Suurten kasvigenomien evoluutio ja tutkiminen

Timo Kumpula

LuK-tutkielma

Oulun yliopisto

Genetiikan ja fysiologian tutkimusryhmä

Toukokuu 2016

Sisällys

1. Johdanto

Geneettinen tutkimus on edistynyt nopeasti viimeisen kahdenkymmenen vuoden aikana. Ihmisen genomi saatiin sekvensoitua vuonna 2001 useiden miljardien dollarien rahallisen panostuksen ja vuosien työn tuloksena (Lander ym. 2001). Nyt uuden sukupolven sekvensointimenetelmien (Next Generation Sequencing, NGS) myötä jokainen voi sekvensoida koko genominsa muutamalla tuhannella eurolla. Joka viikko julkaistaan usean uuden lajin koko genomin kattavia analyysejä ja saamme yhä enemmän tietoa genomin yksityiskohtaisesta toiminnasta. Populaatioiden sisällä pystytään jo vertailemaan eri yksilöitä genomin tasolla (Altshuler ym. 2010). Pelkkien geenisekvenssien tutkimisesta on siirrytty monimutkaisten tapahtumaketjujen tulkintaan, kuten epigeneettisten tekijöiden ja ei-koodaavan DNA:n vaikutuksiin geenien ilmenemisessä. NGS-menetelmät mahdollistavat transkriptien tunnistamisen ja luokittelun ilman aiempaa tietoa niitä koodaavista geeneistä, mikä helpottaa niiden tutkimista (Dijk ym. 2014).

Yhä edullisemmat sekvensointilaitteet mahdollistavat aina vain useampien yliopistojen ja yksityisten laboratorioiden omatoimisen genomien tutkimuksen ja kartoittamisen. Bioinformatiikan merkitys suurten aineistomäärien käsittelyssä on lisääntynyt. Riittävän aineiston kerääminen ei ole enää ongelma, vaan pikemminkin sen tulkinta. Genomien sekvensointiin ja koostamiseen tuottavat lisähaastetta erityisesti polyploidia, toistuvat elementit ja heterotsygotia. Kasvien genomit ovat yleensä eläinten genomeja haastavampia tutkia, koska monilla kasvilajeilla on todella suuret genomien koot, suuri määrä toistuvia elementtejä ja haastavia kokonaisia tai osittaisia kromosomien kaksinkertaistumisia (Hamilton & Buell ym. 2012).

Ensimmäinen kasvi, jonka koko genomi saatiin sekvensoitua, oli lituruoho (*Arapidopsis thaliana*) vuonna 2000. Se on hyvä malliorganismi, koska sillä on pieni diploidi noin 135 Mb (Mega base, 10^6 emäsparia) genomi (2n=10). Lisäksi sen kasvattaminen on laboratorio-olosuhteissa vaivatonta (Kaul ym. 2000). Nopeasta menetelmien kehityksestä huolimatta suurigenomisten kasviryhmien tutkiminen on edelleen haastavaa. Tällaisia kasviryhmiä ovat muun muassa havupuut ja viljakasvit. Havupuilla genomien koot vaihtelevat 20–40 Gb (Giga base, 10^9 emäsparia) välillä, eli ne ovat noin kymmenen kertaa ihmisen genomia suurempia (n. 3 Gb) (Neale ym. 2014). Viljakasvien genomit ovat hieman pienempiä, mutta silti esimerkiksi vehnän genomi on n. 17 Gb (Mayer ym. 2014). Näillä lajiryhmillä perinteiset genomin koostamis- ja tulkitsemismenetelmät ovat tehottomia ja tuottavat paljon virheitä. Viimeaikoina etenkin näiden suurigenomisten lajiryhmien tutkimisessa on saavutettu merkittäviä edistysaskelia.

Tässä tutkielmassa käsittelen suurten genomien erityispiirteitä ja menetelmiä, joilla niihin liittyviä haasteita on pyritty ratkaisemaan. Tarkastelen erityisesti genomin sekvensointiin ja koostamiseen liittyviä kysymyksiä, koska ne ovat tutkimuksen kannalta kriittisiä vaiheita. Kerron myös minkälaisista elementeistä suuret genomit muodostuvat ja pohdin syitä niiden kehittymiseen. Perehdyn tarkemmin yhteen suurigenomiseen lajiryhmään, jonka merkitys ihmiselle on todella suuri: havupuihin. Tästä lajiryhmästä käsittelen vielä erityisesti mäntyä ja kuusta, koska niiden genomeista on saatu viimevuosien tutkimuksissa paljon uutta merkittävää tietoa.

Mistä suuret genomit muodostuvat ja miten ne ovat kehittyneet?

Genomin kokoa kuvataan usein C-arvon avulla. Se kertoo kuinka paljon eliön yksi haploidi solu, esimerkiksi sukupuolisolu, sisältää DNA:ta pikogrammoissa (1C). Yksi pikogramma DNA:ta vastaa karkeasti 978 miljoonaa emäsparia (noin 1 Gb) (Dolezel ym. 2003). Diploideilla organismeilla C-arvo ja genomin koko tarkoittavat samaa asiaa. Polyploidisten lajien kohdalla C-arvon määritelmä vaikeutuu, koska silloin se voi tarkoittaa myös useamman kuin yksinkertaisen kromosomiston massaa (Greilhuber ym. 2005). Usein genomin koko kuitenkin ilmoitetaan massan sijaan suoraan emäsparien lukumääränä.

C-arvo vaihtelee eri lajien välillä todella paljon. Myös hyvin läheistä sukua toisilleen olevien lajien välillä erot genomin koossa voivat olla suuria. Tämä on yleistä muun muassa kasveilla, joilla näin voi helposti käydä koko genomin kaksinkertaistumisen (Whole Genome Duplication, WGD) tai hybridisaation avulla tapahtuneen lajiutumisen seurauksena. Toisaalta myös yksisoluisilla eliöillä voi olla todella suuria genomeja. Tämä pätee erityisesti tumallisilla alkueliöillä, mutta myös bakteereilla voi olla suuria genomeja. Suurin sekvensoitu bakteerigenomi on maaperässä elävän *Ktedonobacter racemifer*in n. 14 Mb genomi (Chang ym. 2011). Ristiriitaa genomin koon ja eliön monimutkaisuuden välillä kutsutaan C-arvo paradoksiksi (Lee & Kim 2014).

Kasvikunnan sisällä lajikohtainen genomin koko vaihtelee kaikista taksoneista eniten. Ne voivat vaihdella jopa 2348-kertaisesti pienimmästä genomista (*Genlisea margaretae* (1C = 0,063 Gb)) suurimpaan genomiin (*Paris japonica* (1C = 148,88 Gb)). Koppisiemenisten kasvien keskimääräinen genomin koko (1C) on 5,81 Gb ja paljassiemenisten 18,16 Gb. Suurimmalla osalla sekvensoiduista kasvilajeista genomin koko on n. 500 Mb (Lee & Kim 2014). DNA:ta löytyy kasveista tuman lisäksi myös mitokondrioista ja kloroplasteista.

Suuret genomit eivät sisällä keskimääräisen kokoista genomia enempää geenejä, vaan suurin osa genomista muodostuu proteiineja koodaamattomista toistuvista sekvensseistä. Niitä löytyy runsaasti muun muassa osittain tai kokonaan kaksinkertaistuneista kromosomeista, koska tyypillisesti vain pieni osa kromosomista koodaa proteiineja. Esimerkiksi metsäkuusella (*Picea abies*) korkeintaan 2,6 % genomista muodostuu geenien kaltaisista sekvensseistä (Nystedt ym. 2014). Muita yleisiä ei-koodaavia alueita ovat muun muassa pseudogeenit, transposonit, intronit ja lyhyet toistojaksot (esim. mikrosatelliitit). Organismin genomin kokoon vaikuttaakin ratkaisevasti muun muassa sen alttius ”kerätä” itseensä näitä toistuvia elementtejä. Esimerkiksi maissi (*Zea Mays*) ja riisi (*Oryza sativa*) ovat molemmat diploideja viljakasveja, joilla on melkein sama määrä kromosomeja (maissilla 20 ja riisillä 24) ja geenejä (n. 35000). Kuitenkin maissin genomin koko on 2300 Mb ja riisin vain 415Mb (Lee & Kim 2014). Maissin genomista jopa lähes 85 % muodostuu erilaisista transposoneista (Schnable ym. 2009).

1.1. Kromosomi- ja kromosomistomutaatiot

Kromosomeihin ja kromosomistoihin liittyvät mutaatiot ovat nopein tapa lisätä genomin kokoa. Kromosomien määrä eri kasvilajeilla vaihtelee aina neljästä (2n=4) jopa yli kuuteen sataan (2n=640). Pelkkä kromosomien suuri määrä ei kuitenkaan kerro mitään genomin koosta. Esimerkiksi paljassiemenisillä kasveilla, joilla suuret genomit ovat yleisiä, valtaosa lajeista on diploideja ja varsin vähäkromosomisia (2n=14-28). Kromosomien lukumäärä voi olla hyvin yhtäläinen suvun sisällä, esimerkiksi kaikista tutkituista 232 mäntyheimon (Pinaceae) lajista kaikilla paitsi yhdellä 2n=24. Kromosomien koot voivat kuitenkin vaihdella paljon eri lajien välillä (Heslop-Harrison & Schwarzacher 2011).

