä voi olla useita vaikutuksia genomin toimintaan. LTR-retrotransposoneiden aktiivisuus voikin johtaa esimerkiksi geenien hiljenemiseen. Mahdollisesti haitallisten vaikutuksien vuoksi transposoituvien elementtien aktiivisuutta vastaan on kehittynyt epigeneettisiä menetelmiä, jotka hillitsevät transposoituvien elementtien aktiivisuutta. Kaikkia transposoituvia elementtejä ei kuitenkaan hiljennetä ja joskus ne voivat olla isännälleen myös hyödyllisiä. Kasvien genomeista merkittävä osa koostuu transposoituvista elementeistä. Suurin transposoituvien elementtien ryhmä kasveilla on LTR-retrotransposonit. Myös männyillä (*Pinus*) on todettu esiintyvän runsaasti LTR-retrotransposoneita.

LTR-retrotransposoneita voidaan löytää DNA-sekvensseistä käyttäen eri ohjelmistoja. Ohjelmistojen toiminta perustuu mallinnukseen LTR-retrotransposonin rakenteesta. Eri ohjelmistojen avulla pystytään myös luokittelemaan löydetyt LTR-retrotransposonit tarkemmin superperheisiin ja perheisiin.

Työ perustuu aiemmin maissille luotuun menetelmään, jolla saadaan etsittyä LTR-retrotransposoneita. Työn tarkoituksena oli ottaa selville, voidaanko alun perin maissille suunniteltua LTR-retrotransposoneiden hakumenetelmää käyttää myös havupuille. Työssä aineistona on käytetty sekä männyistä saatua genomisekvenssiaineistoa sekä transkriptien ekspressioaineistoa.

Mäntyjen genomiaineistosta on etsitty LTR-retrotransposonit bioinformaattisia menetelmiä käyttäen. Aluksi on etsitty missä LTR-retrotransposonit sijaitsevat, sitten niiden sisäisiä ominaisuuksia on annotoitu sekä yritetty löytää tarkempi annotaatio elementeille jo olemassa olevasta tietokannasta. Jos tietokannasta ei löytynyt annotaatiota, on annotaatio tehty käyttämällä hyödyksi klusterointimenetelmiä sekä useita eri ohjelmistoja.

Lisäksi LTR-retrotransposonit on etsitty männyn transkriptien ekspressioaineistosta. Perustuen aiemmin julkaistuun ekspressiotutkimukseen männyllä, on otettu selville missä solukoissa löytyneet LTR-retrotransposonit ovat olleet ylössäädeltyjä sekä millainen niiden ekspressiotaso eri solukoissa on ollut.

1. Johdanto

Transposoituvat elementit ovat DNA:ta, joka kykenee siirtymään uuteen paikkaan genomissa. Usein siirtymisprosessi johtaa elementtien duplikoitumiseen (Feschotte, Jiang, & Wessler, 2002). Transposoituvat elementit pystyvät myös replikoitumaan isännässään itsenäisesti (Wells & Feschotte, 2020). Transposoituvat elementit pystyvät siirtymään genomissa uuteen paikkaan ja ne voivat liikkua myös lajien välillä (Burt & Trivers, 2008, s. 228). Täten ne ovat levinneet lähes kaikkiin eukaryootteihin (Wicker ym., 2007). Transposoituvat elementit voivat muuttaa kromosomien rakennetta (chromosomal rearrangements). Niistä voi silti olla myös hyötyä isäntäeliölle valinnan suosimina (Burt & Trivers, 2008, s. 229). Barbara McClintock löysi transposoituvat elementit maissista vuonna 1950 (McClintock, 1950).

Merkittävä osa kasvien genomeista voi koostua transposoituvista elementeistä, sillä niiden määrä vaihtelee *Arabidopsis thalianan* 15 %:n ja *Liliaceaen* 90 %:n välillä (Sabot & Schulman, 2006). Kasvien transposoituvista elementeistä suurin osa on LTR-retrotransposoneita (Grandbastien, 2015), joista osa on todennäköisesti kehittynyt infektiokyvyn menettäneistä retroviruksista (Boeke & Stoye, 1997).

1.1. Työn tutkimuskysymykset ja hypoteesit

Työn tarkoituksena on keskittyä LTR-retrotransposoneiden esiintymiseen männyillä ja tarkastella missä solukoissa LTR-retrotransposonit ekspressoituvat. Tutkimuskysymyksenä työssä on, soveltuvatko koppisiemenisille tehdyt transposoituvien elementtien annotaatiomenetelmät havupuille. Työn hypoteeseina ovat, että useimmat löydetyt LTR-retrotransposonit eivät ole aktiivisia, ja että transkriptien ekspressiotaso eri solukoiden välillä vaihtelee. Lisäksi työssä käsitellään LTR-retrotransposoneiden biologista merkitystä männyille.

1.2. Transposoituvat elementit ja niiden esiintyminen genomissa

Transposoituvat elementit (TE) voidaan jakaa luokkiin I ja II. Kyseinen Finneganin kehittämä luokkajako on ollut käytössä vuodesta 1989 asti. Transposoituvan elementin sijoittuminen luokkaan perustuu siihen, onko siirtymisessä välimolekyylinä RNA vai DNA. Luokassa I eli retrotransposoneissa välimolekyylinä on RNA. Luokassa II eli DNA-transposoneissa välimolekyylinä toimii DNA (Finnegan, 1989; Wicker ym., 2007). Wicker ym. (2007) ovat kehittäneet transposoituvien elementtien jaottelua eteenpäin. Heidän mukaansa jako tulee tehdä vielä kahden pääluokan jälkeen alaluokkiin, lahkoihin sekä superperheisiin.

Luokkaan I eli retrotransposoneihin voidaan luokitella muun muassa käänteiskopioijan fylogenian perusteella viisi lahkoa. Lahkot ovat LTR (long terminal repeat), LINE (long interspersed nuclear element), SINE (short interspersed nuclear element), DIRS (*Dictyostelium* intermediate repeat sequence) ja PLE (*Penelope*-like element). Lahkot puolestaan sisältävät superperheitä. Esimerkiksi LTR-lahko sisältää *Copia*- ja *Gypsy*-superperheet, joiden erona on *pol*:in *rt*- ja *int*-geenien sijainti. Lisäksi LTR-lahko sisältää *ERV*-, *Retrovirus*- sekä *Bel-Pao*-superperheet (Wicker ym., 2007). Kasveilla merkittävimmät superperheet ovat *Gypsy* ja *Copia* (Friesen, Brandes, & Heslop-Harrison, 2001).

LTR-retroelementit koodaavat yleensä yhteensä viittä proteiinia, joista kaksi on rakenneproteiineja ja kolme entsyymejä. LINE:t puolestaan koodaavat usein kahta proteiinia (Burt & Trivers, 2008, s. 230–239). Avoin lukukehys (ORF, open reading frame) tarkoittaa lukukehystä, jonka kodonit koodaavat aminohappoja (Krebs, Goldstein, & Kilpatrick, 2017, s. 29). LINE:jen ensimmäinen ORF sitoo RNA:ta (Kolosha & Martin, 1997) ja toinen ORF koodaa käänteiskopioijaa (RT) sekä DNA-endonukleaasia (EN, Weiner, 2002). SINE:t puolestaan eivät koodaa yhtäkään proteiinia. SINE:t käyttävät siirtyessään hyödykseen LINE:jä (Burt & Trivers, 2008, s. 238).

Transposoitumismekanismi sisältää kaikilla luokan I transposoituvilla elementeillä RNA-välivaiheen. Luokassa I TE:n lisääntyminen alkaa siten, että RNA muodostetaan transkription avulla käyttämällä transkription templaattina genomissa esiintyvää TE:tä. Tämän jälkeen RNA käänteiskopioidaan DNA:ksi. Tämä siis tarkoittaa, että kun kierto käydään yhdesti läpi, TE-kopioiden määrä lisääntyy yhdellä (Wicker ym., 2007). Menetelmä toimii siis kopioi ja liitä -mekanismilla (Sabot & Schulman, 2006).

Luokassa II tarvitaan DNA-välivaihe, jotta transposoituminen voi tapahtua. Tällöin transposaasi leikkaa transposoituvan elementin sekä liittää sen uuteen kohtaan genomissa (Bennetzen & Wang, 2014). Luokkaan II eli DNA-transposoneihin kuuluu kaksi alaluokkaa. Ensimmäiseen alaluokkaan kuuluvat lahkot TIR (terminal inverted repeat) sekä vain tyrosiinirekombinaasia koodaava Crypton. TIR:ien luokittelu superperheisiin perustuu niiden päissä oleviin käänteisiin toistojaksoihin sekä kohdealueen duplikaatioon (target site duplication, TSD), joka on TE:n liittymisessä syntyvä lyhyt toistoalue. TIR:eihin kuuluvat muun muassa superperheet *Mutator* ja *PiggyBac*. DNA-transposoneiden toiseen alaluokkaan kuuluvat puolestaan lahkot Helitron ja Maverick sekä saman nimiset superperheet (Wicker ym., 2007). DNA-transposonit ovat yksinkertaisimpia transposoituvia elementtejä. Pituudeltaan DNA-transposonit ovat yleensä 1–10 tuhatta emästä. Lisäksi ne koodaavat useimmiten yhtä

proteiinia, transposaasia (Burt & Trivers, 2008, s. 230). DNA-transposoneita löytyy lähes jokaiselta eukaryootilta (Wicker ym., 2007).

Erilaiset transposoituvat elementit sijaitsevat eri osissa genomia. Maissilla alhaisen rekombinaation alueelta löytyy usein LTR-retrotransposoneita ja ei-LTR-retrotransposonit puolestaan sijaitsevat rekombinaation kannalta aktiivisimmalla alueella (Stitzer, Anderson, Springer, & Ross-Ibarra, 2021). Heslop-Harrison ym. (1997) ovat esittäneet useita syitä, miksi retrotransposoneita esiintyy joillakin alueilla vähemmän. Esimerkiksi on mahdollista, että tietylle alueelle liittyessään ne olisivat letaaleja tai estäisivät yksilön lisääntymisen. Voi myös olla, että tietyille genomin alueille liittyessään retroelementit poistetaan tai mahdollista insertiokohtaa ei ole. Myös DNA:n konformaatio voi estää liittämisen (Heslop-Harrison ym., 1997).

Vastaavasti on useita syitä sille, miksi joiltakin alueilta löytyy enemmän retroelementtejä. Yksi syy voi olla, että retrotransposonit ovat olennainen osa genomia. Toinen syy on, että elementit eivät välttämättä lisäänny käänteiskopioinnin kautta vaan esimerkiksi epätasaisen tekijäinvaihdunnan (unequal crossing over) avulla. Kolmas syy voisi olla, että genomissa on tiettyjä alueita, jotka suosivat elementtien liittymistä (Heslop-Harrison ym., 1997).

1.3. LTR-retrotransposonit

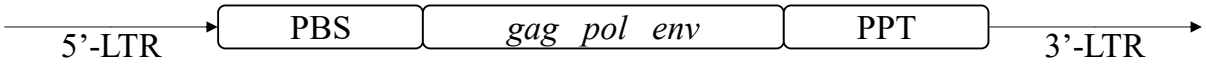
1.3.1. Rakenne ja toimintatapa

LTR-retrotransposonit sisältävät kaksi ORF:ia. *Gag*:in tehtävänä on tuottaa proteiinia, jota tarvitaan nukleokapsidin muodostamiseen ja siten käänteiskopioimiseen. Pol puolestaan on polyproteiini, joka sisältää asparagiiniproteinaasi-domeenin, käänteiskopioijan (reverse transcriptase, RT) ja RNAasi H:n. Yhdessä RT:n ja RNAasi H:n tehtävänä on käänteiskopiointi. Lisäksi pol sisältää myös integraasin, jonka tehtävänä on liittää uusi LTR-retrotransposonikopio genomiin (Sabot & Schulman, 2006).

LTR-retrotransposonit voidaan jakaa eri ryhmiin ORF:ien läsnäolon perusteella. Luokittelussa tarkastellaan, onko ryhmällä *gag*- tai *pol*-ORF:eja. *Gypsy* sisältää Gag:in ja retrovirusten kuoridomeeni-ORF:ia (envelope domain) muistuttavan ENV-kaltaisen domeenin (Sabot & Schulman, 2006). Retroviruksilla *env*-geeni mahdollistaa toisten solujen infektoimisen, koska geenin proteiini mahdollistaa retroviruksen poistumisen isäntäsolusta. Ei vielä tiedetä, mitä *env*-domeenit tekevät kasveilla soluissa tai niiden ulkopuolella (Schulman,

2013). ENV-kaltainen domeeni sisältää asparagiiniproteinaasi-domeenin (aspartic proteinase, AP), RT-RNAasi H:n sekä integraasin järjestyksessä 5'-päästä kohti 3'-päätä. *Copia* puolestaan sisältää samassa järjestyksessä *gag*:in sekä *ap*:n, mutta sen integraasi ja RT-RNAasi H sijaitsevat toisinpäin verrattuna *Gypsyyn* (Sabot & Schulman, 2006).

Kun uusia LTR-retrotransposoneita etsitään, etsityillä alueilla on omat tehtävänsä. LTR-retrotransposoneita etsiessä voidaan tunnistaa dimerisaatiosignaali (dimerization signal, DIS), alukkeen kiinnittymiskohta (primer binding site, PBS), paketointisignaali (packaging signal, PSI), polypuriinijuoste- (polypurine tract eli PPT) sekä integraasi eli INT-signaalit. PBS on apuna, kun syntetisoidaan cDNA:n miinusjuostetta. Plusjuosteen alukkeena puolestaan on PPT. Polypuriinijuoste sekä alukkeen kiinnittymiskohta ovat pituudeltaan 20–25 nukleotidia (Wicker ym., 2007). LTR-retrotransposoneilla esiintyy 5'- ja 3'-päissä pitkiä terminaalisia toistoalueita (LTR), joita tarvitaan transkriptiossa (Sabot & Schulman, 2006, Kuva 1).



Kuva 1. LTR-retrotransposonin rakenne. Perustuu artikkeleihin McCarthy & McDonald (2003) sekä Wicker ym. (2007).

LTR-retrotransposoneiden elinkierto muistuttaa retrovirusten kuten HIV:n elinkiertoa. Elinkierron alussa mRNA transkripoidaan ja Gag sekä Pol syntetisoidaan. Tämän jälkeen Pol jaetaan osiin käyttämällä apuna AP:ta. Osista muodostuu AP, RT-RNAasi H sekä integraasi. RNA dimerisoidaan kissing-loop -mekanismilla ja pakataan. Sitten käänteiskopiointi alkaa. Kun ensimmäinen cDNA-juoste on syntetisoitu, RNA hajotetaan ja toinen juoste muodostetaan. Integraasi liitetään LTR:iin ja syntyneet juosteet voidaan liittää genomiin uuteen paikkaan (Sabot & Schulman, 2006).

