De novo sequencing, annotation, and characterization of the genome of *Lavandula angustifolia* (Lavender)

Radesh Nattamai Malli Pooranachandhiran M.Sc. M.Phil.

Centre for Biotechnology

Submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Faculty of Mathematics and Science, Brock University

St. Catharines, Ontario

Abstract

Lavender (*Lavandula angustifolia*) is a perennial plant native to the Mediterranean region, best known for its essential oil (EOs) that have numerous applications in the pharmaceutical, cosmetic and perfume industries. We performed sequencing of the *L. angustifolia* genome and report a detailed analysis of the assembled genome, focusing on genome size, ploidy, and repeat content. The lavender genome was estimated to be around 870 Mbp (1C=0.96 pg) using a quantitative PCR method. Genome size was further validated through analysis of raw genome sequences using Kmergenie, providing a conclusive end to the lavender genome size dispute. The repeat element composition of the genome was analyzed using *de novo* (RepeatModeler) and library-based methods (RepeatMasker) and was estimated to be around 45% of the full genome or ~57% of the non-gap genome sequences. Further characterization revealed Long Terminal Repeat (LTRs) retrotransposons as the major repeat type, which contribute to ~18% of the genome, followed by DNA transposons at ~8.5% of the genome. Interestingly, unlike most other plant genomes, the lavender genome has many more *Copia* than *Gypsy* elements, both showing a trend of recent increasing activity. Furthermore, these LTRs, especially *Copia* elements, have shown active participation in gene function including genes for essential oil production, with *Copia* elements contributing to ~30 % of the coding DNA sequence (CDS) regions, in addition to promoter, intron and untranslated (UTR) regions. The lavender genome also has an unusually high number of miniature inverted-repeat transposable elements (MITEs) compared to other model plant genomes, with the number being ~88,000, which is close to that (~90,000) of the much larger maize genome. Analysis also revealed the lavender genome with a high proportion at polyploidy level, which is strongly biased towards regions containing essential

oil genes, with polyploidization events in the lavender genome occurred between 16 to 41 Mya. In conclusion, our results reveal the lavender genome to be highly duplicated and with past and ongoing active retrotransposition, making the genome optimized for EO production.

Keywords: *Lavandula angustifolia*, Illumina sequencing, genome size estimation, *de novo* genome assembly and annotation, essential oil pathways.

Acknowledgments

I am very grateful to my supervisor, Dr. Ping Liang, for his endless support and guidance throughout my time in his lab. I will always be appreciative of his patience and encouragement to teach me many fundamental things regarding this subject. I would also like to thank my co-supervisor Dr. Soheil Mahmoud from UBC for his support, encouragement and valuable feedback for my project and thesis. I would also like to thank my committee members Dr. Heather Gordon and Dr. Vincenzo Deluca for their insights and suggestions. I am very appreciative of all my colleagues in the lab, including Daniel W. Tang, Calvin Sjaarda, Jina Nanyakkara, Zakia Dahi and Ilona Hilson for always being there for me when I needed. Finally, I would like to thank with gratitude, the love and unconditional support from my wife, my lovely children, family and friends.

Contents

List of Figures

List of Tables

List of Abbreviations

1-deoxyxylulose-5-phosphate synthase (DXS)

4-hydroxy-3-methylbut-2-enyl diphosphate (HMBPP)

4-hydroxy-3-methylbut-2-enyl diphosphate reductase (HDR)

Base pair (bp)

Basic Local alignment search tool (BLAST)

Benchmarking Universal Single Copy Orthologues (BUSCO)

BLAST-like alignment tool (BLAT)

Coding DNA Sequence (CDS)

Dimethylallyl diphosphate (DMAPP)

Essential oils (EOs)

Farnesyl pyrophosphate synthase (FPPS)

Flow cytometry (FCM)

Geranylgeranyl pyrophosphate synthase (GGPPS)

GFF: General Feature Format

HMBPP reductase (HDR)

HMBPP synthase (HDS)

IPP isomerase (IPPI)

Isopentenyl diphosphate (IPP)

Long Terminal Repeat (LTR)

Messenger Ribonucleic acid (mRNA)

Methylerythritol Pathway (MEP)

Mevalonate Pathway (MVA)