Polyploidialla tarkoitetaan useamman kuin kahden haploidin kromosomiston muodostamaa genomia. Sen määrä voi vaihdella kolminkertaisesta triploidiasta jopa yli kymmenkertaisiin kromosomistoihin. Kasveilla se on huomattavasti eläimiä yleisempää. Se on merkittävä evoluutiota nopeuttava tekijä ja on edelleen tärkeä osa kasvien lajiutumista (Heslop-Harrison & Schwarzacher 2011). On arvioitu, että noin puolet koppisiemenisistä kasveista on polyploidisia. Aineistojen perustella voidaan sanoa, että toistuvat WGD-tapahtumat ovat tavallinen osa kasvien evolutiivista historiaa (Soltis ym. 2015). Paljassiemenisillä polyploidia on paljon harvinaisempaa (n. 5 %) ja esimerkiksi havupuista vain noin 1,5 % on polyploidisia (Ahuja & Neale 2005). Muinaisia polyploidisaatioita on vaikea havaita genomista, koska aika hävittää kaksinkertaistumisen jäljet. Esimerkiksi kaksinkertaistuneista geeneistä 70-90 % ehtii palautua takaisin yksinkertaiseen

muotoon ja kromosomien väliset uudelleenjärjestelyt voivat viedä kaksinkertaistuneita sekvenssialueita eri puolille genomia. Kaikki tämä tekee genomin sisäisen syntenian havaitsemisesta todella haastavaa. Syntenialla tarkoitetaan kromosomien alueiden evolutiivista yhteyttä toisiinsa. Joskus polyploidisaatio voi myös pienentää genomin kokoa (Ahuja & Neale 2005). Kokonaisten kromosomien lisäksi myös pienemmät alueet kromosomista voivat monistua.

Polyploidian vaikutukset sekvensointiin ja genomin koostamiseen riippuvat muun muassa onko kyseessä allo- vai autopolyploidia ja kuinka pitkä aika polyploidisaatiosta on kulunut (Hamilton & Buell ym. 2012). Allopolyploidiassa kantalajien täydelliset kromosomistot yhdistyvät tuottaen uuden hybridilajin. Näin voi nopeasti syntyä suurigenomisia uusia lajeja. Esimerkiksi ruisvehnä on allopolyploidinen laji, jossa rukiin ja vehnän kromosomistot ovat yhdistyneet. Hybridit ovat usein alkuperäislajeja kookkaampia ja saattavat olla ominaisuuksiltaan parempia. Syyksi on arveltu esimerkiksi sukusiitosheikkouden vähentymistä (Chen 2010). Lajin fertiliteetti saattaa kuitenkin laskea polyploidian myötä. Parittomat määrät kromosomeja johtavat steriileihin yksilöihin, koska kromosomit eivät mene tasan solunjaossa (Heslop-Harrison & Schwarzacher 2011). Esimerkiksi kaupalliset triploidit banaanit on tarkoituksella jalostettu polyploidisiksi siementen kehittymisen estämiseksi, joten niiden lisääntyminen tapahtuu vegetatiivisesti eli jälkeläiset ovat isäntäkasvin klooneja. Kun eliön oma genomi moninkertaistuu, kyse on autopolyploidiasta.

1.2. Monistuneet geenit ja intronit

Geenien kopioituminen genomin sisällä on tavallinen genomin kokoa kasvattava mekanismi. Suuri osa geeneistä esiintyy genomista kahtena tai useampana kopiona. Niitä voi syntyä usealla eri mekanismilla: jo aikaisemmin käsiteltyjen kokonaisten tai osittaisten kromosomiduplikaatioiden lisäksi myös epätasainen crossing-over ja geenikonversio ovat tavallisia syitä (Ahuja & Neale 2005). Yleensä ajan kuluessa monistunut kopio menettää alkuperäisen funktionsa ja toiminnallisuutensa mutaatioiden kertymisen seurauksena. Tällaista aluetta sekvenssissä kutsutaan pseudogeeniksi. On kuitenkin mahdollista, että mutaatioiden ansiosta jompikumpi geeneistä kehittää täysin uuden funktion ja antaa näin evoluutiolle lisää työkaluja muokata organismeja. Joskus kaksinkertaistuneet geenit voivat myös jäädä yhdessä hoitamaan hieman eri tavalla sitä tehtävää, mitä alkuperäinen geeni yksin suoritti (Walsh 2003). Monistuneiden geenien absoluuttinen merkitys genomin koolle on vähäinen. Esimerkiksi loblollymännyllä (*Pinus taeda)* 2,9 % genomista muodostuu pseudogeeneistä (Neale ym. 2014).

Intronien rooli genomin koon kasvussa on epäselvä. Joidenkin lajien välisissä vertailuissa intronien koko näyttää korreloivan genomin koon kanssa. Esimerkiksi pallokalan (*Fugus rubripes*) intronit ovat pituudeltaan keskimäärin kahdeksasosan ihmisen introneista, mikä vastaa koko genomien välistä suhdetta (n. 400 Mb/n. 3000 Mb) (Ahuja & Neale 2005). Tämä pätee osittain myös kasvikuntaan. Oletuksen mukaisesti viiniköynnöksellä (*Vitis vinifera)* ja *Amborella trichopodalla* intronien keskimääräinen pituus ja genomiin koon suhde on molemmilla noin kahden suhde yhteen (933bp/487Mb ja 1538bp/706Mb) (Taulukko 1, Neale ym. 2014). Toisaalta monilla kasvilajeilla intronit eivät pitene genomin koon kasvaessa. Taulukon 1 kasvilajeista näkee helposti, että useimmissa tapauksissa intronien pituus ja genomin koko eivät korreloi. Esimerkiksi loblollymännyllä ja metsäkuusella on arvioitu olevan melkein samankokoiset genomit (20148Mb/19600Mb), mutta loblollymännyn intronit ovat keskimäärin melkein kolme kertaa pidemmät (2741bp/1020bp). Ainakin kasvikunnassa siis näyttäisi siltä, että genomin koolla ei ole vahvaa yhteyttä intronien kokoon. Myös intronien absoluuttinen merkitys genomin koolle on pieni. Loblollymännyllä vain noin 0,3 % genomista muodostuu introneista, joista jopa 60 % on erilaisia toistuvia elementtejä (Neale ym. 2014). Toisaalta Nystedt ym. 2013 esittivät, että pitkän aikavälin genomin hidas kasvu on johtanut havupuiden poikkeuksellisen pitkiin introneihin.

Taulukko 1. Kuuden sekvensoidun kasvilajin genomisten tunnuslukujen vertailu (Neale ym. 2014). Ylhäällä näkyy kasvilajin tieteellinen nimi. Vasemmalla näkyvät tunnusluvut ovat ylhäältä alaspäin: Genomin koostettu koko emäspareina (Mb), kromosomien lukumäärä, GC-pitoisuus, transposonien osuus genomista, geenien lukumäärä, keskimääräinen geenin eksonien yhteispituus emäspareina, keskimääräinen intronin pituus emäspareina ja genomin pisimmän intronin pituus emäspareina. Lisämerkintöinä metsäkuusen (*Picea abies*) genomin todellinen estimoitu koko on 19,6 Gb (a), geenien lukumäärä tarkoittaa kokonaisia varmistettuja yli 150 emäksen geenejä (b) ja metsäkuusen geenien lukumäärä perustuu Congenie-projektin (Nystedt ym. 2013) korkean ja keskitason luottamuksen saaneisiin geeneihin (c).

	Pinus taeda	*Picea abies* [8]	*Arabidopsis thaliana* [21]	*Populus trichocarpa* [21]	*Vitis vinifera* [21]	*Amborella trichopoda* [22]
Genome size (assembled) (Mbp)	20,148	12,019[a]	135	423	487	706
Chromosomes	12	12	5	19	19	13
G + C content (%)	38.2	37.9	35.0	33.3	36.2	35.5
TE content (%)	79	70	15.3	42	41.4	N/A
Number of genes[b]	50,172	58,587[c]	27,160	36,393	25,663	25,347
Average CDS length (bps)	965	723	1102	1143	1095	969
Average intron length (bps)	2,741	1,020	182	366	933	1,538
Maximum intron length (bps)	318,524	68,269	10,234	4,698	38,166	175,748

[a]Estimated genome size is 19.6 Gbp.
[b]Number of full-length genes >150 bp in length and validated through current annotations.
[c]High and medium confidence genes from the Congenie project [8].