LTR-retrotransposoneiden transposoitumista kontrolloidaan useassa vaiheessa. Transkription aloitus tapahtuu 5'-päässä sijaitsevien U3- sekä R-domeenien välistä. Transkription lopetus tapahtuu 3'-päässä sijaitsevien R- ja U5-domeenien välissä. Proteiinisynteesin aikana kontrolloidaan, paljonko Gag- ja Pol-proteiineja syntetisoidaan suhteessa toisiinsa. Gag-proteiinia tarvitaan, jotta saadaan tehtyä viruksenkaltaisia partikkeleita (virus-like particles, VLPs). Kun viruksenkaltaisia partikkeleita kootaan, proteaasi aktivoidaan ja se johtaa Gag-Pol -polyproteiinin muodostumiseen. Käänteiskopiointivaihetta säädellään isännän tRNA:n saatavuuden avulla. Käänteiskopioinnin seurauksena syntyy DNA:ta, jonka päissä esiintyvät identtiset toistojaksot. Muodostuneen DNA:n päitä käsitellään ja retrotransposoni liitetään osaksi isännän DNA:ta (Grandbastien, 1998).

1.3.2. Retrotransposoneiden alkuperä sekä leviäminen

Kaikilla retroelementeillä on RT-domeeni, jota apuna käyttäen voidaan tutkia elementtien alkuperää. RT-domeenien perusteella on rakennettu fylogeneettinen puu, jonka perusteella RNA-viruksille sekä retroelementeille löytyy yhteinen kantamuoto. Kaikki retroelementit eivät kuitenkaan varsinaisesti ole kehittyneet RNA-viruksista. Retroelementeillä on havaittu kaksi kehityslinjaa. Toisesta kehityslinjasta ovat muodostuneet ei-LTR-retrotransposonit ja muita ryhmiä kuten ryhmän II intronit. Ensimmäisestä kehityslinjasta puolestaan ovat kehittyneet hepadnavirukset, jonka jälkeen LTR:t ovat kehittyneet ja LTR-retrotransposoneiden *Copia*-ryhmä on muodostunut. Caulimoviruksilla ja LTR-retrotransposoneiden *Gypsy*-ryhmällä on myös yhteinen kantamuoto. Lisäksi samasta kehityslinjasta ovat kehittyneet retrovirukset, kun niille on muodostunut solusta poistumisen mahdollistava *env*-geeni (Xiong & Eickbush, 1990).

Transposoituvia elementtejä voidaan nykyisin havaita lähes kaikissa eukaryooteissa (Wicker ym., 2007). Vertikaalinen geenien siirtyminen viittaa vanhemmalta jälkeläiselle tapahtuvaa geenien siirtymistä, kun taas horisontaalinen viittaa yksilöiltä toisille tapahtuvaa siirtymistä (Deniz, Frost, & Branco, 2019). Kasveilla transposoituvien elementtien levittäytyminen on voinut tapahtua horisontaalisesti kolmella eri tavalla. Yksi tapa on, että kasvit ovat lähekkäisessä kontaktissa ja DNA:ta siirtyy kasvilta toiselle (Fortune, Roulin, & Panaud, 2008). Siirto tapahtuu tällöin parasiittisesta kukkivasta kasvista isäntäkasviin (Mower, Stefanović, Young, & Palmer, 2004). Toinen tapa on horisontaalinen transposoituminen hybridin kautta. Siinä kaksi eri lajin kasvia, josta toisessa on TE, ensin hybridisoituvat, TE aktivoituu ja tapahtuu takaisinristeytys. Takaisinristeytyksen avulla saadaan aikaan TE:n siirtyminen lajilta toiselle. Kolmas tapa TE:n siirtymiseen kasvien välillä on vektorin kautta. Kasveilla ei kuitenkaan tiedetä tiettyä vektoria, joka aiheuttaa tämän (Fortune ym., 2008). Tiedetään kuitenkin, että hyönteiset, jotka purevat kasveja tai laajalle levinneet virukset voivat saada aikaan geenien siirtymisen horisontaalisesti (Friesen ym., 2001). Vektorin kautta tapahtuvassa TE:n siirtymisessä TE aktivoituu ensimmäisessä lajissa ja siirtyy virukseen, joka puolestaan siirtyy toisen lajin yksilöön. Tämän jälkeen viruksessa oleva TE saadaan integroitua toisen lajin genomiin ja täten TE on siirtynyt lajilta toiselle (Fortune ym., 2008).

1.3.3. TE:hin kohdistuvat valintapaineet ja genomin koon muutokset

Vaikka genomista merkittäväkin osa voi olla transposoituvia elementtejä, ne eivät kuitenkaan välttämättä ole aktiivisia eli ne eivät kokoaikaa transposoidu tai tuota mRNA:ta. Syynä

epäaktiivisuuteen on transposoitumisen kyky aiheuttaa mutaatioita. Epäaktiivisuus voi johtua joko transkription estämisestä tai siitä, että elementit ovat genomissa lyhyinä sirpaleisina osina (Fultz, Choudury, & Slotkin, 2015). Baucom ym. (2009) tutkivat LTR-retrotransposoneissa tapahtuvaa valintaa riisillä. Löydetyt LTR-retrotransposonit esiintyivät epigeneettisesti hiljentyneinä ja ne oli hiljennetty jo silloin kun ne olivat uusia (Baucom, Estill, Leebens-Mack, & Bennetzen, 2009).

LTR-retrotransposoneihin kohdistuu riisillä monenlaista valintaa. Asiaa on tutkittu nukleotididiversiteetin avulla. Tuloksista voitiin päätellä, että luultavasti LTR-retrotransposoneiden geeneihin vaikuttaa pääosin puhdistava valinta (purifying selection), koska variaatio sijaitsee synonyymisissä nukleotidikohdissa (Baucom ym., 2009).

Riisin LTR-retrotransposoneista on löytynyt myös mahdollisesti valinnan pyyhkäisyjen jälkiä (selective sweeps). Tämä tulee ilmi siitä, että tutkijat löysivät positiivisen korrelaation geenidiversiteetin ja insertioajan välillä. Kun geenidiversiteettiä mitattiin synonyymisen diversiteetin avulla, pyyhkäisyt tulivat ilmi *Gypsy*-perheiden käänteiskopioija- ja integraasigeeneissä. Vanhemmat perheet kokevat luultavasti positiivista valintaa tai heikompia valinnan rajoituksia, koska ei-synonyymisten ja synonyymisten kohtien suhde kasvaa perheen iän mukaan. Todennäköisesti kyse on heikommista valinnan rajoituksista, koska kyseisten kohtien suhde on enimmäkseen alle yksi (Baucom ym., 2009).

Riisillä tutkittiin myös *Copia*- ja *Gypsy*-superperheissä olevien integraasin ja käänteiskopioijaan liittyvää valintaa. Suurin osa sekvensseistä oli tutkimusten mukaan puhdistavan valinnan alla (Baucom ym., 2009). On todettu, että *Copia*-elementit eivät ole genomissa jatkuvasti aktiivisia, vaan ne aktivoituvat ajoittain. Genomin koon muutokset liittyvät siihen, kuinka pitkiä nämä aktiiviset kaudet ovat olleet elementeillä sekä siihen, miten nopeasti retrotransposoneista syntynyttä DNA:ta poistetaan genomissa (Wicker & Keller, 2007). Genomikoon suurenemista on tutkittu kasveilla ja genomikoon muutos on yhdistetty siihen, kuinka mittava transpositionaalinen vaihe on ollut. Jokaisesta tutkitusta lajista (*A. thaliana*, *A. lyrata*, viiniköynnös, soijapapu, riisi, aroluste, durra ja maissi) löydettiin retrotransposoitumista. Aktiivisuutta lähiajoilta ei kuitenkaan löydetty kuin muutamalta LTR-retrotransposoniperheeltä. On siis voitu esittää TE:iden genomievoluutiomalli, jonka mukaan yksi tai muutama retrotransposoniperhe ovat aktiivisia vaiheittain. Tutkimuksessa eri lajeilla oli eri genomikoko (El Baidouri & Panaud, 2013).

1.3.4. Epigenetiikka TE:n hiljentämisessä

On kaksi erilaista reittiä, jotka voivat aloittaa transposoituvien elementtien hiljentämisen. Reitit ovat homologiasta riippuva hiljentämisen aloitus (homology-dependent initiation of silencing) sekä homologiasta riippumaton hiljentämisen aloitus (homology-independent initiation of silencing). Erona reiteissä on, ovatko TE:n hiljentämisessä syntyneet pienet häiritsevät RNA:t eli siRNA:t (small interfering RNAs) homologisia aktiivisten transposoituvien elementtien kanssa vai eivät. siRNA:t toimivat homologiasensoreina, eli niiden avulla voidaan tunnistaa TE:itä, joilla on samanlaiset sekvenssit. Samanlaisten sekvenssien tunnistamisen avulla voidaan varmistaa, että saman sekvenssin omaavat aktiiviset TE:t hiljennetään (Fultz ym., 2015).

DNA-metylaatio on epigeneettinen prosessi, jossa S-adenosyylimetioniinilta siirretään metyyli sytosiinin 5-kohtaan DNA-metyylitransferaasien katalysoimana (Tollefsbol, 2017). DNA:n metylaatiolla on useita eri vaikutuksia genomissa. Ne voivat esimerkiksi säädellä geenien ekspressiota. Lisäksi DNA:n metylaatio voi vaikuttaa transposoneiden hiljentämiseen ja vaikuttaa kromosomien interaktioon. RNA:n ohjaama DNA-metylaatioreitti eli RdDM-reitti johtaa kasveilla DNA-metylaatioon. Metylaatioon tarvitaan kyseisellä reitillä useita proteiineja sekä pieniä häiritseviä RNA:ita (Zhang, Lang, & Zhu, 2018). Kasveilla metylaatio voi esiintyä eri sytosiinikonteksteissa, joita ovat CG, CHG sekä CHH. Tällöin H tarkoittaa joko nukleotidia A, T tai C (Gallego-Bartolomé, 2020).

Metylaatiota voidaan ylläpitää eri DNA-metyylitransferaasientsyymien avulla ja sitä voidaan poistaa passiivisen demetylaation kautta. Tämä tarkoittaa, että DNA-metylaatiota ei vain yksinkertaisesti ylläpidetä, jolloin DNA:n replikaation seurauksena DNA-metylaatio vähenee. Vaihtoehtoisesti metylaatiota voidaan vähentää DNA-demetylaasien avulla, jolloin kyseessä on aktiivinen DNA:n demetylaatio (Zhang ym., 2018).

1.3.5. Vaikutus geenien toimintaan

Transposoituvat elementit voivat vaikuttaa monin tavoin geenien toimintaan. Esimerkiksi ne voivat aiheuttaa säätelyyn liittyviä mutaatioita (regulatory mutations). Syynä voi olla se, että usein transposoituvissa elementeissä esiintyy säätteleviä osia (Bennetzen & Wang, 2014).

Transposoituvan elementin osia voi liittyä geenin promoottoriin. TE:n säätelyelementit voivat olla geenille epäsopivia, jolloin säätely on haitallista ja luonnonvalinta poistaa kyseisen alleelin. Toinen vaihtoehto on, että promoottoriin liittyy TE, joka on samanlainen kuin mitä promoottorissa on jo osana eli säätelydomeeni duplikoituu. Tällainen

tilanne voi olla neutraali. Jos lisäyksestä ei ole hyötyä, alkuperäinen säätelyalue voi poistua tai sitten lisätty TE poistuu. Ei siis voi päätellä suoraan, että TE on luonut säätelyä promoottoriin. Todellisuudessa on mahdollista, että usea TE on liittynyt alueelle ja on mahdollista sattumasta johtuen, että alkuperäinen säätelyalue poistetaan. Tilanteessa, jossa geenin promoottorialueella on alkuperäinen säätelydomeeni ja lisäksi liitetty neutraali TE, alkuperäinen domeeni tai uusi TE poistetaan 50 % todennäköisyydellä (Bennetzen & Wang, 2014).

Transposoituvat elementit voivat aiheuttaa useita erilaisia toiminnallisia sekä rakenteellisia muutoksia. Yhtenä vaikutuksena on insertionaalinen mutageneesi. Silloin transposoituva elementti insertoi itsensä eksoniin. Insertionaalinen mutageneesi voi myös tapahtua siten, että transposoituva elementti kiinnittyy voimistajaan (enhancer) tai repressoriin. Transposoituva elementti voi myös tuoda genomiin uutta tietoa luomalla uuden voimistajan tai repressorin. Lisäksi on mahdollista, että transposoituva elementti saa aikaan joko uuden transkription aloituskohdan tai uuden promoottorin. Transposoituvat elementit voivat aiheuttaa myös uuden silmukointikohdan (splice site), uudelleenjärjestäytymistä, retropositiota sekä transduplikaatiota. Retropositio tarkoittaa, että mRNA käänteiskopioidaan käyttäen apuna retrotransposonin entsyymiä. Tämän jälkeen tapahtuu integraatio uudelle paikalle. Transduplikaatio tarkoittaa, että transposoituva elementti saa aikaan geenin duplikaation, jonka jälkeen geeni siirtyy uuteen paikkaan. Lisäksi eksaptaatio ja epigeneettinen säätely ovat mahdollisia. Transposoituvan elementin tilanteessa eksaptaatio tarkoittaa sitä, että transposoituva elementti pikemminkin alkaisi edistää isännän menestystä (adaptive function) sen sijaan, että se pyrkisi edistämään omaa lisääntymistään (Lisch, 2013).

Kasveilta on löytynyt aktiivisia LTR-retrotransposoneita, jotka vaikuttavat fenotyyppiin (Xiao, Jiang, Schaffner, Stockinger, & van der Knaap, 2008; Studer, Zhao, Ross-Ibarra, & Doebley, 2011). *Rider*-nimisen LTR-retrotransposonin on todettu vaikuttavan tomaatin muotoon. On todettu, että kun *Riderin* transkriptio alkaa ensimmäisestä LTR:stä kromosomissa 10, sen transkriptio ei lopukaan toisen pään LTR:ään, vaan jatkuu eteenpäin. Kun mRNA, käänteiskopiointi ja cDNA ovat valmiita, *Rider* on siirtynyt kromosomissa 7 sijaitsevaan *DEFL1*-geenin intoniin. Seurauksena on tomaatin muodon muuttuminen pyöreästä soikeaksi (Xiao ym., 2008).

Myös maissista on löytynyt funktionaalinen transposoituva elementti. *Hopscotch*-retrotransposonin on havaittu vaikuttavan siihen, että domestikoidulla maissilla on enemmän apikaalidominanssia kuin edeltäjällään teosintillä. Transposoituva elementti on liittynyt *teosinte branched1* -geenin (*tb1*) säätelyalueelle, jossa se on aiheuttanut voimistunutta geenieksprssiota (Studer ym., 2011).