Next-generation sequencing (NGS)

Optimal Paired-End Read Assembler (OPERA)

Polymerase chain reaction (PCR)

Quantitative Polymerase chain reaction (qPCR)

Terpene synthases (TPS)

Transposable Elements (TEs)

Whole genome sequencing (WGS)

Chapter I Introduction

I.1 The lavender plant and its essential oils

Lavenders belong to the *Labiatae* (*Lamiaceae*) family of plants, which originated in and around the Mediterranean region. In recent times, lavender plants have become prevalent around the world in diverse regions like Bulgaria, United States of America, Australia, China and India due to their commercial value as sources of essential oils (EOs) and as ornamental plants (Giray, 2018). The most common type of lavender is called the English lavender known scientifically as *Lavandula angustifolia* or *Lavandula officinalis* as called previously (Upson and Andrews, 2005). In Latin, the species name *angustifolia* refers to the narrow-leaves of the plant and *officinalis* refers to its medicinal properties (Denner, 2009). Lavenders are long-living perennial plants with a life span of about 10 years. They can adapt to extreme temperature fluctuations, are resistant to frost, drought, common plant pathogens. The lavenders grow in most climatic conditions, but grow best in well-drained soils with full sunlight (Wells et al., 2018). A typical lavender plant grows to a height of 40-80 cm and produces short-stemmed flowers at the top of spikes or stems. The EO yield of *L.angustifolia* per plant is lower when compared to other lavender species like *L. x intermedia*, but higher in EO quality (Kara and Baydar, 2013). The cultivation of most lavender species can be done by seed propagation or by clonal reproduction, although some species (e.g., *L. x intermedia*) do not produce seed and must be vegetatively propagated. Clonal reproduction provides plant uniformity for large-scale cultivation to generate new individuals without altered genotype, whereas seed-based propagation requires a lot of time for germination. Some seeds may have low germination capacity and are not suitable for

cultivation of sterile and hybrid species, e.g., *L. x intermedia* which is a hybrid between *L. angustifolia* and *L. latifolia* (Upson and Andrews, 2005).

I.2 Lavender EOs: chemical composition and metabolic pathways

With an estimated production rate of over 1500 metric tons / annum (mostly extracted from *L. angustifolia* and *L. x intermedia*), lavender EOs significantly contribute to the multi-billion-dollar flavor and fragrance industry worldwide (Giray, 2018). The quality of lavender oils and hence the market price is determined by the chemical composition. Various factors like climatic condition, growing method, harvesting, transport and storage determine the quality of EOs (Giray, 2018).

I.2.1 The chemical composition and applications of EOs

Lavender EOs are comprised mainly of monoterpenes (Białon et al., 2019) with smaller amounts of a few sesquiterpenes (Sarker et al., 2013; Mendoza-Poudereux et al., 2017). The key constituents of lavender EOs are linalool, linalyl acetate, 1,8-cineole, β-ocimene, terpinen-4-ol, and camphor with their relative levels varying among species. EOs obtained from *L. angustifolia* by steam distillation are chiefly composed of linalyl acetate (3,7-dimethyl-1,6-octadien-3-yl acetate) (51%), linalool (3,7-dimethylocta-1,6-dien-3-ol) (35%) (Prashar et al., 2004), and lower concentrations of other components, namely lavandulol, 1,8-cineole, lavandulyl acetate, and camphor (Koulivand et al., 2013). Among the EO constituents, the monoterpene camphor has a large effect on the EO quality. The greater the amount of camphor the lower the quality. High quality EOs have 0.5-1% camphor whereas EOs with 5-10% camphor are unsuitable for many applications including incorporation into foods. Other minor components

can also significantly affect oil quality. EOs are generally extracted by steam distillation of flower heads, although other methods, such as solvent extraction and extraction in liquid CO_2, are also used. Because many high-yield lavender species (e.g., *L. x intermedia*) produce high quantities of undesired constituents (such as camphor), there is a great interest in industry for enhancing EO yield and quality in the lavender (Wells et al., 2018).