1.3. Peräkkäiset toistojaksot

Peräkkäisillä toistojaksoilla tarkoitetaan satelliitti DNA:ta, mikrosatelliitteja ja minisatelliitteja. Ne ovat GC-rikkaita toistojaksoja, joita esiintyy runsaasti varsinkin kromosomien sentromeeri- ja telomeerialueilla. Satelliitit ovat sadasta jopa useaan tuhanteen emäkseen pitkiä, kun taas minisatelliitit ovat kymmenestä sataan emäkseen ja mikrosatelliitit kahdesta yhdeksään emäkseen (Neale ym. 2014). Ne ovat useimmiten järjestäytyneet moneksi peräkkäiseksi kopioiksi samaa sekvenssiä ja saavat aikaan kromatiinin tiukemman pakkautumisen, joten niillä on suuri merkitys esimerkiksi solunjakautumisessa. Jopa miljoonan emäksen mittainen alue voi koostua yhdestä tuhansia kertoja toistuvasta satelliittisekvenssistä. Kasveilla suurin osa peräkkäisistä toistojaksoista on minisatelliitteja, mutta niiden merkitys genomin kokoon on melko vähäinen. Esimerkiksi loblollymännyn genomista ne muodostavat vain n. 2,86 % (Neale ym. 2014).

1.4. Transposonit

Transposoneilla (Transposable Elements, TEs) tarkoitetaan sellaisia sekvenssielementtejä, joille on kehittynyt kyky omatoimisesti liikkua ja monistua genomin sisällä. Tutkimusmenetelmien nopea kehitys on mahdollistanut niiden tutkimuksen aivan uudella tavalla. Ne jaetaan kahteen pääluokkaan, yhdeksään heimoon ja 29 perheeseen kopiointimekanismin, geenirakenteen ja sekvenssin samankaltaisuuden perusteella (Lee & Kim 2014). Pääluokat eroavat toisistaan kopiointimekanismin perusteella: Luokan I transposonit ovat retrotransposoneja, jotka kopioituvat koodaamalla mRNA:ta, joka käänteiskopioijaentsyymin katalysoimana tekee uuden identtisen kopion sekvenssistä johonkin satunnaiseen kohtaan genomia. Usein myös itse entsyymi on retrotransposonin koodaamaa. Mekanismi muistuttaa hyvin paljon retrovirusten (esim. HIV-virus) kopiointimekanismia. Luokan II transposonit ovat DNA-transposoneja, jotka käyttävät hyväkseen ”leikkaa-liitä” – periaatetta. Ne eivät siis koodaa mRNA:ta, vaan kopioituvat käyttämällä hyväkseen transposaasi-entsyymejä (Lee & Kim 2014). Retrotransposonit jaetaan kahteen luokkaan: LTR (long terminal repeat)-retrotransposoneihin ja non-LTR-retrotransposoneihin. Molempia löytyy kasvien genomeista, mutta LTR-retrotransposonit ovat yleisempiä (Lee & Kim 2014). LTR-retrotransposonit ovat läheistä sukua varsinaisille retroviruksille, koska ainoa merkittävä ero on, että retrovirukset koodaavat suojakuoren proteiinia (Envelope protein, ENV).

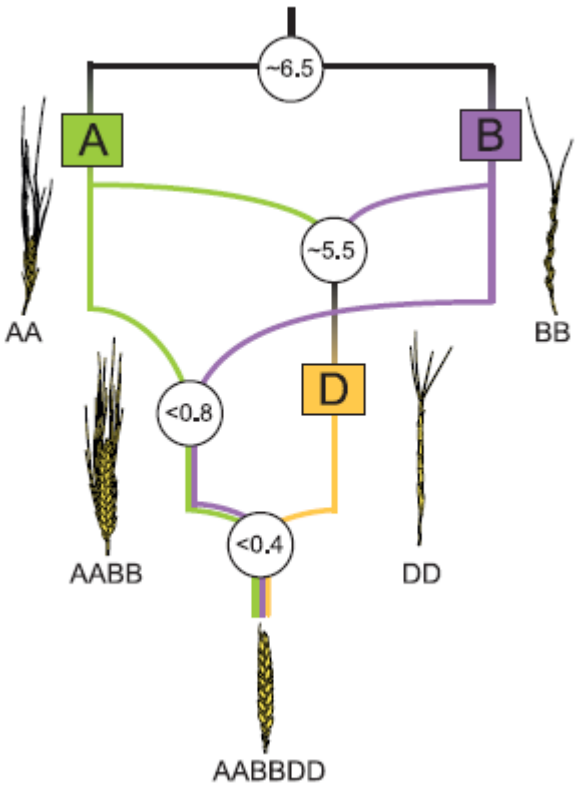
Transposonit voivat saada aikaan genomista plastisuutta aiheuttamalla kromosomimutaatioita ja lisäämällä alleelien erikoistumista (Lee & Kim 2014). Niiden yleisin vaikutus on kuitenkin genomin koon hidas kasvaminen, koska niiden kopioluku voi nousta todella korkeaksi genomin sisällä.

Suurin osa eukaryoottien proteiineja koodaamattomasta DNA:sta muodostuukin erityyppisistä transposoneista. Esimerkiksi loblollymännyllä 62 % genomista muodostuu retrotransposoneista, joista 70 % on LTR-retrotransposoneja (Neale ym. 2014). Maissin genomista 75 % on LTR-retrotransposoneja ja 8,6 % DNA-transposoneja. Yhteensä ne muodostavat genomista jopa lähes 85 % (Schnable ym. 2009). LTR-retrotransposonipareja käytetään tutkimuksessa muun muassa insertioiden ajoittamiseen.

1.5. Mahdollisia selityksiä suurten genomien kehittymiselle

Kaikkien eliölajien genomit kokevat muutoksia ajan kuluessa ja ovat siten jatkuvassa muutostilassa. Genomin koko voi kasvaa tai pienentyä, eivätkä ne ole toisiaan poissulkevia vaihtoehtoja. Kasvien kyky torjua toistuvien elementtien kertymistä vaikuttaa pitkällä tähtäimellä ratkaisevasti genomin kokoon. Nystedt ym. (2013) esittivät, että havupuiden genomien suuren koon taustalla olisi juuri näiden mekanismien heikko toimintateho verrattuna esimerkiksi koppisiemenisiin. Muita selityksiä voisivat olla intronien koon jatkuva kasvu ja pseudogeenien suuri määrä. Toistuvien elementtien torjuntakoneiston heikko teho havupuilla voi selittyä ainakin osittain 24-nukleotidin sRNA:n matalalla ekspressiolla (Nystedt ym. 2013). Näiden RNA-molekyylien yksi tärkeä tehtävä on metyloida toistuvien elementtien sekvenssejä, jolloin niiden leviäminen ja kopioituminen hidastuu. Metsäkuusen (*Picea abies*) kromosomit ovat kaikki kehittyneet lähes samankokoisiksi, mikä viittaa siihen, että niiden kasvulle voi olla jokin fyysinen este (Nystedt ym. 2013).

Toistuvien elementtien hitaan kasautumisen lisäksi etenkin polyploidisaatio on merkittävä tekijä suurten genomien evoluutiossa. Esimerkiksi heksaploidi leipävehnä on muodostunut kolmen eri kromosomiston yhdistymisen tuloksena (Kuva 1, Marcussen ym. 2014). Tuoreet tutkimukset myös osoittavat, että itse asiassa myös havupuiden historiassa on tapahtunut genomin kokoa kasvattaneita polyploidisaatioita (Li ym. 2015). Polyploidit lajit ovat myös saattaneet saada valintaedun runsaasta geneettisestä materiaalista. Fawcett ym. 2009 esittivät, että suurella osalla koppisiemenisiä kasveja nuorin WGD-tapahtuma ajoittuu liitukauden massasukupuuttoon noin 66 miljoonaa vuotta sitten, jolloin noin 60 % kasvilajeista kuoli sukupuuttoon. Tämä tukee teoriaa siitä, että polyploidiset lajit olisivat olleet parempia sopeutumaan vaihteleviin ilmasto-olosuhteisiin (Fawcett ym. 2009).



Kuva 1. Leipävehnän (*Triticum aestivum; AABBDD)* fylogeneettinen historia (Marcussen ym. 2014). Kantalajin jakautuminen ja kolmen hybridisaatiotapahtuman arvioidut ajankohdat miljoonina vuosina on merkitty kuvaan valkoisilla ympyröillä. Vehnän eriytyminen yhteisestä esi-isästä A ja B sukulinjoihin (*Triticum* ja *Aegilops*) tapahtui n. 6,5 miljoonaa vuotta sitten. Ensimmäinen hybridisaatio tapahtui n. 5,5 miljoonaa vuotta sitten A ja B linjojen välillä ja johti genomilinjan D syntyyn. Seuraava hybridisaatio tapahtui hieman alle 0,8 miljoonaa vuotta sitten kahden läheisen sukulaisen välillä *Ae. speltoides* (BB) ja *T. urartu* (AA), joka johti allotetraploidisen emmervehnän *T.turgidum* (AABB) syntyyn polyploidisaation kautta. Leipävehnä syntyi allopolyploidisaation seurauksena kolmannessa hybridisaatiossa hieman alle 0,4 miljoonaa vuotta sitten emmervehnän ja *Ae. tauschii* (DD) välillä. Kolme diploidia sukulinjaa on merkitty kuvaan erivärisillä viivoilla.