1.3.6. TE:n aktiivisuus kasvien solukoissa

Transposoituvien elementtien aktiivisuutta on tutkittu kasvien eri solukoissa. Transposoituvien elementtien ekspression havaittiin olevan siitepölyssä ylössäädeltyä, mutta muissa solukoissa transposoituvat elementit joko olivat vain vähän aktiivisia tai eivät olleet ollenkaan aktiivisia *Arabidopsisilla*. Lisäksi maissilla hiljennetyn TE:n havaittiin aktivoituvan uudestaan siitepölyssä (Slotkin ym., 2009). Myös riisillä eniten TE:n aktiivisuutta havaittiin siitepölyssä. Jonkin verran aktiivisuutta havaittiin myös muun muassa nuorissa ja vanhoissa lehdissä. Nuorissa juurissa aktiivisuutta ei havaittu, mutta vanhemmissa juurissa sitä havaittiin (Nobuta ym., 2007). Transposoituvien elementtien uudelleenaktivaatio siitepölyn kasvullisessa ytimessä (pollen vegetative nucleus) on luultavasti johtunut heterokromatiinin muodostumiseen liittyvien geenien sekä siRNA:n biogeneesigeenien alassäätelystä (Slotkin ym., 2009). Bioottinen sekä abioottinen stressi voi myös johtaa retrotransposoneiden aktiivisuuteen kasveissa. Myös ympäristötekijät, kuten protoplastin eristys ja solukkoviljely, voivat aiheuttaa aktiivisuutta (Grandbastien, 1998).

1.4. Havupuiden LTR-retrotransposonit

Gypsy- ja *Copia*-retroelementtejä esiintyy runsaasti paljassiemenisillä (Friesen ym., 2001), joihin kuuluvat havupuut, Gnetales, käpypalmut (cycads) ja *Ginkgo*. Havupuulajeja on noin 630 ja pohjoisella pallonpuoliskolla ne muodostavat suuria taloudellisesti merkittäviä havupuumetsiä. Myös monissa muissa ekosysteemeissä havupuut ovat avainlajeja ja ne muodostavat suuren määrän biomassaa sekä osallistuvat merkittävästi fotosynteesiin (Nystedt ym., 2013).

Nystedt ym. (2013) analysoivat kuusen (*Picea abies* (L.) Karst) genomia. LTR-retrotransposonit vastasivat suurinta osaa löydetyistä TE:istä. *Gypsyyn* kuuluvia elementtejä oli eniten (35 %) ja toiseksi eniten *Copiaa* (16 %, Nystedt ym., 2013). Cossu ym. (2017) tutkivat myös *P. abiesin* LTR-retrotransposoneita. Tutkimuksessa käytettiin käänteiskopioijan domeenia. Paralogisia *Copia*-sekvenssejä löytyi 670 ja paralogisia *Gypsy*-sekvenssejä 1410. Fylogeneettisten menetelmien avulla LTR-retrotransposoniryhmiä löytyi 23. Näistä suurin osa (16) kuului *Gypsyyn* ja loput 7 kuuluivat *Copiaan*. Evolutiivisesti ryhmät olivat samankaltaisempia ryhmien sisällä kuin niiden välillä (Cossu ym., 2017).

Wegrzyn ym. (2014) tutkivat loblollymännyn (*Pinus taeda* L.) genomia ja annotoivat sen. Kaiken kaikkiaan 58,8 % genomista koostui retroelementeistä ja 1,04 % DNA-

transposoneista. Lisäksi he löysivät toistojaksoja, joita ei luokiteltu. Määrällisesti kokopitkiä LTR-retrotransposoneita löytyi 179 367 kappaletta. Kokopitkät sekä osittaiset LTR-retrotransposonit kattoivat 41,68 % koko genomista. *Gypsy*- ja *Copia*-elementtejä löytyi lähes saman verran. *Gypsy*-elementtejä löytyi 2,5 Gbp, joka vastaa 10,98 % koko genomista ja *Copiaa* 2,1 Gbp eli 9,14 % koko genomista. Kokopitkiä *Gypsy*-elementtejä löytyi 1,14 % ja *Copia*-elementtejä 0,89 % genomista (Wegrzyn ym., 2014). Stevens yms. (2016) tutkivat sokerimännyn (*P. lambertiana* Dougl.) LTR-retrotransposoneita. Tuloksena oli, että *P. lambertianan* genomista 79 % on transposoituvia elementtejä, joista 67 % on LTR-retrotransposoneita (Stevens ym., 2016).

Copian ja *Gypsyn* insertoitumisaikoja tutkittiin havupuilla. Tutkimuksissa tuli ilmi, että *Copia*-perhe on laajentunut erityisesti katajassa (*Juniperus communis*) ja *Gypsy* euroopanmarjakuusessa (*Taxus baccata*). Aktiivisuutta ei löytynyt *P. abiesissa* edellisen viiden miljoonan vuoden aikana. Kun tutkittiin *P. abiesta* ja *P. glaucaa*, selvisi että tutkituista 68:sta insertiosta viisi tapahtui noin 13–20 miljoonaa vuotta sitten kehityslinjojen eroamisen jälkeen. 63 insertiota puolestaan tapahtui ennen kehityslinjojen eroamista. Johtopäätöksenä oli, että LTR-retrotransposoneiden on täytynyt kasaantua kymmenien tai satojen miljoonien vuosien aikana. Ilmeisesti elementtejä on poistettu genomista vain vähän ja lähinnä niiden lisääntyminen on johtunut insertioista (Nystedt ym., 2013). LTR-retrotransposoneiden mediaani insertoitumisaika *P. lambertianalla* oli tutkimuksessa 16 miljoonaa vuotta sitten (Stevens ym., 2016).

Nystedt ym. (2013) mukaan havupuiden genomin evoluutiossa on tapahtunut 12 kromosomin koon suureneminen, kun koppisiemeniset erosivat kehityslinjasta. Kromosomien laajentuminen on tapahtunut kuitenkin hitaasti, mutta tasaiseen tahtiin. Kromosomien laajentumisen syynä voi olla *Gypsyn* ja *Copian* aktiivisuus. Suuret intronit ovat syntyneet TE:iden aktiivisuuden vuoksi, kun TE:t ovat liittyneet geeneihin. Tämän lisäksi TE:n aktiivisuus on myös aiheuttanut pseudogeenejä muun muassa sen vuoksi, että TE:t ovat kopioituneet geenien keskelle (Nystedt ym., 2013).

2. Aineisto ja menetelmät

2.1. LTR-retrotransposoneiden etsintä

LTR-retrotransposoneita pystytään etsimään DNA-sekvensseistä käyttämällä tiettyjä ohjelmistoja. *LTRharvest*-ohjelmisto (Ellinghaus, Kurtz, & Willhoeft, 2008) käyttää aiempiin

tutkimuksiin (McCarthy & McDonald, 2003; Kalyanaraman & Aluru, 2006) pohjautuvaa mallia LTR-retrotransposonin rakenteesta. Ohjelma tarkistaa esimerkiksi LTR:n pituudet ja ohjelmaan voi syöttää minimi- ja maksimipituuden. Lisäksi ohjelma tarkistaa paljonko on pituus 3'-kohdealueduplikaation (TSD, target site duplication) ja 3'-LTR:n välillä. Ohjelma perustuu malliin, jossa LTR-retrotransposonissa on kohdealueduplikaatiot 5'-LTR:stä yläjuoksuun (upstream) sekä 3'-LTR:stä alajuoksuun (downstream). Tavallisesti nämä ovat hyvinkin identtisiä sekvenssiltään. Myös 5'- ja 3'-LTR:t ovat lähes identtisiä sekvenssiltään. Molemmat voivat kuitenkin evoluution myötä hieman muuttua (Ellinghaus ym., 2008).

LTRharvestin (Ellinghaus ym., 2008) käytössä on useita eri vaiheita. Alussa ohjelma käyttää hyödyksi GenomeToolsia (GenomeTools, 2014), joka tallentaa tietokoneelle indeksoidun muodon ohjelmaan syötetystä aineistosta. Aineiston indeksointi mahdollistaa eri parametrien kokeilun nopeasti. Tämän jälkeen etsitään samanlaisia toistoja sekvenssistä. Myöhemmin löydetyistä kandidaattipareista voidaan etsiä LTR-motiiveja, jotka ovat palindromisia tai TSD:tä, jos käyttäjä niin haluaa. Sitten jäljellä olevista kandidaattipareista tarkistetaan pituudet ja LTR:ien välimatka. Myös sekvenssien identtisyys tarkistetaan. Jäljellä olevat kandidaattiparit luokitellaan ohjelmiston mukaan LTR-retrotransposoneiksi (Ellinghaus ym., 2008).

LTRharvest on sopiva genomikooltaan suurten eliöiden tutkimiseen. Se ei kuitenkaan kykene löytämään elementtejä, joista puuttuu toisesta päästä LTR (solo LTRs) tai sellaisia LTR:iä, joissa on suuri insertio (Ellinghaus ym., 2008).

LTRdigestiä (Steinbiss, Willhoeft, Gremme, & Kurtz, 2009) voidaan käyttää mahdollisten LTR-retrotransposoneiden annotoimiseen. *LTRdigest* käyttää samankaltaista mallia LTR-retrotransposoneille kuin *LTRharvest*, mutta *LTRdigestin* mallissa on hieman lisäominaisuuksia, sillä ohjelmassa on otettu huomioon myös PBS, yksi proteiinidomeeni *gag:*iin, kolme proteiinidomeenia *pol:*iin ja PPT (Steinbiss ym., 2009).

2.2. PIER-tietokanta

Wegrzyn ym. (2013) ovat koostaneet PIER-tietokannan (Pine Interspersed Element Resource), jota voidaan käyttää esimerkiksi toistokirjastona, kun etsitään transposoituvia elementtejä homologiaan perustuen (Wegrzyn ym., 2014). Tietokanta rakennettiin siten, että *P. taedasta* eristetystä genomisesta DNA:sta koostettiin fosmideja. Tutkimuksessa oli myös 10 BAC-kloonia, jotka Sanger-sekvensoitiin ja koostettiin. Näihin aineistoihin suoritettiin toistoalueiden etsintä. Tandemitoistoalueet tunnistettiin BAC- ja fosmidisekvensseistä. Myös valkokuusen

(*P. glauca*) ja *T. mairein* BAC:it analysoitiin. Muokattu kasvien toistoaluetietokanta (CPRD) muodostettiin, joka sisälsi viisi puista aikaisemmin tunnistettua elementtiä sekä muista kasveista havaittuja elementtejä. Muodostetun toistoaluetietokannan avulla tunnistettiin homologian avulla levittäytyneet (interspersed) toistoalueet. Levittäytyneet toistoalueet tunnistettiin BAC:eista ja fosmideista myös sekvenssien linjautumiseen itsensä kanssa tai rakenteellisiin ominaisuuksiin perustuvalla *de novo* -menetelmällä. Kun levittäytyneet toistoalueet oli löydetty, homologiaan ja *de novo* -menetelmään perustuvat tulokset yhdistettiin vertaamalla niitä CPRD:n homologiseen tietokantaan. Annotaatio ja luokittelu tehtiin manuaalisesti perustuen Wicker ym. (2007) esittämään luokitteluun. Elementit luokiteltiin perheisiin ja uudet hyvälaatuiset toistoalueet tallennettiin PIER-tietokannaksi (Wegrzyn ym., 2013).

2.3. Transkriptomiikka

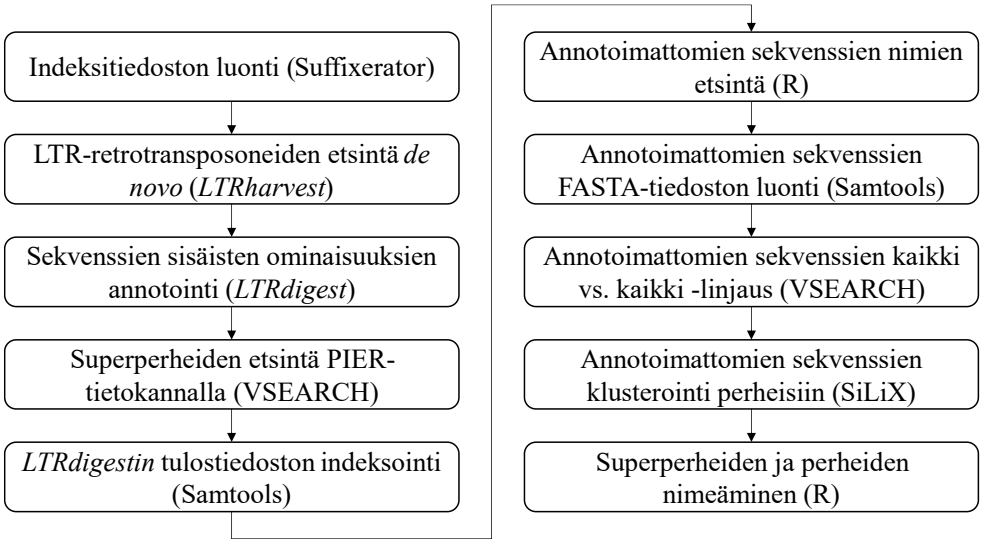
Soluissa tiettynä ajankohtana tuotettuja transkriptejä voidaan kutsua yhteisnimityksellä transkriptomi. Transkriptomin avulla voidaan tutkia genomin toiminnallisuutta. Yleisesti menetelmä toimii niin, että RNA:sta tehdään ensin cDNA-kirjasto. Tämän jälkeen luodaan transkriptomi, joka voidaan luoda linjaamalla sekvensoinnista saadut luvut (reads) joko ilman referenssigenomia (*de novo*) tai sen kanssa. Aineistosta on mahdollista saada selville geenien ekspression määrä (Wang, Gerstein, & Snyder, 2009).

2.4. Oma aineisto

Mäntynäytteet transkriptomi-FASTA:a varten on kerätty Punkaharjulta vuonna 2016. Näytteitä on kerätty viidestä eri solukosta ja tutkittavana oli kuusi yksilöä. Yksilöt eivät olleet keskenään sukua. Solukkonäytteet saatiin neulasesta, nilasta, kasvullisesta silmusta (vegetative bud), alkiosta ja megagametofyytistä. Alkiosta ja megagametofyytistä eristettiin mRNA ja neulasesta, silmusta ja nilasta eristettiin RNA, josta kerättiin mRNA. Sitten valmistettiin RNA-kirjasto ja poolattiin 6–12 kirjastoa Illumina NextSeq500:lla (pair-end) ja sekvensoitiin Illuminan Mid-Output Kit:llä (Ojeda ym., 2019). Lisäksi aineistona oli *P. lambertianan* v1.5 (https://treegenesdb.org/FTP/Genomes/Pila/v1.5/genome/, 2018), *P. taedan* v2.01 (https://treegenesdb.org/FTP/Genomes/Pita/v2.01/genome/, 2017) ja *P. sylvestrisin* (Jarkko Salojärvi, julkaisematon) genomit. Kaikki aineistot saatiin valmiiksi filtteröityinä ja käsiteltyinä FASTA-tiedostoina, jotka sisältävät otsikkorivin sekä sekvenssiaineiston.