2. Genomin sekvensointi ja koostaminen

Genomin sekvensoinnilla tarkoitetaan DNA:n emäsjärjestyksen selvittämistä (poikkeuksena RNA-virukset). Sen lopputuloksena saadaan yksi tai useampi sekvenssi, eli lineaarinen esitys tutkittavan molekyylin rakenteesta. Sekvensointimenetelmät ovat kehittyneet vuonna 1977 keksitystä Sanger-sekvensoinnista nopeasti toisen ja kolmannen sukupolven menetelmiin.

Genomin koostaminen (assembly) tarkoittaa sekvensoitujen genomin osien yhdistelyä konsensussekvenssiksi. Sen toteuttamiseen on kolme päätapaa. Ensimmäiset kokonaiset

kasvigenomit koostettiin 2000-luvun alussa sekvensoimalla bakteeriklooneja (Bacterial Artificial Chromosomes, BACs), joihin oli liitetty tutkittavia inserttejä (Clone-By-Clone-menetelmä, CBC). CBC-menetelmällä tutkittavan genomin kromosomeista saadaan laadittua fyysinen kartta (mapping) eli jokaisella sekvenssillä on tunnettu paikka kromosomissa. Muun muassa tästä syystä koostettu konsensussekvenssi on yleensä korkealaatuista (Bolger ym. 2014). Haittapuolena on suuri työmäärä ja korkea hinta. Jättipoppelin (*Populus trichocarpa*) genomi koostettiin 2006 käyttämällä uutta koko genomin satunnaista sekvensointia (Whole Genome Shotgun-menetelmä, WGS) (Tuskan ym. 2006). Siinä koko genomi pilkotaan ja sekvensoidaan satunnaisesti lyhyissä osissa (lukemat). Lukemat asetetaan samankaltaisuuden perusteella yhteen, jolloin saadaan pitempiä jatkuvia sekvenssejä (kontiki). Kontikit voidaan edelleen yhdistää samankaltaisuuden perusteella skaffoldeiksi. Koostamisen päämääränä voi olla jokaisen kromosomin esittäminen yhtenä kokonaisena skaffoldina. Kolmas tapa on erotella kromosomit yksitellen erilleen sekvensointia ja koostamista varten. Tämä on helpottanut etenkin polyploidisten kasvilajien genomien koostamista, koska silloin tiedetään mistä kromosomista lukemat ovat peräisin (Bolger ym. 2014).

Tärkeintä koostamisen onnistumisen kannalta on lukemien pituus. Lyhyet lukemat (n. 100-250bp) tuottavat etenkin paljon toistuvaa sekvenssiä sisältävien suurten genomien kohdalla ongelmia koostamisvaiheessa, koska koostamisohjelmat eivät tiedä, ovatko alleelit olleet alun perin peräkkäin vai rinnakkain. Lopputuloksena toistuvat alueet koostetaan konsensussekvenssiin helposti virheellisesti (Fuller ym. 2009). Tämä on suuri ongelma etenkin WGS-pohjaisissa koostamisissa ja niiden onnistuminen vaatiikin usein jo olemassa olevan referenssigenomin hyödyntämistä (Bolger ym. 2014). Ratkaisuna tähän ongelmaan uudet sekvensointimenetelmät pyrkivät samaan aikaan pidempiä lukemia.

Sekvensointi- ja koostamismenetelmissä on tärkeää huomioida lähtö DNA:n riittävä määrä ja laatu, sekä siitä tehtävien DNA-kirjastojen onnistuminen. Tyypillisesti lähtö-DNA valmistellaan kirjastoksi fragmentoimalla se sekvensoinnille riittävän lyhyiksi pätkiksi esimerkiksi ultraäänellä monistamalla se ensin esimerkiksi polymeraasiketjureaktiolla (PCR). Harsuuntuneet päät korjataan ja sekvensseihin lisätään adaptereja ligaasien ja transferaasien avulla polymeraasia, sekä erilaisia sekvensointialustoja varten (Fuller ym. 2009). Erimittaisista fragmenteista valikoidaan usein halutun mittaiset sekvenssit liian lyhyiden pätkien ja ylimääräisten adapterien karsimiseksi (Dijk ym. 2014).

Pitkien DNA-molekyylien koostamisen helpottamiseksi molekyylin molemmista päistä voidaan sekvensoida lyhyt pätkä sekvenssiä (paired end reads). Jos sekvensoitavan molekyyli pituus tunnetaan onnistuneen laboroinnin tuloksena, niin koostamisvaiheessa sekvensoitujen päiden välinen etäisyys auttaa lukemien ryhmittelyssä ja tulkitsemisessa. Jos molekyyli on esimerkiksi 500 emäksen mittainen ja konsensussekvenssissä lukemat ovatkin 200 emäksen päässä toisistaan, niin silloin joko lähtö-DNA:ssa on tapahtunut insertio tai konsensussekvenssissä deleetio. Tämä toimii hyvin vielä alle tuhannen emäksen sekvensseissä, mutta pitemmissä täytyy käyttää monivaiheisempaa kirjastoa (mate-pair library) (Henson ym. 2012).

2.1. Uuden sukupolven sekvensointimenetelmät (NGS)

DNA:n uuden sukupolven sekvensointimenetelmillä tarkoitetaan sekä toisen että kolmannen sukupolven menetelmiä. Toisen sukupolven sekvensointimenetelmät ovat tällä hetkellä jo kattavasti koko tiedeyhteisön käytössä ja niillä suoritettu sekvensointi on suurten genomien kohdalla yleisesti hyväksytty standardi. Kolmannen sukupolven menetelmiä kehitetään jatkuvasti nopeaa tahtia, mutta niiden käyttö on vielä selvästi toisen polven menetelmiä harvinaisempaa. Ne ovat kuitenkin jo käytössä kaikista edistyneimmissä laboratorioissa. Sanger-sekvensoinnista siirtyminen NGS-menetelmiin 2000-luvun lopulla on mullistanut genomien koostamisen.

Tällä hetkellä kaikista suosituimpina toisen polven menetelminä voidaan pitää Illuminaa, sekä Life Technologiesin Ion Torrenttia. Näistä kahdesta Illumina on saavuttanut vahvemman aseman suuremman suoritustehon ja halvemman hinnan ansiosta (Dijk ym. 2014). Molemmissa sekvensointi perustuu siihen, että yksijuosteisen kohde-DNA:n vastinjuoste syntetisoidaan useassa rinnakkaisessa reaktiossa käyttämällä hyväksi joko DNA-polymeraasia tai ligaasientsyymejä (sequencing by synthesis-periaate, SBS) (Fuller ym. 2009). Ne ovat molemmat tehokkaita ja suhteellisen halpoja menetelmiä, joiden avulla tutkittavasta genomista saadaan syntetisoitua suuri määrä suhteellisin lyhyitä, noin 100-250 nukleotidin mittaisia lukemia.

Kolmannen sukupolven menetelmät pyrkivät saamaan aikaan pidempiä lukemia. Niissä ei enää koosteta sekvenssiä miljoonien lyhyiden lukemien avulla, vaan ideana on sekvensoida pitkiä DNA-molekyylejä reaaliajassa. DNA:ta ei siis tarvitse monistaa ja valmistella kirjastoksi spesifien reagenssien avulla, vaan jo yhdestä DNA-molekyylistä voidaan saada tarvittava määrä tietoa (Dijk ym. 2014). Oxford nanopore ajaa DNA-molekyylin yksijuosteisena sensitiivisen kalvon läpäisevän ohuen nanohuokosen läpi samalla mitaten eri emästen aiheuttamia jännitemuutoksia.

Pacific Bioscience käyttää SBS-menetelmää lukemalla kaivoissa tapahtuvan synteesin lähettämiä signaaleja. Se saa aikaan pitkiä, jopa 10-15 Kb mittaisia lukemia, mutta toisaalta menetelmässä syntyy myös paljon virheitä (noin 14 %) (Roberts ym. 2013). Muilla johtavilla menetelmillä virhe-esiintyvyys on 0,1-1 % luokkaa. Pacificin virhemalli on kuitenkin muista menetelmistä poiketen täysin satunnainen, joten konsensussekvenssistä saadaan suuresta virhe-esiintyvyydestä huolimatta hyvälaatuista (Roberts ym. 2013). Virheet siis jakautuvat tasaisesti ympäri sekvenssiä, joten koostamisvaiheessa ne saadaan hyvällä todennäköisyydellä korjattua vertailemalla useita lukemia toisiinsa. Lisäksi tulosten kokonaispeittävyys on erinomainen, sillä sekvensointireaktion reagenssit mahdollistavat myös vahvojen GC-alueiden sekvensoinnin (Roberts ym. 2013).

Uusimpana ja edistyksellisimpänä kolmannen polven metodina voidaan pitää 10x Genomicsin GemCode™ -menetelmää. Siinä muodostetaan aluksi paljon lyhyitä lukemia, mutta toisin kuin muissa menetelmissä, jokainen lukema saa lähtösolusta tai -molekyylistä riippuvan oman yksilöllisen molekyyliviivakoodin. Viivakoodien avulla samasta lähtömolekyylistä tai – solusta peräisin olevat lyhyet lukemat saadaan yhdistettyä yli 100 Kb mittaisiksi lukemiksi. Pitkien lukemien koostaminen on lyhyitä helpompaa ja näin saadaan toistuvien alueiden ongelma osittain ratkaistua. Pitkät lukemat voidaan myös viivakoodien avulla järjestää siten, että haplotyypit voidaan tunnistaa. Tiedetään siis kumpi kromosomi on tullut isältä ja kumpi äidiltä (10x Genomics™ 2016).

Erilaisten sekvensointimenetelmien vertailussa on kuitenkin otettava huomioon, että eri menetelmät eivät sovellu tasaisesti kaikenlaisille tutkimuksille. Esimerkiksi Pacific Biosciencen menetelmä sopii erityispiirteidensä ansiosta erinomaisesti muun muassa pienten genomien (kuten virusten ja bakteerien), sekä korkeiden GC-pitoisuuksien sekvensoimiseen (Roberts ym. 2013).

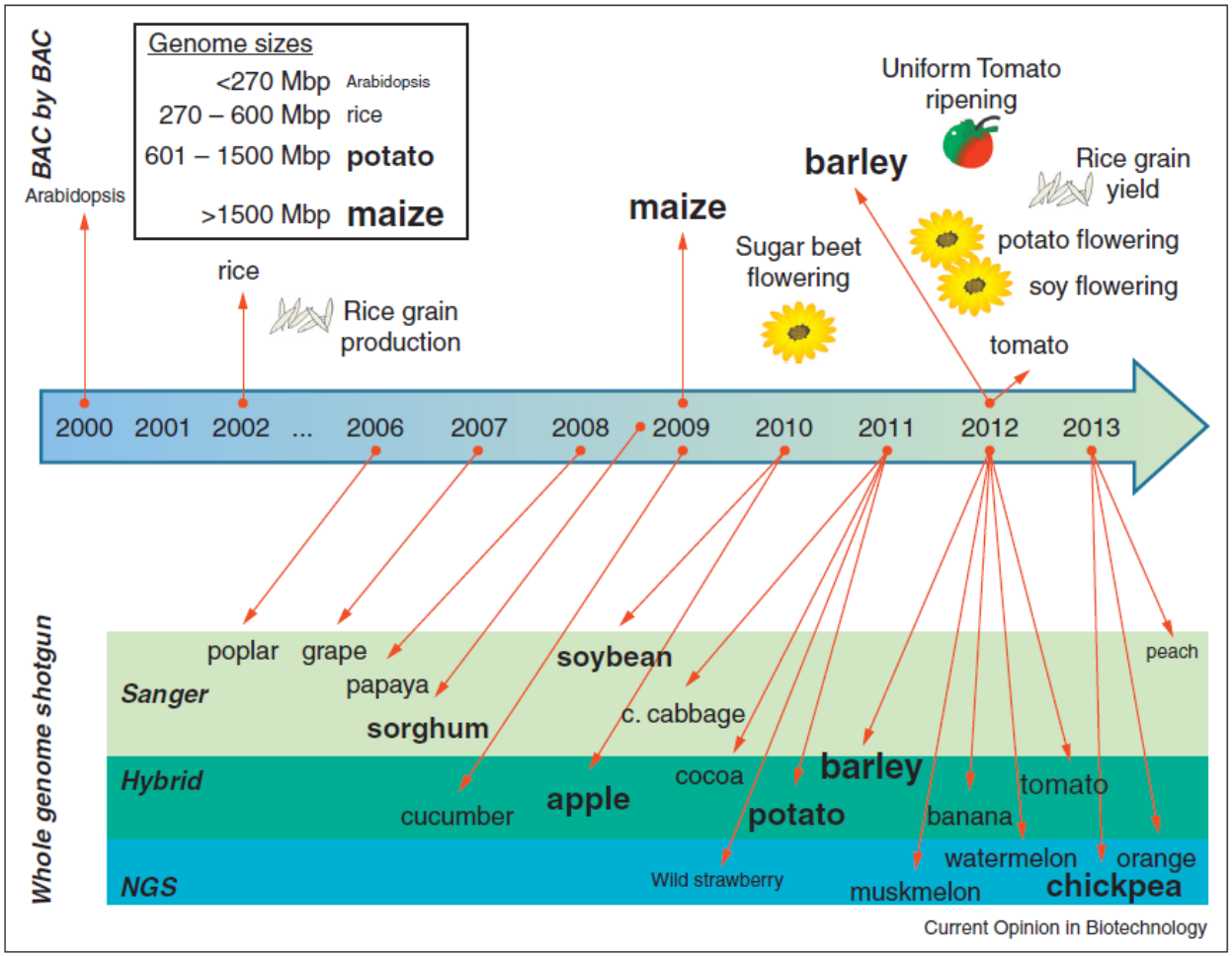
2.2. Kasvigenomien koostaminen

Kasvien genomit ovat vaativia koostettavia. Merkittävin haaste on toistuvien elementtien suuri määrä, mikä vaikeuttaa genomin luotettavaa koostamista jopa parhailla algoritmeilla. Esimerkiksi Clone-by-Clone-(CBC)menetelmällä sekvensoidun maissin genomista jopa 85 % muodostuu transposoneista, mikä tekee koko genomin sekvensoinnista lyhyen lukukehyksen menetelmillä lähes mahdotonta (Hamilton & Buell ym. 2012). WGS-menetelmillä sekvensoiduista kasvigenomeista jää helposti puuttumaan vaikeasti sekvensoitavia alueita. Esimerkiksi paralogisten (äskettäin duplikoituneiden) geenien koostaminen on genomin eri osien kaksinkertaistumisten takia äärimmäisen vaikeaa, koska koostamisohjelmat eivät osaa erotella niitä eri osiin genomia

(Hamilton & Buell ym. 2012). Kolmannen sukupolven menetelmät ja esimerkiksi fosmidikirjastot pyrkivät vastaamaan tähän ongelmaan. Fosmidipohjaisissa menetelmissä haluttu DNA-insertti kloonataan fosmidivektoriin, joka siirretään isäntäbakteerin (usein *E.coli*) sisälle. Monistuvat bakteerikloonit muodostavat fosmidikirjaston (fosmid pools, 100-6000 fosmidia/kirjasto), jotka voidaan edelleen sekvensoida pienissä osissa ja koostaa yksitellen (Nystedt ym. 2013). Niihin voidaan liittää jopa 40 Kb inserttejä, joiden koostaminen on varsin helppoa, koska niiden sijainti kromosomeissa tunnetaan. Fosmidien heikkous on muiden CBC-menetelmien tapaan suuri työmäärä ja korkea hinta.

Yksinkertaisia genomeja voidaan helposti koostaa suoraan WGS-menetelmillä, mutta monimutkaiset kasvigenomit vaativat usein eri menetelmien yhdistelyä. Nämä hybridimenetelmät tuottavat usein hyviä tuloksia, koska eri menetelmät tukevat toistensa heikkouksia. Ensimmäinen hybridimenetelmillä koostettu kasvigenomi oli vuonna 2009 kurkun (*Cucumis sativus*) genomi, jossa yhdistettiin Sanger-sekvensoiduista BAC-vektoreista (3,9 kertainen peittävyys) saatu aineisto Illuminalla sekvensoituun WGS-aineistoon (69,3 kertainen peittävyys) (Huang ym. 2009) (Kuva 2, Bolger ym. 2014). Toinen ajankohtaisempi esimerkki hybridimenetelmien käytöstä on loblollymännyn valtavan genomin selvittäminen yhdistelemällä kolmea erilaista kirjastotyyppiä (Neale ym. 2014). Ensimmäinen puhtaasti NGS-menetelmillä selvitetty kasvigenomi oli ahomansikka (*Fragaria vesca*), jolla on suhteellisen yksinkertainen diploidi seitsemän kromosomin genomi (240Mb, 2n=14) (Shualev ym. 2011).

Usein suurten kasvigenomien koostaminen kokonaisina ei ole käytännöllistä tai maksaa liikaa. Edullisempi ja yksinkertaisempi vaihtoehto on sekvensoida pelkkä eksomi (Warr ym. 2015). Eksomilla tarkoitetaan genomin pelkkiä proteiineja koodaavia osia (eksoneita). Koko genomin sekvensoimiseen verrattuna eksomisekvensoinnilla saadaan kohteen eksonit selvitettyä huomattavasti paremmalla peittävyydellä, koska voidaan keskittyä vain pieneen osaan genomia. Pienempi aineisto on myös helpompi analysoida ja prosessoida koneellisesti (Warr ym. 2015). Sen toteuttaminen vaatii kuitenkin DNA-koettimien (probe) valmistamisen etukäteen olemassa olevan referenssigenomin pohjalta. Myös lähilajin referenssigenomia voidaan hyödyntää. Ennen sekvensointivaihetta koettimet ”kaappaavat” eksonit hybridisaation avulla ja loput genomista pestään pois.