2.5. LTR-retrotransposoneiden etsiminen

Työ suoritettiin käyttämällä erilaisia LTR-retrotransposoneita etsiviä ohjelmia. Menetelmät olivat samankaltaiset kuin mitä on aiemmin käytetty maissille (Anderson, Stitzer, Brohammer ym., 2019a). Työssä käytettiin hyödyksi maissin transposoituvien elementtien löytämiseen kehitettyjä menetelmiä (https://github.com/mcstitzer/w22_te_annotation/, 2018; Springer ym., 2018). Tämän työn menetelmät on esitetty Kuvassa 2.



Kuva 2. Työn menetelmät.

2.5.1. Indeksitiedoston luonti

P. sylvestrisin, *P. taedan* ja *P. lambertianan* genomiaineistot käsiteltiin yksi kerrallaan käyttäen GenomeToolsin sisältämää Suffixerator-ohjelmaa (GenomeTools, 2014), koska seuraavassa välivaiheessa käytettävä ohjelma vaatii lähtötiedostoksi indeksimuotoisen tiedoston (Ellinghaus, Kurtz, & Willhoeft, 2012). Parametreinä käytettiin -tis -suf -lcp -des -ssp -dna, joilla määritettiin millaisia output-tiedostoja luodaan (.tis, .suf, .lcp, .des ja .ssp) sekä se, että aineisto on syötetty ohjelmaan DNA:na. Lisäksi ohjelman muistinkäyttöä rajoitettiin parametrilla -memlimit 125GB tai 165GB.

2.5.2. LTR-retrotransposoneiden *de novo* -etsintä

Suffixeratorista (GenomeTools, 2014) tuloksena saadut indeksitiedostot syötettiin *LTRharvest*-ohjelmaan (Ellinghaus ym., 2008), jotta saatiin etsittyä LTR-retrotransposoneita *de novo*. Ohjelman ajossa käytettiin parametria -gff3, jotta saadaan tulokset GFF3-tiedostoon. GFF3-

tiedostomuoto on tekstitiedosto, joka sisältää yhdeksän saraketta. Tiedostomuodossa on esimerkiksi sarake, joka sisältää tunnisteen sekä alku- ja loppukoordinaatit (https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md, 2020). Lisäksi käytettiin muun muassa oletusparametreja -minlenltr ja -maxlenltr, jotka määrittävät haettavan LTR:n pituudeksi 100–1000. Oletusparametrit -mindistltr ja -maxdistltr etsivät LTR:ien aloituskohtia minimietäisyydellä 1000 ja maksimietäisyydellä 15000, sekä oletusparametri -similar tarkoittaa, että oletuksena samankaltaisuuden raja on ollut 85 %.

2.5.3. Sekvenssien sisäisten ominaisuuksien annotointi

Seuraavaa vaihetta varten käytettiin ensin GenomeToolsin (GenomeTools, 2014) gff3-ohjelmaa parametrilla -sort, jotta saatiin aiemmin luotu gff3-tiedosto järjestettyä löydetyn LTR-retrotransposonin sijainnin mukaan. *LTRdigestin* (Steinbiss ym., 2009) käyttöön tarvittiin HMM-profiilit, joiden avulla voidaan etsiä proteiiniperheiden tunnettuja domeeneja käyttämällä hyödyksi todennäköisyyslaskennallisia profiili-HMM:iä (profile hidden Markov models, Eddy, 1998; Fischer ym., 2015). HMM-profiileina käytettiin GyDB Collection HMM (Llorens ym., 2011). Tämän jälkeen *LTRdigest* ajettiin parametreilla -seqnamelen 50 -hmms, eli määritettiin sekvenssien nimien pituudeksi maksimissaan 50 merkkiä ja hakuun sisällytettiin HMM-profiilit.

2.5.4. Superperheiden etsintä

Seuraavassa vaiheessa käytettiin VSEARCH-ohjelmaa (Rognes, Flouri, Nichols, Quince, & Mahé, 2016). Haku tehtiin parametreillä --usearch_global -db --id 0.8 --blast6out --strand both --query_cov 0.8 --target_cov 0.8 --top_hits_only --threads 40. Tietokantana (-db) käytettiin PIER-tietokantaa (Wegrzyn ym., 2013). Tässä vaiheessa käytettiin parametria usearch_global, koska haluttiin tehdä globaali parittainen linjaus tietokannan ja aineistojen välille (Vsearch manual, 2019). Linjauksella tarkoitetaan sitä, että sekvenssit järjestetään niin, että löydetään samankaltaisia alueita (Haque, Aravind, & Reddy, 2009). Linjaus tehtiin, jotta saatiin sekvenssit annotoitua. Parametreissä esiintyy 80–80–80 -sääntö, jonka perusteella transposoituvat elementit voidaan jakaa perheisiin. Säännön mukaan sekvenssin pituuksien täytyy olla vähintään 80 emäsparia, niiden sekvenssien identiteetin tulee olla vähintään 80 % vähintään 80 % pituudesta, jotta transposoituvat elementit voidaan luokitella kuuluviksi samaan perheeseen (Wicker ym., 2007). –top_hits_only merkitsee tulokset vain korkeimman

identiteettiprosentin mukaan. --threads 40 sallii ohjelman ajon 40 prosessoriytimellä. Parametri --strand both valittiin, koska haluttiin tarkistaa molemmat juosteet ja --blast6out määrittää muodon, jossa tulostiedosto saadaan ohjelmasta.

2.5.5. *LTRdigestin* tulostiedoston indeksointi

Kaikkia potentiaalisia LTR-retrotransposonisekvenssejä ei pystytty annotoimaan PIER-tietokannan (Wegrzyn ym., 2013) avulla. *LTRdigestin* tulostiedostosta luotiin indeksitiedosto Samtoolsin (Danecek ym., 2021) faidx-ohjelman avulla.

2.5.6. Annotoimattomien sekvenssien nimien etsintä

R:n (R Core Team, 2021) käytössä hyödynnettiin työssä RStudio-ohjelmaa (RStudio Team, 2021). R:ään luettiin Samtoolsilla (Danecek ym., 2021) luotu indeksitiedosto sekä VSEARCH:n (Rognes ym., 2016) tulostiedosto. R:ssä tarkistettiin, mitkä sekvenssit on pystytty annotoimaan PIER:in (Wegrzyn ym., 2013) avulla vastaavasti kuten aiemmin on tehty (https://github.com/mcstitzer/w22_te_annotation/blob/master/ltr/families/get_nofams.R, 2018). Tulokset kirjoitettiin tiedostoon, joka sisältää sellaisten sekvenssien nimet, joita ei pystytty vielä jaottelemaan yhteenkään perheeseen. Tämän jälkeen R:llä tehdyt tulostiedostot tuli muuttaa vielä Windowsin rivien lopetuksesta UNIX:n rivien lopetusmuotoon.

2.5.7. FASTA-tiedoston luonti annotoimattomista sekvensseistä

Samtoolsilla (Danecek ym., 2021) luotiin FASTA-tiedosto, joka sisältää sekvenssit, joita ei vielä pystytty annotoimaan PIER:in (Wegrzyn ym., 2013) avulla.

2.5.8. Annotoimattomien sekvenssien kaikki vastaan kaikki -linjaus

VSEARCH:lla (Rognes ym., 2016) tehtiin allpairs_global -haku, joka tekee globaalin kaikki vastaan kaikki -parittaislinjauksen (Vsearch manual, 2019). Parametreina olivat -allpairs_global, -blast6out, -id 0.8, -query_cov 0.8, -target_cov 0.8 ja --threads 40. Jälleen blast6out määritti tulostiedoston muodon ja linjaus tehtiin 80–80–80 -säännön mukaisesti, jotta saatiin loput aineistosta jaettua omiin perheklustereihin (Wicker ym., 2007).

2.5.9. Sekvenssien klusterointi perheisiin

Aiemmin annotoimattomien sekvenssien klusterointi perheisiin suoritettiin SiLiX:iä (Miele, Penel, & Duret, 2011) käyttäen. Tällöin parametreina olivat -i 0.8, -r 0.8, –net sekä -f LTR, jolla saatiin perheiden nimet alkamaan kirjaimilla LTR.

2.5.10. Superperheiden ja perheiden nimeäminen

Tämän jälkeen luotiin tiedostot, jotka sisältävät sekä jo aiemmin PIER:in (Wegrzyn ym., 2013) avulla saadut annotaatiot, sekä pystyttiin vielä luokittelemaan puuttuvat elementit tarkemmin superperheisiin R-koodin (https://github.com/mcstitzer/w22_te_annotation/blob/master/combine_all_LTRs.R, 2018) avulla. Superperheisiin luokittelu perustui siihen, mitkä proteiinidomeenit tunnistettiin aiemmin *LTRdigestillä*. Löydetyt LTR-retrotransposonit nimettiin tarkemmin käyttäen apuna tiedostoa, jossa oli maissin transposoituvien elementtien perheiden nimet (http://ftp.gramene.org/CURRENT_RELEASE/gff3/zea_mays/repeat_annotation/, 2018).

2.5.11. Transkriptomi-FASTA

Männyn transkriptomi-FASTA (Ojeda ym., 2019) käsiteltiin toisella tavalla kuin mäntyjen genomiaineistot. Alussa männyn transkriptomi-FASTA jaettiin kolmeen osaan, koska teknisistä syistä aineistoa ei pystytty käsittelemään kokonaisena. Jokainen kolmesta osasta käsiteltiin Suffixeratorissa (GenomeTools, 2014), jossa parametrina oli -memlimit 75GB, ja *LTRharvestissa* (Ellinghaus ym., 2008). Myös *LTRdigest*-vaiheeseen vaadittava sort-välivaihe (Steinbiss, Willhoeft, Gremme, & Kurtz, 2012) suoritettiin kaikille kolmelle transkriptomi-FASTA:lle. *LTRdigest*-vaiheessa männyn transkriptomi-FASTA:lle tehtiin HMM-haku, kuten muille aineistoille, mutta lisäksi tehtiin myös tRNA-haku, jota varten tRNA-geenitietoja sisältävästä GtRNAdb2-tietokannasta (Chan & Lowe, 2016) ladattiin oletushakusanoilla kaikki tulokset. Täten saatiin ladattua 451 327 tRNA:ta. Hakutulokset sisälsivät kuitenkin myös tyhjiä sekvenssejä, jotka poistettiin. Yhteensä poistettiin 9 489 tyhjää hakutulosta, eli lopullisessa tietokannassa käytössä oli jäljellä 441 838 tRNA-sekvenssiä. *LTRdigest*-ajo ajettiin parametreilla -seqnamelen 150 -hmms -pptlen 10 30 -pbsoffset 0 3 -trnas. -seqnamelen oli tässä ajossa pidempi, koska sekvenssien nimet olivat transkriptomi-FASTA:ssa pidempiä kuin muissa aineistoissa, -pptlen määrittää polypuriinijuosteen vaihteluvälin pituuden ja -pbsoffset

määrittää kuinka kaukana alukkeen kiinnittymiskohta saa olla LTR:n reunasta. Polypuriinijuosteen vaihteluvälin sekä alukkeen kiinnittymiskohdan etäisyyden arvot perustuvat esitettyyn esimerkkiin ohjelman käytöstä (Steinbiss ym., 2012). Myös VSEARCH-haku usearch_global -parametrilla suoritettiin männyn transkriptomi-FASTA:lle kolmessa osassa. Kolmelle transkriptomi-FASTA:lle luotiin indeksit Samtoolsin (Danecek ym., 2021) avulla. Vielä annotoimattomien sekvenssien nimien selvitys suoritettiin R:ssä aiemmin mainitulla tavalla ainoastaan transkriptomi-FASTA:n kolmannelle osalle, koska se oli ainoa osa tiedostosta, josta löytyi VSEARCH:n haussa PIER:in avulla yhtäkään nimettyä sekvenssiä. Myös seuraava Samtools-vaihe, jossa luotiin FASTA-tiedosto vielä annotoimattomista sekvensseistä, tehtiin vain transkriptomi-FASTA:n kolmannelle osalle. Männyn transkriptomi-FASTA:n kolmannellekaan osalle ei tehty myöhempiä vaiheita, koska ne eivät onnistuneet johtuen tiedoston jakamisesta useaan osaan.

2.5.12. Löydettyjen LTR-retrotransposoneiden analysointi

Käytetty PIER-tietokanta sisälsi muun muassa transposoituvan elementin nimen ja superperheen. Kun tehtiin VSEARCH-haku, kuten kohdassa 2.5.4. tehtiin, tuloksena saatiin lista sekvensseistä, joista löytyi transposoituva elementti sekä kyseisille elementeille nimet PIER:in mukaan nimettynä. Tämän jälkeen tarkistettiin mihin superperheeseen mäntyjen genomiaineistosta löytyneet transposoituvat elementit kuuluvat PIER-tietokannan avulla. Saadut tulostiedostot ladattiin R:ään, jossa tarkistettiin mitä superperheitä mäntyjen genomeista sekä transkriptomi-FASTA -tiedostoista on löytynyt. Tässä vaiheessa suoritettiin manuaalinen filtteröinti *Gypsy*- tai *Copia*-superperheisiin kuuluville elementeille sekä LTR-retrotransposoneille, joiden luokittelu ei onnistunut. Kuvat superperheisiin luokittelusta tehtiin R:n pie-funktiota käyttäen.

2.6. Solukkoekspressioanalyysi

Solukkoekspressiota tutkittaessa käytettiin aineistoa aiemmasta julkaisusta (Cervantes, Vuosku, & Pyhäjärvi, 2021), jossa tutkittiin eri solukoissa tapahtunutta ylössäätelyä. Artikkelissa on tutkittu, onko ylössäätelyä tapahtunut alkiossa, silmussa, megagametofyytissä, neulasessa tai nilassa. Tässä työssä on käytössä sama ekspressioaineisto ja tavoitteena oli tutkia, mitkä männyn transkriptomi-FASTA -tiedostoista löydetyt potentiaaliset LTR-retrotransposonit ovat ylössäädeltyjä aiempien tulosten perusteella.

Solukkoekspressioanalyysiä varten männyn transkriptomi-FASTA:sta luotiin lista kaikkien sekvenssien nimistä, joista oli löytynyt aiemmissa välivaiheissa potentiaalisia LTR-retrotransposoneita. Tämän jälkeen sekvenssien nimiä lyhennettiin ja alkuperäisestä ekspressioaineistosta (Cervantes ym., 2021) valittiin transkriptit, jotka tämän työn perusteella ovat LTR-retrotransposoneita sekä lisättiin alkuperäiset otsikot. R:ään luettiin solukkoekspressiotiedot transkripteille, jotka oli tunnistettu LTR-retrotransposoneiksi sekä alkuperäinen solukkoekspressioaineisto (Cervantes ym., 2021). LTR-retrotransposonilista siis sisältyy myös kaikkien transkriptien aineistoon.