Kuva 2. Kasvigenomien sekvensoinnin ja niistä johdettujen sovellusten aikajana (Bolger ym. 2014). Kuvasta näkyy genomin sekvensointivuosi, laji ja käytetty tekniikka. Janan yläpuolella olevat kasvit on sekvensoitu CBC-menetelmillä ja alapuolella WGS-menetelmillä (Sanger, NGS). Hybriditekniikalla sekvensoidut kasvit on merkitty janan molemmille puolille. Ero puhtauden NGS-menetelmien ja hybridimenetelmien välillä on kuitenkin häilyvä, koska monet NGS-projektit ovat vaatineet olemassa olevien Sanger-sekvenssien hyödyntämistä. Kuvaan on merkitty erikoissymboleilla merkittäviä genomitiedosta johdettuja sovelluksia, kuten esimerkiksi perunan kukkimiseen liittyvän geenireitin selvittäminen.

Kasvigenomien tutkimista vaikeuttaa myös se, että monet kasvit ovat ristisiittoisia ja yksittäisen kasvin genomi on varsin heterotsygoottinen. Koostaminen helpottuu, jos genomi on joko haploidissa tai homotsygoottisessa muodossa. Myös heterotsygoottisten yksilöiden sekvensointi on mahdollista, mutta yleensä korkea heterotsygotian aste johtaa vaikeuksiin ja epätarkkuuksiin koostamisessa. Usein heterotsygotian määrää pyritään vähentämään tulosten parantamiseksi. Esimerkiksi viiniköynnöksen (*Vitis vinifera*) genomi saatiin selvitettyä tuottamalla tietoisesti todella sisäsiittoisia sukulinjoja (Jaillon ym. 2007). Polyploidiset kasvilajit ovat erityisen haastavia. Tähän mennessä kaikki yritykset polyploidisten kasvien genomien sekvensoimiseksi ja koostamiseksi ovat perustuneet joko ploidian vähentämiseen tai kromosomien fyysiseen erotteluun. Hyvä esimerkki

ploidian vähentämisestä on perunan (*Solanum tuberosum L.*) referenssigenomi. Suurin osa kaupallisista perunalajikkeista on tetraploideja, mutta niiden koostaminen osoittautui niin hankalaksi, että referenssigenomi saatiin vasta erikoisesta kaksinkertaisesti monoploidista lajikkeesta. Kun saatua yksittäistä referenssigenomin haplotyyppiä verrattiin diploidiin perunalajikkeeseen huomattiin, että diploidin perunan kromosomistot erosivat enemmän keskenään (Potato Genome Sequencing Consortium 2011). Myös toisen läheisen lajin apuna käyttäminen voi toimia. Kaupallinen mansikka (*Fragaria* x *ananassa*) on allo-oktaploidi ja sitä voidaan pitää yhtenä monimutkaisimmista viljelyskasveista. Koostamistyön helpottamiseksi diploidi ahomansikka (*Fragaria vesca*) sekvensoitiin usean genomin yhtäaikaisen sekvensoinnin välttämiseksi (Shulaev ym. 2011).

Hyvä esimerkki kromosomien fyysisestä erottelusta on alloheksaploidi leipävehnä (*Triticum aestivum*, genomin koko 17 Gb). Sen lähes kaikki 42 kromosomia koostettiin yli 60-kertaisella peitolla erottelemalla kromosomien käsivarret yksitellen sekvensointia varten. Tutkimuksessa hyödynnettiin vehnän erikoisia kromosomimutanttilinjoja, joissa esimerkiksi tietty kromosomi esiintyi yksinkertaisena (Mayer ym. 2014).

Geenien järjestyksen ja kytkentäryhmän selvittäminen voi olla myös haasteellista. Jos pitkien skaffoldien koostaminen ei ole onnistunut, niin erilliset sekvenssit kannattaa ”ankkuroida” kromosomin geneettiseen karttaan käyttämällä hyväksi tunnettuja geneettisiä markkereita. Esimerkiksi kurkulla tämän lähestymistavan ansiosta 73 % saaduista skaffoldeista saatiin liitettyä geneettiseen karttaan (Huang ym. 2009). Toinen tapa on hyödyntää apuna lajin lähisukulaista ja käyttää niiden välistä synteniaa hyväksi ankkuroidessa saatuja sekvenssejä geneettiseen karttaan. Tätä on käytetty hyväksi muun muassa perunalla, jonka genomin koostamisessa käytettiin hyväksi tomaatin tunnettuja sekvenssejä (Potato Genome Sequencing Consortium 2011).

2.2.1. Koostamisen laadulliset kriteerit ja tyypillisimmät tunnusluvut

Genomin koostamisen laatuun tulee kiinnittää erityistä huomiota. Tutkimusta tehdessä täytyy olla perillä aineiston laatukriteereistä virheellisen lähdeaineiston välttämiseksi. Omien tulosten laadullinen arviointi helpottaa johtopäätöksien tekemistä.

Yleisimmin käytettyjä tilastollisia laadunarvioimismenetelmiä ovat kontikien ja skaffoldien N50/L50 – tunnusluvut, aukkojen prosentuaalinen osuus aineistosta, sekä koostamisen kokonaispeittävyys (Yandell & Ence 2012). Kontikien N50-tunnusluku lasketaan asettamalla ensin

kaikki määritellyt kriteerit täyttävät kontikit pituusjärjestykseen pisimmästä lyhimpään. Sitten pisimmästä alkaen kontikien pituuksia aletaan laskea yhteen, kunnes yhteissumma vastaa puolta kaikkien kontikien yhteispituudesta. N50 on tämän kumulatiivisen luettelon lyhimmän kontikin pituus. N50-tunnusluku saadaan skaffoldeille samalla periaatteella. Mitä suurempi N50, sitä paremmin koostaminen on onnistunut. On kuitenkin tärkeää huomata, että huonosti tehty koostaminen, jossa sekvenssit on vain pakotettu sopimaan yhteen voi saada hämäävän suuren N50-arvon (Yandell & Ence 2012). Näiden virheellisten tulosten tunnistaminen on laadukkaan tutkimuksen edellytys. Kontikien laatukriteerit voivat vaihdella paljon eri projektien välillä. Kontikien ja skaffoldien koostamisessa liian lyhyet lukemat karsitaan yleensä pois samoin kuin yksittäiset havainnot ilman muun aineiston tukea (Yandell & Ence 2012). L50-tunnusluku määräytyy samalla periaatteella kuin N50-tunnusluku, mutta yhden havainnon pituuden sijasta mittana käytetään puolivälin saavuttamiseen tarvittujen kontikien tai skaffoldien lukumäärää.

Prosentuaalinen aukkojen osuus kontikien ja skaffoldien välissä on toinen merkittävä koostamisen laatua kuvaava tunnusluku. Sekvensointiteho laskee pitkien lukemien loppupuolella ja osa alueista on niin toistojaksojen kyllästämää, että kontikien ja skaffoldien väliin jää yleensä jonkin verran sekvensoimattomia emäksiä, jotka merkitään usein N-merkinnällä. DNA voi myös olla huonolaatuista (esimerkiksi muinainen DNA), mikä lisää sekvensoimattomien alueiden osuutta aineistosta. Yleensä aukkoja on sitä vähemmän, mitä pitempiä lukemia koostamisessa on käytetty. Joissain tutkimuksissa kaikki aukot ekstrapoloidaan viidenkymmenen N:n mittaisiksi (Yandell & Ence 2012). Pyrkimyksenä on saada mahdollisimman pieni aukkojen osuus koko aineistosta. Kahden koostamisen vertailussa aineistojen N50-arvot voivat olla lähellä toisiaan, mutta prosentuaalinen aukkojen osuus voi vaihdella huomattavasti ja paljastaa näin laadukkaamman referenssigenomin.

Kolmas laatua kuvaava tunnusluku on kokonaispeittävyys. Sitä voidaan ajatella joko koko genomia koskevana tai pelkkiä geenejä koskevana. Genomin kokonaispeitolla tarkoitetaan koostamisen lopputuloksena saatua tunnettua sekvenssiä suhteessa koko genomin muilla menetelmillä (esim. virtaussytometrialla) arvioituun kokoon. Geenipeitolla tarkoitetaan sekvensoitujen geenien osuutta koko genomin geeneistä. Useimmiten genomissa on ainakin jonkin verran vaikeasti sekvensoitavia runsaasti toistojaksoja sisältäviä alueita, joten ei ole mitenkään hälyttävää, jos koostamisen kokonaispeitto jää 90–95 % tietämille. Tällaista tulosta voi pitää hyvänä. Vaikeasti

sekvensoitavilla alueilla on yleensä vähän geenejä, joten geenipeittävyys saadaan yleensä kokonaispeittävyyttä paremmaksi (Yandell & Ence 2012).

Usein koostetun referenssigenomin laatua testataan lopuksi vertaamalla siihen satunnaisia bakteerivektoreihin insertoituja saman genotyypin sekvenssejä, jolloin tarkoituksena on saada mahdollisimman suuri samankaltaisuus. Näin saadaan selville onko koostettu genomi konsistenssi itsensä kanssa. Toinen yleinen tapa on verrata koostetusta genomista löytyviä hyvin konservoituneita geenejä olemassa oleviin tietokantoihin ja katsoa, kuinka hyvin ne on saatu koostettua (Parra ym. 2007).

3. Havupuiden genomien piirteitä

Havupuut ovat tärkeä lajiryhmä mitattuna millä mittarilla tahansa. Ne ovat hallinneet monia maaekosysteemejä etenkin pohjoisella pallonpuoliskolla miljoonia vuosia ja niiden taloudellinen merkitys metsäteollisuudelle on suuri. Havupuu soveltuu muun muassa rakennusmateriaaliksi ja energianlähteeksi. Koko maapallon biomassasta ne muodostavat jopa yli 80 % (Neale & Kremer 2011). Pelkästään loblollymäntyä istutetaan vuosittain yli 1,5 miljardia geneettisesti paranneltua tainta (Neale ym. 2014).

Havupuut kuuluvat paljassiemenisiin kasveihin. Ne ovat ristisiittoisia ja lisääntyvät yleensä ilmassa leviävän siitepölyn avulla. Tästä seuraa suuri efektiivinen populaatiokoko ja korkea heterotsygotian aste. Nukleotidikorvautumiset tapahtuvat kuitenkin koppisiemenisiä lajeja hitaammin, minkä selityksenä voi olla pitkä elinikä (kymmenistä satoihin vuosiin) (Nystedt ym. 2013). Tutkimuksen kannalta kaikista tärkein havupuiden ominaisuus on niiden siemenen megagametofyytin haploidi solukko, joka toimii kehittyvän alkion ravintona. Haploidin solukon sekvensointi ja koostaminen on diploidia solukkoa yksinkertaisempaa, koska ei tarvitse huolehtia vastinkromosomien välisistä eroista. Diploidin solukon kohdalla jo pieni ero kromosomien välillä voi aiheuttaa kahden erillisen kontikin muodostamisen, joiden erottelu esimerkiksi kaksinkertaistuneista sekvensseistä on lähes mahdotonta (Zimin ym. 2014). Vain kolmen havupuulajin genomisekvenssit on julkaistu (metsäkuusi, valkokuusi ja loblollymänty). Suuri genomin koko, pitkät intronit ja toistuvan sekvenssin suuri osuus genomista ovat havupuiden genomien tärkeimpiä erityispiirteistä.

Kattavan referenssigenomin luominen näille hankalasti koostettaville lajeille mahdollistaa niiden ominaisuuksiin vaikuttavien geenien tutkimisen. Erityisen kiinnostavia ominaisuuksia ovat muun

muassa kyky kestää tauteja, tuholaisia ja stressiä, sekä puun muodostuminen. Myös lajien välinen vertailu on kiinnostavaa. Esimerkiksi metsäkuusen sekvenssistä on löytynyt erittäin fragmentoituneina koppisiemenisillä yleisiä FT-geenejä (FLOWERING LOCUS T) (Nystedt ym. 2013). Kiinnostavana uutena havaintona Li ym. 2015 esittivät, että havupuiden ja muiden paljassiemenisten kasvien historiassa on tapahtunut merkittäviä niiden evoluutioon vaikuttaneita WGD-tapahtumia, joita aikaisemmat tutkimukset eivät olleet onnistuneet huomaamaan.

3.1. Valko- ja metsäkuusi

Valkokuusi (*Picea glauca*) ja metsäkuusi (*Picea abies*) ovat ensimmäiset kokonaan sekvensoidut havupuut (Biroll ym. 2013) (Nystedt ym. 2013). Ne julkaistiin lähes samaan aikaan huhtikuussa 2013. Valkokuusi sekvensoitiin pelkästään NGS-menetelmillä, jolloin sekvensoinnin hinta saatiin todella alhaiseksi. Toisaalta saatu sekvenssi ei ollut yhtä laadukasta kuin hybridi- tai CBC-menetelmillä, koska esimerkiksi sekvenssin liittäminen suoraan fyysiseen karttaan ei onnistu. Metsäkuusella sekvensointi tapahtui monivaiheisen hybridimenetelmän kautta, jossa yhdistettiin neljällä eri menetelmällä saatua aineistoa: WGS-menetelmällä sekvensoidut haploidista ja diploidista solukosta saadut kirjastot, diploidista solukosta tehty fosmidikirjasto, sekä kaikista solun RNA-molekyyleistä muodostettu RNAseq-kirjasto (Nystedt ym. 2013). Koostamisen lopputuloksena noin 63 % proteiineja koodaavista geeneistä saatiin koostettua hyvällä yli 90 % peittävyydellä ja 96 % osittaisella yli 30 % peittävyydellä. Suuren genomin haasteellisuudesta kertoo kuitenkin se, että vain noin 25 % genomista saatiin koostettua yli 10 Kb skaffoldeihin. Metsäkuusen genomin yksi erityispiirre on useiden pitkien intronien (pisin jopa noin 68 Kb) löytyminen genomista. Noin joka kymmenennellä geenillä on yli 5 Kb introneita (Nystedt ym. 2013).

3.2. Loblollymänty

Loblollymännyn (*Pinus taeda*) genomi on tähän mennessä suurin sekvensoitu ja koostettu genomi (koko noin 20,1 Gb). Sekvensointi tapahtui uniikin WGS-hybridimenetelmän avulla, joka yhdisteli parhaat puolet eri menetelmistä (Neale ym. 2014). Koostamisessa käytettiin apuna kolmea erilaista kirjastotyppiä: haploidista megagametofyytistä saatua 200-700 bp:n pair-end kirjastoa, diploidista neulassolukosta tehtyä 1-5,5 Kb mate-pair kirjastoa ja 35-40 Kb DiTag fosmidikirjastoa. Megagametofyytin haploidi solukko antaa hyvät edellytykset koostamisen perustaksi, mutta siitä eristetyn haploidin DNA:n määrä ei kuitenkaan riitä laadukkaaseen koostamiseen. Neulasien diploidi solukko tarjosi tarvittavan määrän lähtö-DNA:ta koostamisen viemiseksi loppuun hyvällä

peittävyydellä ja fosmidikirjastot auttoivat genomin fyysisen kartan muodostamisessa (Neale ym. 2014). Lopputuloksena saatu referenssigenomi on yksi korkealaatuisimmista suurista genomeista, joita on onnistuttu sekvensoimaan ja koostamaan tähän päivään mennessä. Skaffoldien N50-arvoksi saatiin kahden koostamisvaiheen jälkeen jopa 66,9 Kb ja CEGMA:n (Core Eukaryotic Genes Mapping Approach) universaaleista merkkigeeneistä 203 löytyi referenssigenomista täysmittaisena. CEGMA on kokoelma geenejä, jotka ovat vahvasti konservoituneita lähes kaikilla taksoneilla ja joita voidaan täten käyttää koostamisen laadun testaamiseen (Parra ym. 2007). Täysimittaisten geenien osuus vastasi jopa 91 % kaikista löydetyistä merkkigeeneistä (Zimin ym. 2014).

Loblollymännyn genomista toistuvat elementit muodostavat jopa 82 %. Suurin osa toistuvista elementeistä on kuitenkin onneksi muinaisia, joten koostamisalgoritmit osaavat erotella ne kerääntyneiden erojen perusteella melko hyvin erillisiksi kopioiksi (Zimin ym. 2014). Kiinnostavana havaintona introneista jopa 60 % muodostuu toistuvista elementeistä. Havupuilla on myös erityisen vähän mikrosatelliitteja koppisiemenisiin kasveihin verrattuna, joista poikkeavan suuri määrä on heptanukleotideja (seitsemän nukleotidin mittaisia). Männyillä on todella pitkät telomeerit (jopa 57 kb, *Pinuslongaeva*), mikä yhdessä sentromeerien kanssa todennäköisesti selittää heptanukleotidien suuren määrän (Neale ym. 2014).

4. Tulevaisuudennäkymiä

Suurigenomisista kasviryhmistä etenkin viljakasvien merkitys tulee kasvamaan, koska ne tuottavat suurimman osan maapallon syötävästä biomassasta. Vehnän viljely aloitettiin jo noin 10000 vuotta sitten ja se on pääsääntöinen ruuan lähde noin kolmannekselle maapallon väestöä (Mayer ym. 2014). Jatkuva väestönkasvu asettaa jalostajille haasteita muun muassa vehnän satoisuuden lisäämiseksi jopa 70 prosentilla vuoteen 2050 mennessä (Mayer ym. 2014). Voimistuva ilmastonmuutos voi heikentää satoisuutta ja lisätä vaihtelevia sääoloja useilla alueilla luoden tarpeen erilaisia sääoloja kestäville lajikkeille. Kattavat tiedot vehnän genomista ovat ehdoton edellytys tehokkaalle jalostukselle ja perinteisestä jalostuksesta tulisi vähitellen siirtyä tarkkaan genomitietoon pohjautuvaan jalostukseen. Viimeisimmät tarkat tiedot vehnän genomista mahdollistavat erilaisten mutaatioiden tehokkaan tunnistamisen, koska ne voidaan havaita ilman suurta vaikutusta kasvin fenotyyppiin. Tämä tulee todennäköisesti mahdollistamaan uusien erilaisten kasviyksilöiden käytön jalostuksessa ja avaamaan näin koko polyploidin vehnän

genomisen potentiaalin (Borrill ym. 2015). Genomitiedon käyttö jalostuksessa ja tutkimuksessa tulee todennäköisesti olemaan kiivaan keskustelun aihe, koska alalla liikkuu paljon rahaa ja investointeja. Yksityisten genomitietoa keräävien firmojen on päätettävä antavatko ne tietojaan julkiseen käyttöön, vai pitävätkö ne omana tietonaan.