2.6.1. Ylössäätelyn frekvenssi solukkoa kohti

Solukkoekspressioaineistossa (Cervantes ym., 2021) oli merkitty solukoittain ja transkripteittäin, onko kyseinen transkripti kyseisessä solukossa ylössäädelty. Jos transkripti oli solukossa ylössäädelty, taulukossa oli arvo 1. Jos transkripti ei ollut solukossa ylössäädelty, taulukossa oli arvo 0. Ylössäätelyn frekvenssi solukkoa kohti laskettiin jokaista solukkoa kohti erikseen. Lasku suoritettiin R:ssä siten, että jaettiin solukkoa kohti ylössäädeltyjen transkriptien määrä yhteensä kaikkien solukon transkriptien lukumäärällä. Lasku suoritettiin koko alkuperäiselle aineistolle (Cervantes ym., 2021) sekä pienemmälle aineistolle, joka sisälsi vain LTR-retrotransposonit. Ylössäätelyn kuvaaminen R:llä suoritettiin ggplot2-paketilla (Wickham, 2016).

2.6.2. Ylössäätelyn tilastollinen testaus

Aineistoille suoritettiin z-testi solukoittain käyttämällä BSDA-paketin (Arnholt & Evans, 2017) z.test-funktiota, laskettiin kaikkien transkriptien ylössäätelyn suhteellisen osuuden keskiarvot solukoittain, LTR-retrotransposoneiden ylössäätelyn suhteellisten osuuksien keskiarvot solukoittain, kaikkien transkriptien ylössäätelyn suhteellisten osuuksien 95 %:n luottamusvälit, LTR-retrotransposoneiden ylössäätelyn suhteellisten osuuksien 95 %:n luottamusvälit, Cohenin d ja Cohenin d:n 95 %:n luottamusvälit sekä pääteltiin efektikoko Cohenin d:n perusteella. Z-testi valittiin, koska aineisto oli binääristä. Welchin kahden näytteen t-testi laskettiin funktiolla t.test neulasen ja nilan ylössäädellyille LTR-retrotransposoneille, jotta saatiin selville, oliko näiden ryhmien keskiarvojen välillä tilastollisesti merkittävä ero.

Cohenin d laskettiin käyttämällä R:n effectsize-paketin (Ben-Shachar, Lüdecke, & Makowski, 2020) funktiota cohens_d. Tulosten tulkinta suoritettiin paketin funktiolla

interpret_d. Paketissa $d < 0.2$ tarkoittaa hyvin pientä efektikokoa, $0.2 \leq d < 0.5$ vastaa pientä efektikokoa, $0.5 \leq d < 0.8$ vastaa keskisuurta efektikokoa ja $d \geq 0.8$ vastaa suurta efektikokoa.

2.6.3. Transkriptien ekspressio solukkoa kohti

Ekspressiotaso solukkoa kohti otettiin alkuperäisestä solukkoekspressioaineistosta (Cervantes ym., 2021) sekä alkuperäiselle aineistolle että pienennetylle aineistolle, joka sisälsi löydetyt LTR-retrotransposonit. Molemmat aineistot luettiin R:ään ja jokaista solukkoa kohti otettiin luonnollinen logaritmi. Kuvaajat tehtiin R:n hist-funktiolla, jossa käytettiin parametria freq = FALSE, jotta saatiin tehtyä tiheyskuvaajat. Näin saatiin luotua kuvaajat, jotka ilmaisevat ekspressiotasoja jokaista tutkittua solukkoa kohti.

2.6.4. Transkriptien ekspressiotason tilastollinen testaus

Solukoiden transkriptien ekspressiotasoille suoritettiin samat tilastolliset testit kuin ylössäätelyn suhteellisen osuudelle, paitsi että z-testin sijasta käytettiin t-testiä. T-testi valittiin, koska aineisto oli jatkuvaa. T-testi suoritettiin käyttämällä R:n t.test-funktiota. Keskiarvot, luottamusvälit, Cohenin d, Cohenin d:n luottamusvälit sekä efektikoko määritetiin samoilla menetelmillä kuin aiemmin. ANOVA suoritettiin aov-funktiolla. ANOVA-tulokset tulkittiin summary-funktiolla. Tarkempi tieto solukoiden välisten keskiarvojen eroista selvitettiin Tukey Honest Significant Differences -testillä funktiolla TukeyHSD.

2.7. N50 ja N90

Kontigien N50:llä voidaan ilmaista, kuinka kattava genomin koostaminen (assembly) on ollut. Mitä suurempi N50 on, sitä kokonaisempi koostaminen on ollut. Käytännössä N50-arvo tarkoittaa pituutta, jota pidemmät kontigit kattavat 50 % kooston emäksistä (Pevsner, 2015, s. 395). Aineistoille laskettiin N50- ja N90-arvot käyttämällä CNEr-paketin (Tan, Polychronopoulos, & Lenhard, 2019) funktioita N50 ja N90.

3. Tulokset

P. sylvestrisin transkriptomi-FASTA:ssa oli 787 820 kontigia, *P. sylvestrisin* genomiaineistossa 3 954 508, *P. lambertianan* genomiaineistossa 16 610 ja *P. taedan* genomiaineistossa 1 755 249 kontigia. N50-arvo oli *P. sylvestrisin* genomiaineistolle 12 543, *P. lambertianan*

genomiaineistolle 2 917 154 ja *P. taedan* genomiaineistolle 110 557. N90-arvo oli männyn genomiaineistolle 1 569, *P. lambertianan* genomiaineistolle 712 812 ja *P. taedan* genomiaineistolle 7 360. N50-arvoista voidaan päätellä, että *P. lambertianan* genomiaineisto oli kokonaisin ja *P. sylvestrisin* hajanaisin. N50- tai N90-arvoja ei pystytty määrittämään *P. sylvestrisin* transkriptomi-FASTA:lle.

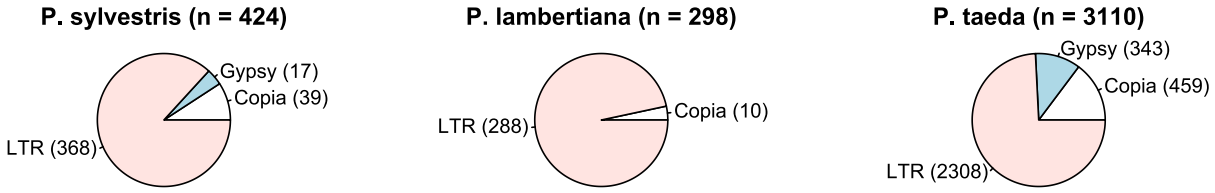
Männyn ekspressioaineiston kaikista kolmesta osasta löytyi yhteensä 13 025 potentiaalista LTR-retrotransposonia, *P. sylvestrisistä* 319 702, *P. lambertianasta* 915 449 ja *P. taedasta* 705 552. PIER-tietokannan avulla saatiin annotoitua *P. sylvestrisin* transkriptomi-FASTA:sta 3, *P. sylvestrisistä* 1 132, *P. lambertianasta* 323 ja *P. taedasta* 7 381 LTR-retrotransposonia. Aiemmin kolmeen jaetusta transkriptomi-FASTA:sta vain yhdestä osasta löytyi edellä mainitut kolme LTR-retrotransposonia. Annotoituja LTR-retrotransposoneita tarkasteltaessa huomattiin, että PIER:istä oli löytynyt elementtejä, joita aineistossa ei pitäisi olla, koska ne eivät kuulu LTR-ryhmään. Kun nämä vääräksi tiedetyt tulokset poistettiin manuaalisesti, jäi filtteröinnin jälkeen *P. sylvestrisin* transkriptomi-FASTA:an 1, *P. sylvestrisin* genomiaineistoon 424, *P. lambertianan* genomiaineistoon 298 ja *P. taedan* genomiaineistoon 3 110 annotaatiota (Taulukko 1).

Taulukko 1. Löydetyt potentiaaliset LTR-retrotransposonit (LTR-RT) *LTRdigestin*, PIER:in avulla annotoinnin sekä manuaalisen filtteröinnin jälkeen.

	***P. sylvestris* (transkriptomi)**	***P. sylvestris* (genominen)**	***P. lambertiana***	***P. taeda***
Potentiaaliset LTR-RT:t (*LTRdigest*)	13025	319702	915449	705552
Nimetyt LTR-RT:t (PIER)	3	1132	323	7381
Nimetyt LTR-RT:t (manuaalinen filtteröinti)	1	424	298	3110

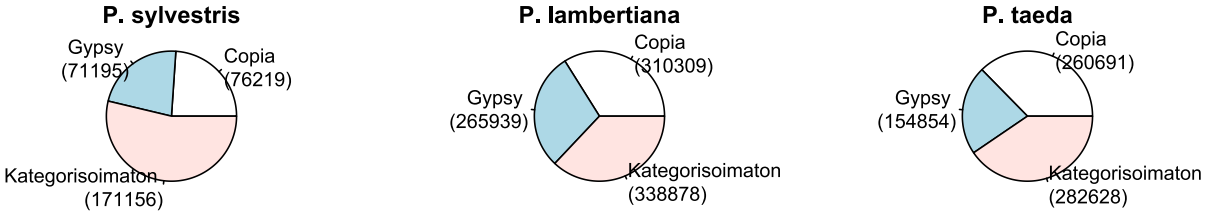
P. sylvestrisin genomiaineistosta onnistui nimetä PIER-tietokannan avulla 17 elementtiä kuuluvaksi *Gypsy*-superperheeseen ja 39 *Copia*-superperheeseen. Lisäksi pystyttiin tunnistamaan 368 elementtiä, jotka olivat LTR-retrotransposoneita, mutta PIER-tietokannassa niitä ei oltu pystytty tarkemmin määrittämään superperheisiin (Wegrzyn ym., 2014). *P. lambertianan* genomiaineistosta löydettiin PIER:in avulla 10 elementtiä, jotka kuuluvat *Copia*-ryhmään ja 288 elementtiä, jotka kuuluvat yleisesti LTR-retrotransposoneihin ilman tarkempaa luokittelua superperheisiin. *P. taedan* genomiaineistosta onnistui nimetä PIER:in avulla 343 elementtiä *Gypsy*-superperheeseen, 459 *Copia*-superperheeseen ja 2 308 elementtiä

yleisesti kuuluvaksi yleisesti LTR-retrotransposoneihin, joista PIER-tietokantaan ei oltu saatu tarkempaa tietoa superperheestä (Wegrzyn ym., 2014, Kuva 3).



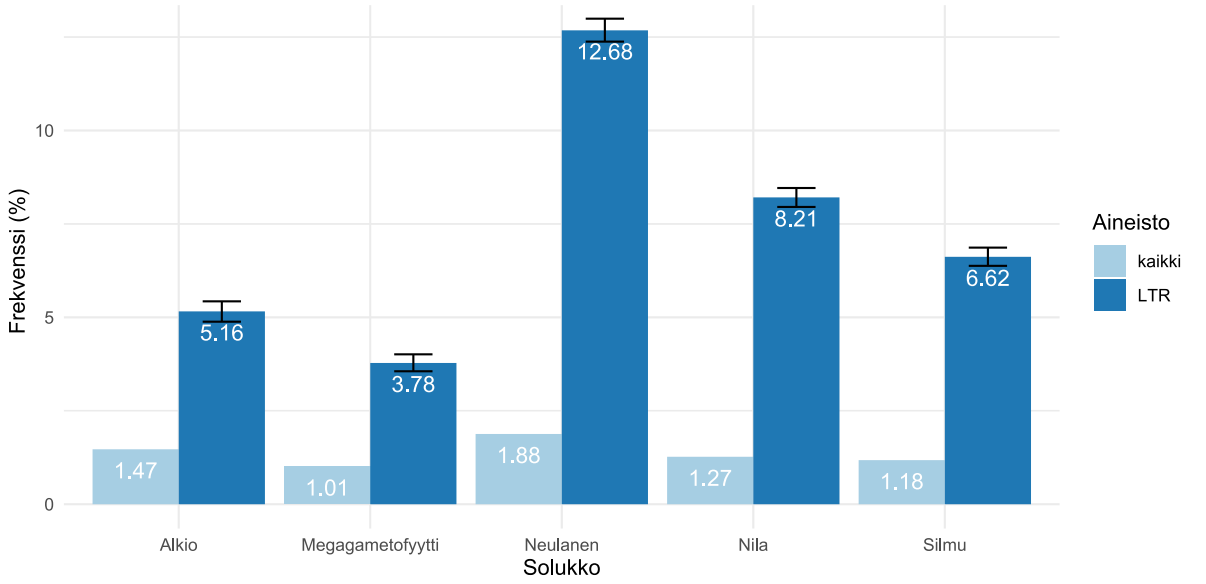
Kuva 3. PIER:in avulla nimetyt superperheet *P. sylvestrisin*, *P. lambertianan* ja *P. taedan* genomiaineistoista manuaalisen filtteröinnin jälkeen.

Suurinta osaa transposoituvista elementeistä ei pystytty annotoimaan PIER:in avulla, vaan luokittelu superperheisiin suoritettiin klusterointimenetelmiä käyttämällä. *P. sylvestrisin* genomiaineistosta saatiin klusteroitua yhteensä 156 190 perheeseen 318 570 transposoituvaa elementtiä. Elementeistä ei pystytty luokittelemaan n. 53,73 %. 147 414 elementtiä saatiin siis luokiteltua tarkemmin superperheisiin. *Gypsy*-superperheeseen saatiin luokiteltua 71 195 elementtiä (n. 22,35 % kaikista) ja *Copia*-superperheeseen 76 219 elementtiä (n. 23,93 % kaikista). *P. lambertianan* genomiaineistosta saatiin klusteroitua 915 126 elementtiä 344 438 eri perheeseen. Klusteroiduista elementeistä 37,03 % jäi luokittelemattomiksi. 576 248 elementtiä saatiin siis luokiteltua superperheisiin. Kaikista klusteroiduista elementeistä 33,91 % eli 310 309 elementtiä luokiteltiin kuuluvaksi *Copia*-superperheeseen. 29,06 % eli 265 939 elementtiä luokiteltiin kuuluvaksi *Gypsy*-superperheeseen. *P. taedan* genomiaineistosta saatiin klusteroitua yhteensä 698 173 elementtiä 241 623 eri perheeseen. Klusteroiduista elementeistä 40,48 % jäi luokittelematta tarkemmin superperheisiin. 415 545 elementtiä saatiin siis luokiteltua superperheisiin. Kaikista klusteroiduista elementeistä 37,34 % kuului *Copia*-superperheeseen (260 691 elementtiä) ja 22,18 % eli 154 854 elementtiä luokiteltiin *Gypsy*-superperheeseen (Kuva 4). Tarkempaa luokittelua perheisiin ja superperheisiin ei suoritettu männyn transkriptomi-FASTA:n LTR-retrotransposoneille, koska jo GenomeTools-vaiheessa aineisto täytyi jakaa osiin ohjelmien ajamisen mahdollistamiseksi. Tämän seurauksena LTR-retrotransposoneiden luokittelu on kaikissa tiedostoissa alkanut alusta, ja myöhemmässä vaiheessa olisi ollut vaikeaa yhdistää löytyneitä elementtejä samaan tiedostoon. Koska myöhemmät vaiheet perustuvat 80–80–80 -säännön (Wicker ym., 2007) mukaiseen linjaukseen, ositettujen transkriptomi-FASTA -tiedostojen linjaustulokset eivät olisi olleet totuudenmukaisia. Tämä johtuu siitä, että linjausta ei olisi voinut tehdä kaikkien aineistossa esiintyvien LTR-retrotransposonisekvenssien kesken.



Kuva 4. Kaikki *P. sylvestrisin*, *P. lambertianan* ja *P. taedan* genomiaineistoista löytyneet LTR-retrotransposonisuperperheet.

Alkuperäisessä ekspressioaineistossa (Cervantes ym., 2021) oli 715 398 transkriptiä, joista 1,65 % eli 11 806 transkriptiä oli aiempien vaiheiden perusteella LTR-retrotransposoneita. Tässä työssä *P. sylvestrisin* genomiaineistosta löytyneistä 318 570 transposoituvasta elementistä 3,71 % löytyi alkuperäisestä ekspressioaineistosta (Cervantes ym., 2021). Ylössäätely solukkoa kohti määritettiin laskemalla ylössäädeltyjen transkriptien suhteellinen määrä verrattuna kaikkien transkriptien määrään solukkoa kohti. Kuvassa 5 on esitetty ylössäätelyn suhteellinen määrä solukkoa kohti kaikissa transkripteissä (vaaleansininen) ja LTR-retrotransposoneissa, jotka löytyivät solukkoekspressioaineistosta (tummansininen). Alinta ylössäätely oli megagametofyytissä sekä kaikissa transkripteissä (1,01 %), että LTR-retrotransposoneissa (3,78 %). Eniten ylössäätelyä oli neulasessa sekä kaikissa transkripteissä (1,88 %) että LTR-retrotransposoneissa (12,68 %). Alkion kaikissa transkripteissä ylössäätelyä oli 1,47 % ja LTR-retrotransposoneissa 5,16 %. Nilassa ylössäätelyä oli kaikissa transkripteissä 1,27 % ja LTR-retrotransposoneissa 8,21 %. Silmun kaikissa transkripteissä ylössäätelyä oli 1,18 % ja LTR-retrotransposoneissa 6,62 % (Kuva 5). Ylössäätelytulokset solukkoa kohti testattiin tilastollisesti käyttämällä z-testiä. Tulokset olivat tilastollisesti merkittäviä eli tilastollisen testaamisen perusteella kaikkien transkriptien ja LTR-retrotransposoneiden ylössäätelyn frekvenssin keskiarvo eroavat tilastollisesti merkittävästi (Taulukko 2).



Kuva 5. Ylössäädeltyjen transkriptien frekvenssi verrattuna kaikkien transkriptien määrään solukkoa kohti sekä kaikissa transkripteissä (vaaleansininen, Cervantes ym., 2021) että aineistosta löytyneissä LTR-retrotransposoneissa (tummansininen). LTR-retrotransposoneiden ylössäädeltyjen transkriptien frekvensseille on esitetty myös 95 % luottamusväli. Kuvan LTR-retrotransposoniaineisto sisältyy myös kaikkien transkriptien aineistoon.

Taulukko 2. Ylössäätelytuloksista tehdyn z-testin sekä Cohenin d-testin tulokset solukkoa kohti.

Solukko	p-arvo	Ka* (kaikki transkriptit)	Ka (LTR)	Ka:n 95 % CI** (LTR)	Cohenin d	Cohenin d 95 % CI	Efekti-koko
Alkio	<2.2e-16	0.015	0.052	[0.049, 0.054]	0.18	[0.16, 0.20]	hyvin pieni
Mega-gameto-fyytti	<2.2e-16	0.010	0.038	[0.036, 0.040]	0.17	[0.15, 0.19]	hyvin pieni
Neulanen	<2.2e-16	0.019	0.13	[0.12, 0.13]	0.63	[0.55, 0.72]	keski-suuri
Nila	<2.2e-16	0.013	0.082	[0.080, 0.085]	0.48	[0.35, 0.51]	pieni
Silmu	<2.2e-16	0.012	0.066	[0.064, 0.069]	0.38	[0.36, 0.40]	pieni

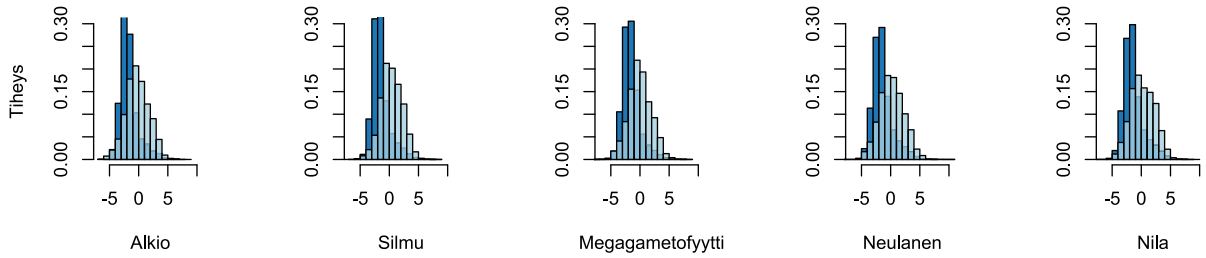
*keskiarvo

**luottamusväli

Welchin kahden näytteen t-testissä neulasen ja nilan ylössäädellyille LTR-retrotransposoneille saatiin p-arvoksi <2,2e-16 ja 95 %:n luottamusväliksi [0.035, 0.055].

Tulokset tarkoittavat, että ryhmien keskiarvot erosivat toisistaan tilastollisesti merkittävästi eli ryhmien välillä oli eroa.

Ekspressiotaso eri solukoissa vaihteli (Kuva 6). Kaikkien transkriptien ja LTR-retrotransposoneiden keskiarvon tilastollista merkittävyyttä testattiin t-testillä. Tulos oli tilastollisesti merkittävä. LTR-retrotransposoneiden ekspressio oli siis koko aineistoa keskiarvoisesti suurempaa (Taulukko 3).



Kuva 6. Ekspressiotaso kaikissa transkripteissä (tummansininen, Cervantes ym., 2021) ja LTR-retrotransposoneissa (vaaleansininen) eri solukoissa.

Taulukko 3. T-testin ja Cohenin d:n tulokset eri solukoiden ekspressiotasoon liittyen.

Solukko	p-arvo	Ka* (kaikki transkriptit)	Ka (LTR)	Ka:n 95 % CI** (LTR)	Cohenin d	Cohenin d 95 % CI	Efekti-koko
Alkio	<2.2e-16	1.21	4.91	[4.41, 5.41]	0.14	[0.12, 0.16]	hyvin pieni
Mega-gameto-fyytti	<2.2e-16	1.24	5.81	[4.98, 6.63]	0.16	[0.14, 0.18]	hyvin pieni
Neulanen	<2.2e-16	1.88	8.34	[7.59, 9.08]	0.05	[0.03, 0.07]	hyvin pieni
Nila	<2.2e-16	1.38	7.71	[6.85, 8.58]	0.18	[0.16, 0.20]	hyvin pieni
Silmu	<2.2e-16	1.27	6.99	[6.29, 7.69]	0.21	[0.19, 0.24]	pieni

*keskiarvo

**luottamusväli

Solukoiden välisen keskiarvojen eron tulkitsemiseksi suoritettiin ANOVA-testi, jonka F-arvo oli 13,76 ja p-arvo 3,26e-11. Koska korjattu p-arvo on alle 0,05, solukoiden transkriptien ekspression keskiarvojen välillä oli tilastollisesti merkittävä ero. Tukey Honest Significance Differences -testin mukaan tilastollisesti merkittävät sovitetut p-arvot olivat solukkoparien alkio-silmu, neulanen-alkio, nila-alkio, neulanen-megagametofyytti sekä nila-megagametofyytti välillä (Taulukko 4).

Taulukko 4. ANOVA:n pohjalta tehty Tukey Honest Significance Differences -vertailu solukoiden välillä.

Solukkopari	Keskiarvojen erotus	95 % luottamusväli	Korjattu p-arvo
Alkio-silmu	-2.08	[-3.53, -0.63]	0.00
Megagametofyytti-silmu	-1.18	[-2.63, 0.27]	0.17
Neulanen-silmu	1.35	[-0.10, 2.80]	0.08
Nila-silmu	0.72	[-0.73, 2.17]	0.65
Megagametofyytti-alkio	0.90	[-0.55, 2.35]	0.44
Neulanen-alkio	3.43	[1.98, 4.88]	0.00
Nila-alkio	2.80	[1.35, 4.26]	0.00
Neulanen-megagametofyytti	2.53	[1.08, 3.98]	0.00
Nila-megagametofyytti	1.90	[0.45, 3.35]	0.00
Nila-neulanen	-0.63	[-2.08, 0.82]	0.76

4. Pohdinta

Työn tarkoituksena oli tutkia, voidaanko aiemmin maissilla käytettyjä LTR-retrotransposoneiden etsimis- ja annotaatiomenetelmiä käyttää havupuille. Tutkimuksessa siis etsittiin ja annotoitiin *P. sylvestrisin* transkriptomiaineistosta sekä *P. sylvestrisin*, *P. lambertianan*, *P. taedan* genomisesta aineistosta LTR-retrotransposoneita. Lisäksi tarkoituksena oli tutkia, missä männyn solukoissa LTR-retrotransposonit ovat aktiivisia. Elementtien löytäminen ja annotaatio onnistui valittujen menetelmien avulla. LTR-retrotransposoneiden ekspressio oli koko aineistoa keskiarvoisesti suurempaa, kun tarkasteltiin ekspressiota eri solukoissa. Lisäksi ylössäädeltyjen transkriptien määrä solukoittain erosi tilastollisesti merkittävästi sen perusteella, oliko kyseessä kaikki transkriptit vai LTR-retrotransposonit. LTR-retrotransposoneiden ekspression määrässä eri solukoiden välillä esiintyi tilastollisesti merkittäviä eroja. Tässä työssä *P. sylvestrisin* genomiaineistosta

löytyneistä transposoituvista elementeistä noin neljä prosenttia löytyi aiemmin julkaistusta ekspressioaineistosta (Cervantes ym., 2021), mikä viittaa siihen, että kovinkaan iso osa *P. sylvestrisin* genomista löytyneistä LTR-retrotransposoneista ei ollut aktiivisia.

Tutkimuksessa oli rajoituksia. Yhtenä rajoitteena oli, että männyn genomi- ja transkriptomiaineisto olivat peräisin eri yksilöistä, joten aineistot eivät olleet suoraan vertailukelpoisia. Ekspressioaineistossa jokainen oma transkripti ei välttämättä ollut peräisin eri LTR-retrotransposoneista, vaan koska aineisto oli peräisin usealta yksilöltä, voi sama elementti olla aineistossa useaan kertaan. Tämä johtuu siitä, että elementti on voinut toisessa yksilössä mutatoitua hieman eri tavalla. LTR-retrotransposoneiden transkripteissä on päissä polyadenylaatiosignaali (Lanciano & Cristofari, 2020) ja ekspressioaineistossa mRNA kerättiin hyödyntäen poly-A -häntiä (Ojeda ym., 2019). Tämän perusteella olisi mahdollista, että ekspressioaineisto todella sisälsi LTR-retrotransposoneita. Tämän työn tuloksista ei siis voi päätellä, sijaitsevatko löydetyt LTR-retrotransposonit geenien sisällä vai niiden ulkopuolella.

Joitakin elementtejä myös annotoitiin virheellisesti PIER:in avulla kuuluvaksi sellaisiin superperheisiin, jotka eivät ole LTR-retrotransposoneita. Kun työssä klusteroitiin nimeämättömät perheet, mukana ei ollut elementtejä, jotka PIER:n perusteella annotoitiin virheellisesti. Työtä voisi parantaa ottamalla selville, miksi tietokannan käyttö johti osittaisiin vääriin tuloksiin.

Työssä ei otettu huomioon sitä, että LTR-retrotransposonit ovat usein sisäkkäin genomissa, varsinkin jos elementtejä on suuri määrä (Jedlicka, Lexa, Vanat, Hobza, & Kejnovsky, 2019). Tulevaisuudessa sisäkkäisyys tulisi ottaa huomioon, jolloin löytyisi mahdollisesti eri määrä LTR-retrotransposoneita kuin tämän työn menetelmillä. Tämä työ oli vasta pohjustusta sille, miten LTR-retrotransposoneiden aktiivisuutta voitaisiin tutkia havupuissa. LTR-retrotransposonit esiintyvät sisäkkäin tietynlaisessa ympäristössä (Jedlicka ym., 2019) eli jos tutkimus olisi tehty ottamalla huomioon sisäkkäisyys, olisi voitu myös tutkia millaisilla alueilla sisäkkäisyyttä esiintyy. Täten olisi voitu saada selville alueita, joissa valinta ei ehkä toimi samalla tavalla kuin muualla. Jos jossain alueella olisi ollut paljon sisäkkäisyyttä, se voisi osoittaa, että kyseisellä alueella ei ole samanlaista valintapainetta poistaa LTR-retrotransposoneita kuin sellaisilla alueilla, joissa sisäkkäisyyttä esiintyy vähän. Olisi myös mielenkiintoista saada selville, oliko esimerkiksi jossakin solukossa enemmän sisäkkäisyyttä LTR-retrotransposoneissa kuin muualla. Tämä voisi edelleen avata LTR-retrotransposoneiden biologista merkitystä havupuissa. Lisäksi rajoituksena työssä oli, että työn tuloksena ei saatu transposoituvien elementtien pituuksia eli ei saatu suoraan selville kuinka ison osan mäntyjen genomeista LTR-retrotransposonit kattavat.