Uudetkin menetelmät vanhenevat nopeasti ja eri menetelmien yhdistely tulee yleistymään. Tämä helpottaa merkittävästi suurten kasvigenomien sekvensointia ja koostamista, kun pidemmät lukemat ovat mahdollisia ja genomi saadaan jaettua pienempiin osiin sekvensointia varten (esim. 10x Genomics ja kromosomien fyysinen erottelu). Havupuiden jalostusta voidaan tehostaa yksilöiden genomisen jalostusarvon hyödyntämisellä fenotyypin sijaan, sillä fenotyypin selvittäminen on vaivalloista esimerkiksi puun laadun ja monimutkaisten stressinsieto-ominaisuuksien selvittämisessä (De La Torre ym. 2014). Uusi havainto havupuiden WGD-tapahtumista tulee todennäköisesti muuttamaan havupuiden evoluution tarkastelua ja tutkimusta (Li ym. 2015). Bioinformatiikan merkitys laadukkaalle koostamiselle ja aineistojen tulkinnalle kasvaa jatkuvasti ja sillä on vaikeuksia pysyä nopeasti kehittyvien menetelmien perässä. Myös aineistojen säilytykseen vaadittavia tallennustiloja ja eri tietokantojen välistä yhteistyötä täytyy lisätä tulevaisuudessa. Emme suinkaan ole kehityksen loppumetreillä, vaan pikemminkin todellinen genominen vallankumous on vasta alkamassa.

5. Kirjallisuusluettelo

Ahuja M & Neale D (2005) Evolution of genome size in conifers. Silvae Genet 54(3): 126-137.

Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, McVean GA ym. & 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319): 1061-1073.

Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Saint Yuen MM, Keeling CI, Brand D, Vandervalk BP, Kirk H, Pandoh P, Moore RA, Zhao Y, Mungall AJ, Jaquish B, Yanchuk A, Ritland C, Boyle B, Bousquet J, Ritland K, MacKay J, Bohlmann J & Jones SJM (2013) Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics 29(12): 1492-1497.

Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B & Mayer KFX (2014) Plant genome sequencing - applications for crop improvement. Curr Opin Biotechnol 26: 31-37.

Borrill P, Adamski N & Uauy C (2015) Genomics as the key to unlocking the polyploid potential of wheat. New Phytol 208(4): 1008-1022.

Chang Y, Land M, Hauser L, Chertkov O, Del Rio TG, Nolan M, Copeland A, Tice H, Cheng J, Lucas S, Han C, Goodwin L, Pitluck S, Ivanova N, Ovchinikova G, Pati A, Chen A, Palaniappan K, Mavromatis K, Liolios K, Brettin T, Fiebig A, Rohde M, Abt B, Goeker M, Detter JC, Woyke T, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk H & Lapidus A (2011) Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium Ktedonobacter racemifer type strain (SOSP1-21(T)). Stand Genomic Sci 5(1): 97-111.

Chen ZJ (2010) Molecular mechanisms of polyploidy and hybrid vigor. Trends Plant Sci 15(2): 57-71.

De La Torre AR, Birol I, Bousquet J, Ingvarsson PK, Jansson S, Jones SJM, Keeling CI, MacKay J, Nilsson O, Ritland K, Street N, Yanchuk A, Zerbe P & Bohlmann J (2014) Insights into Conifer Giga-Genomes. Plant Physiol 166(4): 1724-1732.

Dolezel J, Bartos J, Voglmayr H & Greilhuber J (2003) Nuclear DNA content and genome size of trout and human. Cytom Part A 51A(2): 127-128.

Fawcett JA, Maere S & Van de Peer Y (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci U S A 106(14): 5737-5742.

Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC & Vezenov DV (2009) The challenges of sequencing by synthesis. Nat Biotechnol 27(11): 1013-1023.

Greilhuber J, Dolezel J, Lysak M & Bennett M (2005) The origin, evolution and proposed stabilization of the terms "genome size' and 'C-value' to describe nuclear DNA contents. Ann Bot 95(1): 255-260.

Hamilton JP & Buell CR (2012) Advances in plant genome sequencing. Plant J 70(1): 177-190.

Henson J, Tischler G & Ning Z (2012) Next-generation sequencing and large genome assemblies. Pharmacogenomics 13(8): 901-915.

Heslop-Harrison JS(& Schwarzacher T (2011) Organisation of the plant genome in chromosomes. Plant J 66(1): 18-33.

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Du Y, Li S ym. (2009) The genome of the cucumber, Cucumis sativus L. Nat Genet 41(12): 1275-U29.

Jaillon O, Aury J, Noel B, Policriti A, Clepet C ym. & French-Italian Public (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449(7161): 463-U5.

Kaul S, Koo H, Jenkins J, Rizzo M, Rooney T, Tallon L, Theologis A, Dangl J, Jones J, Chen M, Chory J, Somerville M ym. & Ar Gen In (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408(6814): 796-815.

Lander E, Int Human Genome Sequencing Consortium, Linton L, Birren B, Wetterstrand K, Patrinos A, Morgan M ym. & Int Human Genome Sequencing Conso (2001) Initial sequencing and analysis of the human genome. Nature 409(6822): 860-921.

Lee S & Kim N (2014) Transposable elements and genome size variations in plants. Genomics & informatics 12(3): 87-97.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH & Barker MS (2015) Early genome duplications in conifers and other seed plants. Science advances 1(10): e1501084.

Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, Jakobsen KS, Wulff BBH, Steuernagel B, Mayer KFX, Olsen O & Int Wheat Genome Sequencing (2014) Ancient hybridizations among the ancestral genomes of bread wheat. Science 345(6194): 1250092.

Mayer KFX, Rogers J, Dolezel J, Pozniak C, Eversole K, Feuillet C, Praud S ym. & IWGSC (2014) A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345(6194): 1251788.

Neale DB & Kremer A (2011) Forest tree genomics: growing resources and applications. Nat Rev Genet 12(2): 111-122.

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, Martinez-Garcia PJ, Vasquez-Gross HA, Lin BY, Zieve JJ, Dougherty WM, Fuentes-Soriano S, Wu L, Gilbert D, Marcais G, Roberts M, Holt C, Yandell M, Davis JM, Smith KE, Dean JFD, Lorenz WW, Whetten RW, Sederoff R, Wheeler N, McGuire PE, Main D, Loopstra CA, Mockaitis K, deJong PJ, Yorke JA, Salzberg SL & Langley CH (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol 15(3): R59.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hallman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Kaller M, Luthman J, Lysholm F, Niittyla T, Olson A, Rilakovic N, Ritland C, Rossello JA, Sena J, Svensson T, Talavera-Lopez C, Theissen G, Tuominen H, Vanneste K, Wu Z, Zhang B, Zerbe P, Arvestad L, Bhalerao R, Bohlmann J, Bousquet J, Gil RG, Hvidsten TR, de Jong P, MacKay J, Morgante M, Ritland K, Sundberg B, Thompson SL, Van de Peer Y, Andersson B, Nilsson O, Ingvarsson PK, Lundeberg J & Jansson S (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497(7451): 579-584.

Parra G, Bradnam K & Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genornes. Bioinformatics 23(9): 1061-1067.

Roberts RJ, Carneiro MO & Schatz MC (2013) The advantages of SMRT sequencing. Genome Biol 14(7): 405.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Wing RA, Wilson RK ym. (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science 326(5956): 1112-1115.

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Borodovsky M, Veilleux RE, Folta KM ym. (2011) The genome of woodland strawberry (Fragaria vesca). Nat Genet 43(2): 109-116.

Soltis PS, Marchant DB, Van de Peer Y & Soltis DE (2015) Polyploidy and genome evolution in plants. Curr Opin Genet Dev 35: 119-125.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D ym. (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313(5793): 1596-1604.

van Dijk EL, Auger H, Jaszczyszyn Y & Thermes C (2014) Ten years of next-generation sequencing technology. Trends Genet 30(9): 418-426.

Walsh B (2003) Population-genetic models of the fates of duplicate genes. Genetica 118(2-3): 279-294.

Warr A, Robert C, Hume D, Archibald A, Deeb N & Watson M (2015) Exome Sequencing: Current and Future Perspectives. G3-Genes Genomes Genet 5(8): 1543-1550.

Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, van Ham RCHJ, Visser RGF ym. & Potato Genome Sequencing Consortiu (2011) Genome sequence and analysis of the tuber crop potato. Nature 475(7355): 189-U94.

Yandell M & Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13(5): 329-342.

Zimin A, Stevens KA, Crepeau M, Holtz-Morris A, Koriabine M, Marcais G, Puiu D, Roberts M, Wegrzyn JL, de Jong PJ, Neale DB, Salzberg SL, Yorke JA & Langley CH (2014) Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. Genetics 196(3): 875-890.

10x Genomics kotisivut [http://www.10xgenomics.com/technology/] [Viitattu 7.4.2016]