Tällä hetkellä ei ole vielä julkaistu männyn genomia. Kun julkaisu tehdään, saadaan tietää missä osissa genomia LTR-retrotransposonit sijaitsevat. *P. taedan* ja *P. lambertianan* genomitutkimuksissa on havaittu, että TE:itä pyritään jonkin verran poistamaan geenien läheisyydestä. Tämä on voitu päätellä siitä, että TE:iden määrä lisääntyi mitä kauemmaksi geeneistä siirryttiin. Lisäksi havaittiin, että suurin osa TE-insertioista löytyi introneista (Voronova, Rendón-Anaya, Ingvarsson, Kalendar, & Ruņģis, 2020). Kuitenkin liittyessään eksoniin transposoituva elementti voi aiheuttaa insertionaalista mutageneesiä (Lisch, 2013). Jos tiedettäisiin, missä LTR-retrotransposonit sijaitsevat, voitaisiin tutkia minkä geenien lähellä ne sijaitsevat. Tällöin saataisiin selville, voiko LTR-retrotransposoni aiheuttaa jonkin haitallisen mutaation. Transposoituvien elementtien insertoitumisen eksoniin on esimerkiksi todettu lisäävän alttiutta joillekin patogeeneille *Arabidopsisilla* (Stuart ym., 2016). Tämän vuoksi voisi olla kannattavaa tutkia enemmän transposoituvia elementtejä ja niiden vaikutuksia myös havupuiden jalostuksen kannalta.

Saadut tulokset poikkesivat jonkin verran aiemmin tutkitusta. Wegrzyn ym. (2014) löysivät loblollymännystä (*P. taeda* L.) PIER-tietokantaa käyttäen noin 180 000 kokopitkää LTR-retrotransposonia. Löydetyistä elementeistä noin 27 % kuului *Gypsy*-ryhmään ja noin 21 % *Copia*-ryhmään (Wegrzyn ym., 2014). Tässä työssä kaikista käytetyistä genomiaineistoista löytyi prosentuaalisesti enemmän *Copia*-superperheeseen kuuluvia LTR-retrotransposoneita. *Gypsy*-superperheeseen luokiteltiin kuuluvaksi prosentuaalisesti vähemmän LTR-retrotransposoneita *P. taedalla* ja *P. sylvestrisillä*, mutta *P. lambertianalla Gypsy*-ryhmään luokiteltiin kuuluvaksi enemmän LTR-retrotransposoneita. Wegrzyn ym. (2014) homologian perusteella löytämät kokopitkät LTR-retrotransposonit vastasivat noin 4 % koko *P. taedan* genomista. Tässä työssä *P. taedalle* löytyi elementtejä noin neljä kertaa enemmän. *P. lambertianalle* löydettiin noin viisi kertaa enemmän LTR-retrotransposoneita ja *P. sylvestrisille* löydettiin suunnilleen sama määrä kuin Wegrzyn ym. (2014) löysivät *P. taedalle*.

Erilaiset tulokset löytyneiden LTR-retrotransposoneiden määrässä selittynevät sillä, että menetelmät ovat olleet erit, vaikka käytetty tietokanta olikin sama. Wegrzyn ym. (2014) tekivät LTR-retrotransposonihaun homologiaan perustuen ja he käyttivät PIER-tietokantaa kirjastona. Tässä työssä kirjastona on myös käytetty PIER-tietokantaa, mutta LTR-retrotransposoneiden haku on ensin tehty *de novo*. Tällä hetkellä vaikuttaa, että tietokannoista ei löydy paljoakaan tietoa havupuiden LTR-retrotransposoneista, sillä kovinkaan isoa osaa elementeistä ei saatu annotoitua PIER:in avulla. Työn tuloksena siis löydettiin paljon elementtejä, jotka vaikuttavat LTR-retrotransposoneilta, mutta niitä ei löydy tietokannoista.

Olisi siis syytä parantaa joko jo olemassa olevia tietokantoja, tai kehittää uusi tietokanta, jossa olisi paremmin tietoa juuri havupuiden LTR-retrotransposoneista. Työssä olisi myös voinut kokeilla useamman tietokannan käyttämistä ja verrata niistä saatuja tuloksia toisiinsa. Osa tietokannoista on tällä hetkellä maksumuurin takana (RepBase, Bao, Kojima, & Kohany, 2015). RepBasea ylläpidetään aktiivisesti, joten sieltä olisi voinut löytyä enemmän annotaatioita männyn LTR-retrotransposoneista kuin PIER:istä, joka perustuu yhden artikkelin tuloksiin. Esimerkiksi jonkinlainen joukkoistettu tietokanta voisi mahdollisesti toimia, jotta voitaisiin tuoda yhä useampien tutkijoiden löytämät transposoituvat elementit julkiseen tietoon. Tämä tietenkin vaatii lisätyötä myös tutkijoilta, mutta tuloksena voisi olla arvokas tietolähde. Aktiivinen ylläpito voi myös johtaa siihen, että rajat uuden LTR-retrotransposonin ja vanhan elementin muunnoksen välillä vaihtelevat tietokantojen tekijöiden mukaan. Eri tietokantojen välillä voi siis olla suurta vaihtelua siinä, mikä voidaan määrittää LTR-retrotransposoniksi. Jos rajat määritelmään ovat tällä hetkellä todella tiukat, myös se voi selittää, että tästä työstä ei löytynyt paljoakaan suoraan annotoituja elementtejä.

Transposoituvia elementtejä löytävien ohjelmien suorituskykyä on tutkittu riisillä. *LTRharvestia* käytettäessä sensitiivisyys, eli kuinka hyvin *LTRharvest* annotoi LTR-retrotransposonin, oli noin 98 % ja virheellisten löydösten osuus oli noin 49 % (Ou ym., 2019). On siis mahdollista, että tässäkin työssä suurin osa aineiston sisältämistä LTR-retrotransposoneista on löytynyt, mutta voi olla, että vääriä positiivisia on myös löytynyt merkittävä määrä. LTR_retriever on ohjelma, jolla pystytään poistamaan vääriä positiivisia (Ou & Jiang, 2018). Riisillä todettiin, että ajamalla LTR_retriever (Ou & Jiang, 2018) *LTRharvestin* tulostiedostoilla, saatiin sensitiivisyys pysymään suhteellisen samana (noin 94 %), mutta väärien löydösten osuus laski noin 17 %:iin (Ou ym., 2019). Tämän perusteella myös tässä työssä olisi kannattanut vielä ajaa LTR_retriever *LTRharvestin* jälkeen ja kokeilla, paljonko tulokset olisivat muuttuneet. *LTRharvest* vaikutti kuitenkin soveltuvan käytetylle aineistolla hyvin, koska *P. abiesilla* on aiemmin todettu olevan löydetyistä LTR-retrotransposoneista noin 90 % kokopitkiä elementtejä (Nystedt ym., 2013) ja *LTRharvest* puolestaan on suunniteltu löytämään kokopitkiä LTR-retrotransposoneita (Ellinghaus ym., 2008).

Solukoiden välisessä LTR-retrotransposoneiden ekspressiossa esiintyi eroa. Eniten LTR-retrotransposoneiden ylössäätelyä löytyi neulasesta ja nilasta. Aiemmin maissista on todettu, että iso osa solukkospesifisistä transposoituvien elementtien perheistä on havaittu siitepölystä ja endospermistä. Ekspressiota on havaittu myös alkiossa, lehdessä ja juurissa. Havaitusta transposoituvien elementtien ekspressiosta suurin osa perheistä kuului LTR-retrotransposoneihin (Anderson, Stitzer, Zhou ym., 2019b). Tässä työssä neulasessa esiintyvien

LTR-retrotransposoneiden ylössäätelyn suhteellinen määrä oli suurempaa kuin muissa solukoissa. Jos transkriptomiaineisto olisi pystytty ajamaan yhdessä osassa, olisi voitu tutkia onko neulasessa esimerkiksi yksittäinen perhe todella aktiivinen vai johtuuko neulasen poikkeava tulos jostakin muusta. Myös stressin on havaittu johtavan retrotransposoneiden aktiivisuuteen (Grandbastien, 1998). Neulasesta saatua tulosta voisi myös siis selittää mahdollisesti niiden kokema stressi, jonka seurauksena transposoituvat elementit olisivat voineet aktivoitua.

Saatuihin tuloksiin voi vaikuttaa myös se, että *P. sylvestrisin* transkriptomi-FASTA:n siemenaineisto on kerätty kypsistä siemenistä, joissa alkio on jo muodostunut. Tämän kehitysvaiheen siementen LTR-retrotransposoneiden aktiivisuus on voitu jo estää epigenetiikan avulla. Jos siemenet olisivat olleet tuoreempia, niistä olisi ehkä löytynyt LTR-retrotransposoneiden aktiivisuutta. Myös megagametofyytin solukkonäyte voi olla myöhäistä kehitysvaihetta, mistä johtuen LTR-retrotransposoneiden aktiivisuutta ei havaittu niin paljon kuin olisi ehkä ollut mahdollista. Työssä käytettyä aineistoa ei kuitenkaan oltu suunniteltu siihen, että siitä etsittäisiin LTR-retrotransposoneiden aktiivisuutta. Tulevaisuudessa voisi siis olla kannattavaa kerätä näytteitä ekspressioanalyysiin solukoiden eri kehitysvaiheilta ja tutkia niistä LTR-retrotransposoneiden aktiivisuutta. Aikaisempien tutkimuksien perusteella (Nobuta ym., 2007) sillä on merkitystä, onko solukko vanhaa vai uutta transposoituvien elementtien aktiivisuuden kannalta.

Tulosten mukaan LTR-retrotransposoneiden ekspressio oli koko aineistoa keskiarvoisesti suurempaa. On mahdollista, että transkriptien ekspressio olisi määritetty väärin toistuvista elementeistä johtuen. Jos aineistossa esiintyisi kaksi lähes samanlaista geenikopiota, joista toinen olisi korkeammin ekspressoitunut, koostossa molempien geenien luvut linjattaisiin jompaankumpaan kopioon. Usein ekspressiotaso määritetään mittaamalla lukujen normalisoitu määrä. Tässä tilanteessa toisen geenin ekspressiomäärä arvioidaan virheellisesti liian korkeaksi ja toisessa liian matalaksi (Treangen & Salzberg, 2012). Suurempi LTR-retrotransposoneiden ekspressio työssä voi siis olla tulosta myös linjauksessa syntyneistä virheistä.

Transposoituvien elementtien liittyminen geeneihin on voinut johtaa pseudogeenien syntymiseen (Nystedt ym., 2013). RNA-Seq:llä ei kuitenkaan välttämättä havaita pseudogeenien ekspressiota. Mitä suurempaa ekspressiotaso on, sitä todennäköisemmin se havaitaan (Kalyana-Sundaram ym., 2012). Voi siis olla, että tämän työn ekspressioaineistoon ei oltu saatu kaikkia transposoituvia elementtejä mukaan.

Tulevaisuudessa voisi tutkia myös sitä, paljonko yksittäisiä LTR-retrotransposoneita löytyy tuloksista. Täten saataisiin selville, onko jokin yksittäinen perhe

esimerkiksi todella aktiivinen ja ovatko esimerkiksi useimmat perheet aktiivisia. Voitaisiin myös tutkia mahdollisuutta, onko havupuilla vain muutama aktiivinen perhe. Olisi myös mielenkiintoista tutkia, ovatko tietyt perheet aktiivisia tietyissä solukoissa, vai vaikuttaako perheiden aktiivisuus eri solukoissa satunnaiselta. Tällä tavoin voitaisiin löytää havupuuspesifisiä LTR-retrotransposoneita. Tätä varten transkriptomi-FASTA:n käsittely yhdessä osassa pitäisi saada toimimaan, jotta päästäisiin tutkimaan, missä solukoissa mikäkin löydetty LTR-retrotransposoniperhe on ollut aktiivinen. Työtä voisi laajentaa myös tutkimalla eri transposoituvien elementtien insertioaikoja. Näin voitaisiin saada selville, onko männyilläkin LTR-retrotransposoneiden poistuminen genomista vähäistä, kuten *P. abiesilla* on aiemmin havaittu (Nystedt ym., 2013).

5. Yhteenveto

Työn tutkimuskysymyksenä oli, soveltuvatko koppisiemenisille tehdyt transposoituvien elementtien annotaatiomenetelmät havupuille. Hypoteeseina olivat, että useimmat LTR-retrotransposonit eivät ole aktiivisia ja niiden ekspressiotaso eri solukoiden välillä vaihtelee. Vaikuttaa siltä, että koppisiemenisille suunnitellut annotaatiomenetelmät toimivat myös havupuille. Myös hypoteesit vaikuttavat työn perusteella pitävän paikkansa, kun menetelmänä käytettiin samaa menetelmää kuten aiemmin maissille (https://github.com/mcstitzer/w22_te_annotation/, 2018; Springer ym., 2018; Anderson ym., 2019a). LTR-retrotransposonit onnistuttiin löytämään ja annotoimaan kaikista genomiaineistoista. LTR-retrotransposoneiden solukkospesifinen ekspressio saatiin selville ja solukoiden välillä havaittiin eroa. Tulokset tukevat sitä, että maissille aiemmin kehitetyt LTR-retrotransposoneiden etsimismenetelmät voivat olla hyödyllisiä myös havupuille. Saadut tulokset antavat hyvän pohjan sille, millaisia LTR-retrotransposoneita havupuilta löytyy. Koska tiedetään, miten asiaa voi onnistuneesti tutkia, voidaan tulevaisuudessa ottaa esimerkiksi paremmin selville, mitkä perheet ovat aktiivisia havupuissa. Lisäksi on mahdollista tutkia, mitkä perheet ovat missäkin solukoissa aktiivisia. Työ antaa hyvän pohjan havupuissa esiintyvien LTR-retrotransposoneiden määristä ja siitä, mitä superperheitä havupuissa esiintyy sekä missä solukoissa ne ovat aktiivisia. Työ myös mahdollistaa sen, että tulevaisuudessa voidaan ottaa paremmin selville, mikä on havupuiden LTR-retrotransposoneiden biologinen merkitys. Työn antaman pohjan perusteella voitaisiin tulevaisuudessa tutkia havupuiden LTR-retrotransposoneita lisää erilaisin menetelmin, jotta saadaan vielä tarkempaa tietoa niiden toiminnasta ja merkityksestä.

Kiitokset

Kiitokset ohjaajilleni Tanja Pyhäjärvelle sekä Lumi Viljakaiselle. Kiitos Tanjalle mielenkiintoisista keskusteluista aihepiirin parissa, kirjallisuudesta sekä avusta työn menetelmissä. Kiitos Lumille palautteen antamisesta työhön. Kiitokset myös koko Pyhäjärvi Labille, perheelleni sekä ystävilleni avusta ja tuesta. Kiitos myös Tieteen Tietotekniikan keskus CSC:lle laskennallisista resursseista.

Kirjallisuus

Anderson, S. N., Stitzer, M. C., Brohammer, A. B., Zhou, P., Noshay, J. M., O'Connor, C. H., . . . Springer, N. M. (2019a). Transposable elements contribute to dynamic genome content in maize. *The Plant Journal, 100*(5), 1052–1065. doi:10.1111/tpj.14489

Anderson, S. N., Stitzer, M. C., Zhou, P., Ross-Ibarra, J., Hirsch, C. D., & Springer, N. M. (2019b). Dynamic patterns of transcript abundance of transposable element families in maize. *G3: Genes|Genomes|Genetics, 9*(11), 3673–3682. doi:10.1534/g3.119.400431

Arnholt, A. T., & Evans, B. (2017). BSDA: Basic statistics and data analysis. Saatavilla: https://CRAN.R-project.org/package=BSDA

B73v4.TE.filtered.gff3.gz. (2018). Saatavilla: http://ftp.gramene.org/CURRENT_RELEASE/gff3/zea_mays/repeat_annotation/

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA, 6*, 11. doi:10.1186/s13100-015-0041-9

Baucom, R. S., Estill, J. C., Leebens-Mack, J., & Bennetzen, J. L. (2009). Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. *Genome Research, 19*(2), 243–254. doi:10.1101/gr.083360.108

Bennetzen, J. L., & Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual Review of Plant Biology, 65*, 505–530. doi:10.1146/annurev-arplant-050213-035811

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). Effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software, 5*(56), 2815–2821. doi:10.21105/joss.02815

Boeke, J. D., & Stoye, J. P. (1997). Retrotransposons, endogenous retroviruses, and the evolution of retroelements. Teoksessa: Coffin, J. M., Hughes, S. H., & Varmus, H. E. (toim.), *Retroviruses*. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.

Burt, A., & Trivers, R. (2008). Genes in conflict : The biology of selfish genetic elements. Cambridge, Mass: Harvard University Press.

Cervantes, S., Vuosku, J., & Pyhäjärvi, T. (2021). Atlas of tissue-specific and tissue-preferential gene expression in ecologically and economically significant conifer *Pinus sylvestris*. *PeerJ, 9*, e11781. doi:10.7717/peerj.11781

Chan, P. P., & Lowe, T. M. (2016). GtRNAdb 2.0: An expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Research, 44*(Database issue), D184–D189. doi:10.1093/nar/gkv1309

Combine_all_LTRs.R. (2018). Saatavilla: https://github.com/mcstitzer/w22_te_annotation/blob/master/combine_all_LTRs.R

Cossu, R. M., Casola, C., Giacomello, S., Vidalis, A., Scofield, D. G., & Zuccolo, A. (2017). LTR retrotransposons show low levels of unequal recombination and high rates of intraelement gene conversion in large plant genomes. *Genome Biology and Evolution, 9*(12), 3449–3462. doi:10.1093/gbe/evx260

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., . . . Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10*(2), giab008. doi:10.1093/gigascience/giab008

Deniz, Ö, Frost, J. M., & Branco, M. R. (2019). Regulation of transposable elements by DNA modifications. *Nature Reviews Genetics, 20*(7), 417–431. doi:10.1038/s41576-019-0106-6

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics, 14*(9), 755–763. doi:10.1093/bioinformatics/14.9.755

El Baidouri, M., & Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biology and Evolution, 5*(5), 954–965. doi:10.1093/gbe/evt025

Ellinghaus, D., Kurtz, S. & Willhoeft, U. (2012). LTRharvest : a manual. Saatavilla: http://genometools.org/documents/ltrharvest.pdf

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics, 9*, 18. doi:10.1186/1471-2105-9-18

Feschotte, C., Jiang, N., & Wessler, S. R. (2002). Plant transposable elements: Where genetics meets genomics. *Nature Reviews Genetics, 3*(5), 329–341. doi:10.1038/nrg793

Finnegan, D. J. (1989). Eukaryotic transposable elements and genome evolution. *Trends in Genetics, 5*(4), 103–107. doi:0168-9525(89)90039-5

Fischer, C. N., Carareto, C. M. A., dos Santos, R. A. C., Cerri, R., Costa, E., Schietgat, L., & Vens, C. (2015). Learning HMMs for nucleotide sequences from amino acid alignments. *Bioinformatics, 31*(11), 1836–1838. doi:10.1093/bioinformatics/btv054

Fortune, P. M., Roulin, A., & Panaud, O. (2008). Horizontal transfer of transposable elements in plants. *Communicative & Integrative Biology, 1*(1), 74–77. doi:10.4161/cib.1.1.6328

Friesen, N., Brandes, A., & Heslop-Harrison, J. S. (2001). Diversity, origin, and distribution of retrotransposons (*gypsy* and *copia*) in conifers. *Molecular Biology and Evolution, 18*(7), 1176–1188. doi:10.1093/oxfordjournals.molbev.a003905

Fultz, D., Choudury, S. G., & Slotkin, R. K. (2015). Silencing of active transposable elements in plants. *Current Opinion in Plant Biology, 27*, 67–76. doi:10.1016/j.pbi.2015.05.027

Gallego-Bartolomé, J. (2020). DNA methylation in plants: Mechanisms and tools for targeted manipulation. *New Phytologist, 227*, 38–44. doi:10.1111/nph.16529

Generic feature format version 3 (GFF3). (2020). Saatavilla: https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

GenomeTools. (2014). Saatavilla: http://genometools.org/

Get_nofams.R. (2018). Saatavilla: https://github.com/mcstitzer/w22_te_annotation/blob/master/ltr/families/get_nofams.R

Grandbastien, M.-A. (1998). Activation of plant retrotransposons under stress conditions. *Trends in Plant Science, 3*(5), 181–187. doi:10.1016/S1360-1385(98)01232-1

Grandbastien, M.-A. (2015). LTR retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochimica Et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 1849*(4), 403–416. doi:10.1016/j.bbagrm.2014.07.017

Haque, W., Aravind, A., & Reddy, B. (2009). Pairwise sequence alignment algorithms: A survey. Paper presented at the *Proceedings of the 2009 Conference on Information Science, Technology and Applications,* Kuwait, Kuwait. 96–103. doi:10.1145/1551950.1551980

Heslop-Harrison, J. S., Brandes, A., Taketa, S., Schmidt, T., Vershinin, A. V., Alkhimova, E. G., . . . Harrison, G. E. (1997). The chromosomal distributions of Ty1-*copia* group retrotransposable elements in higher plants and their implications for genome evolution. *Genetica, 100*(1–3), 197–204. doi:10.1023/A:1018337831039

Jedlicka, P., Lexa, M., Vanat, I., Hobza, R., & Kejnovsky, E. (2019). Nested plant LTR retrotransposons target specific regions of other elements, while all LTR retrotransposons often target palindromes and nucleosome-occupied regions: In silico study. *Mobile DNA, 10*, 50. doi:10.1186/s13100-019-0186-z

Kalyanaraman, A., & Aluru, S. (2006). Efficient algorithms and software for detection of full-length LTR retrotransposons. *Journal of Bioinformatics and Computational Biology, 4*(2), 197–216. doi:10.1142/s021972000600203x

Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S., Robinson, D. R., Wu, Y., Cao, X., . . . Chinnaiyan, A. M. (2012). Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell, 149*(7), 1622–1634. doi:10.1016/j.cell.2012.04.041

Kolosha, V. O., & Martin, S. L. (1997). In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. *Proceedings of the National Academy of Sciences of the United States of America, 94*(19), 10155–10160. doi:10.1073/pnas.94.19.10155

Krebs, J. E., Goldstein, E. S., & Kilpatrick, S. T. (2017). Lewin's GENES XII. Burlington, MA: Jones & Bartlett Learning.

Lanciano, S., & Cristofari, G. (2020). Measuring and interpreting transposable element expression. *Nature Reviews Genetics, 21*(12), 721–736. doi:10.1038/s41576-020-0251-y

Lisch, D. (2013). How important are transposons for plant evolution? *Nature Reviews Genetics, 14*(1), 49–61. doi:10.1038/nrg3374

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J. M., Tamarit, D., . . . Moya, A. (2011). The gypsy database (GyDB) of mobile genetic elements: Release 2.0. *Nucleic Acids Research, 39*(Database issue), D70-D74. doi:10.1093/nar/gkq1061

McCarthy, E. M., & McDonald, J. F. (2003). LTR_STRUC: A novel search and identification program for LTR retrotransposons. *Bioinformatics, 19*(3), 362–367. doi:10.1093/bioinformatics/btf878

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences, 36*(6), 344–355. doi:10.1073/pnas.36.6.344

Miele, V., Penel, S., & Duret, L. (2011). Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics, 12*, 116. doi:10.1186/1471-2105-12-116

Mower, J. P., Stefanović, S., Young, G. J., & Palmer, J. D. (2004). Gene transfer from parasitic to host plants. *Nature, 432*(7014), 165–166. doi:10.1038/432165b

Nobuta, K., Venu, R. C., Lu, C., Beló, A., Vemaraju, K., Kulkarni, K., . . . Meyers, B. C. (2007). An expression atlas of rice mRNAs and small RNAs. *Nature Biotechnology, 25*(4), 473–477. doi:10.1038/nbt1291

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., . . . Jansson, S. (2013). The norway spruce genome sequence and conifer genome evolution. *Nature, 497*(7451), 579–584. doi:10.1038/nature12211

Ojeda, D. I., Mattila, T. M., Ruttink, T., Kujala, S. T., Kärkkäinen, K., Verta, J.-P., & Pyhäjärvi, T. (2019). Utilization of tissue ploidy level variation in *de novo* transcriptome assembly of *Pinus sylvestris*. *G3: Genes|Genomes|Genetics, 9*(10), 3409–3421. doi:10.1534/g3.119.400357

Ou, S., & Jiang, N. (2018). LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology, 176*(2), 1410–1422. doi:10.1104/pp.17.01310

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., . . . Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology, 20*, 275. doi:10.1186/s13059-019-1905-y

Pevsner, J. (2015). Bioinformatics and functional genomics. Chichester, West Sussex, UK: Wiley-Blackwell.

Pila v1.5. (2018). Saatavilla: https://treegenesdb.org/FTP/Genomes/Pila/v1.5/genome/

Pita v2.01. (2017). Saatavilla: https://treegenesdb.org/FTP/Genomes/Pita/v2.01/genome/

R Core Team. (2021). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Saatavilla: https://www.R-project.org/

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ, 4*, e2584. doi:10.7717/peerj.2584

RStudio Team. (2021). RStudio: Integrated development environment for R. Boston, MA: RStudio, PBC. Saatavilla: http://www.rstudio.com/

Sabot, F., & Schulman, A. H. (2006). Parasitism and the retrotransposon life cycle in plants: A hitchhiker's guide to the genome. *Heredity, 97*(6), 381–388. doi:10.1038/sj.hdy.6800903

Schulman, A. H. (2013). Retrotransposon replication in plants. *Current Opinion in Virology, 3*(6), 604–614. doi:10.1016/j.coviro.2013.08.009

Slotkin, R. K., Vaughn, M., Borges, F., Tanurdzić, M., Becker, J. D., Feijó, J. A., & Martienssen, R. A. (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell, 136*(3), 461–472. doi:10.1016/j.cell.2008.12.038

Springer, N. M., Anderson, S. N., Andorf, C. M., Ahern, K. R., Bai, F., Barad, O., . . . Brutnell, T. P. (2018). The maize W22 genome provides a foundation for functional genomics and transposon biology. *Nature Genetics, 50*(9), 1282–1288. doi:10.1038/s41588-018-0158-0

Steinbiss, S., Willhoeft, U., Gremme, G. & Kurtz, S. (2012). LTRdigest user's manual. Saatavilla: http://genometools.org/documents/ltrdigest.pdf

Steinbiss, S., Willhoeft, U., Gremme, G., & Kurtz, S. (2009). Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons. *Nucleic Acids Research, 37*(21), 7002–7013. doi:10.1093/nar/gkp759

Stevens, K. A., Wegrzyn, J. L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., . . . Langley, C. H. (2016). Sequence of the sugar pine megagenome. *Genetics, 204*(4), 1613–1626. doi:10.1534/genetics.116.193227

Stitzer, M. C., Anderson, S. N., Springer, N. M., & Ross-Ibarra, J. (2021). The genomic ecosystem of transposable elements in maize. *PLOS Genetics 17*(10), e1009768. doi:10.1371/journal.pgen.1009768

Stuart, T., Eichten, S. R., Cahn, J., Karpievitch, Y. V., Borevitz, J. O., & Lister, R. (2016). Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *eLife, 5*, e20777. doi:10.7554/eLife.20777

Studer, A., Zhao, Q., Ross-Ibarra, J., & Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nature Genetics, 43*(11), 1160–1163. doi:10.1038/ng.942

Tan, G., Polychronopoulos, D., & Lenhard, B. (2019). CNEr: A toolkit for exploring extreme noncoding conservation. *PLOS Computational Biology, 15*(8), e1006940. doi:10.1371/journal.pcbi.1006940

Tollefsbol, T. O. (2017). Chapter 1 - an overview of epigenetics. Teoksessa: Tollefsbol T. O. (toim.), *Handbook of Epigenetics*. Academic Press. doi:10.1016/B978-0-12-805388-1.00001-8

Treangen, T., & Salzberg, S. (2012). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics, 13*(1), 36–46. doi:10.1038/nrg3117

Voronova, A., Rendón-Anaya, M., Ingvarsson, P., Kalendar, R., & Ruņģis, D. (2020). Comparative study of pine reference genomes reveals transposable element interconnected gene networks. *Genes, 11*(10), 1216. doi:10.3390/genes11101216

Vsearch manual. (2019). Saatavilla: https://github.com/torognes/vsearch/releases/download/v2.10.4/vsearch_manual.pdf

W22_te_annotation. (2018). Saatavilla: https://github.com/mcstitzer/w22_te_annotation/

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics, 10*(1), 57–63. doi:10.1038/nrg2484

Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L.-S., Loopstra, C. A., Vasquez-Gross, H. A., . . . Neale, D. B. (2014). Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics, 196*(3), 891–909. doi:10.1534/genetics.113.159996

Wegrzyn, J. L., Lin, B. Y., Zieve, J. J., Dougherty, W. M., Martínez-García, P. J., Koriabine, M., . . . Stevens, K. A. (2013). Insights into the loblolly pine genome: Characterization of BAC and fosmid sequences. *PLOS One, 8*(9), e72439. doi:10.1371/journal.pone.0072439

Weiner, A. M. (2002). SINEs and LINEs: The art of biting the hand that feeds you. *Current Opinion in Cell Biology, 14*(3), 343–350. doi:10.1016/S0955-0674(02)00338-1

Wells, J. N., & Feschotte, C. (2020). A field guide to eukaryotic transposable elements. *Annual Review of Genetics, 54*, 539–561. doi:10.1146/annurev-genet-040620-022145

Wicker, T., & Keller, B. (2007). Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. *Genome Research, 17*(7), 1072–1081. doi:10.1101/gr.6214107

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., . . . Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics, 8*(12), 973–982. doi:10.1038/nrg2165

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., & van der Knaap, E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science, 319*(5869), 1527–1530. doi:10.1126/science.1153040

Xiong, Y., & Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *The EMBO Journal, 9*(10), 3353–3362. doi:10.1002/j.1460-2075.1990.tb07536.x

Zhang, H., Lang, Z., & Zhu, J.-K. (2018). Dynamics and function of DNA methylation in plants. *Nature Reviews Molecular Cell Biology, 19*(8), 489–506. doi:10.1038/s41580-018-0016-z