Some monoterpene constituents of lavender EOs also have bioactive properties. For example, linalyl acetate and linalool have sedative and local anesthetic effects, whereas 1,8-cineole can be used as a spasmolytic, local anesthetic and antibacterial agent. Camphor, α-terpineol, terpenen-4-ol, α- and β-pinene, and *p*-cymene have strong antimicrobial activities (Woronuk et al., 2011). Furthermore, the anticancer and antimutagenic properties of some lavender EO constituents, like perillyl alcohols, show exciting therapeutic prospects (Woronuk et al., 2011). For instance, perillyl alcohol and terpinen-4-ol have been shown to have strong antimutagenic properties against TA98 bacterial cells, and development of topological ointments from lavender oils has been suggested as a skin cancer prevention strategy (Evandri et al., 2005; Dalilan et al., 2013). In addition, multiple examples of lavender EOs, specifically the linalool component, have been shown to impact the nervous system of rats and the rate of metabolism and body weight (Koulivand et al., 2013).

In addition to their applications as cosmetics, fragrances, and medicines, lavender EOs have potential uses as bioinsecticides in controlling different crop- and plant-destroying insects. For example, linalool from lavender plants was found to be toxic against a plant parasite, *Plutella xylostella*, one of the most serious pests of cruciferous crops (Yi et al., 2016). The low risk to environment and humans provides the applications of these plant secondary bioactive compounds a potentially large market.

I.2.2 Essential oil metabolism in lavender

The active ingredients of lavender EOs include mono- and sesquiterpenes, the C_{10} and C_{15} class of terpenoids and isoprenoids, respectively. Like other terpenes, mono- and sesquiterpenes are derived from isopentenyl pyrophosphate (IPP) and dimethyl allyl phosphate (DMAPP) in three stages, including: i) generation of IPP/DMAPP, ii) condensation of IPP/DMAPP to form geranyl pyrophosphate GPP (linear precursor to monoterpenes) and farnesyl pyrophosphate FPP (linear precursor to sesquiterpenes), and iii) conversion of GPP and FPP to mono- and sesquiterpenes, respectively, by terpene synthase enzymes.

I.2.2.1 Generation of IPP/DMAPP

The biosynthesis of IPP and DMAPP occurs via two spatially separated pathways: 1) the mevalonate (MVA) pathway located in the cytosol, and 2) the deoxy-xylulose phosphate (DXP)/2-*C*-methyl-D-erythritol 4-phosphate (MEP) pathway located in the plastid (Chatzivasileiou et al., 2019). The IPP/DMAPP required for monoterpene production are generated *via* the MEP pathway, while the IPP/DMAPP for sesquiterpene biosynthesis are generated by the MVA pathway. As shown in Figure 1, in the MEP pathway, seven enzymes are involved in the formation of IPP/DMAPP from pyruvate and glyceraldehyde-3-P. The IPP/DMAPP pools are also synthesized *via* the cytosolic MVA pathway using six enzymes. For the MVA pathway, acetoacetyl-CoA (AAC) is initially formed from acetyl CoA by AAC thiolase and is subsequently through the MVA pathway converted to mevalonate-5-diphosphate (MVAPP), which is the last step to produce IPP by MVA decarboxylase (MVD). IPP is then isomerized to DMAPP by IPP Isomerase (IPPI). Although the MEP and MVA pathways are

independent, there is evidence indicating that they interact with one another through exchange of the IPP/DMAPP pool, although the trafficking mechanisms for the process are still unknown. Various studies suggest that DXS and HMGR are rate-limiting enzymes in the MEP and MVA pathways, respectively (Zhang et al., 2018).

I.2.2.2 Generation of GPP and FPP from IPP/DMAPP

The condensation of IPP with DMAPP to synthesize geranyl diphosphate (GPP, C_{10}) and farnesyl diphosphate (FPP, C_{15}) is catalyzed by GPP synthase (GPPS) and FPP synthase (FPPS), respectively (Figure 1) that have been described in numerous plants, including lavenders (Sanmiya et al., 1997; Chang et al., 2010; Closa et al., 2010; Chen et al., 2015; Richter et al., 2015; Adal and Mahmoud, 2020). In addition, other prenyltransferase enzymes involved in the production of higher order terpenoids have also been described in many plants (Akhtar et al., 2013; Wang et al., 2016; Kopcsayová and Vranová, 2019; Kusano et al., 2019; Zhou and Pichersky, 2020).

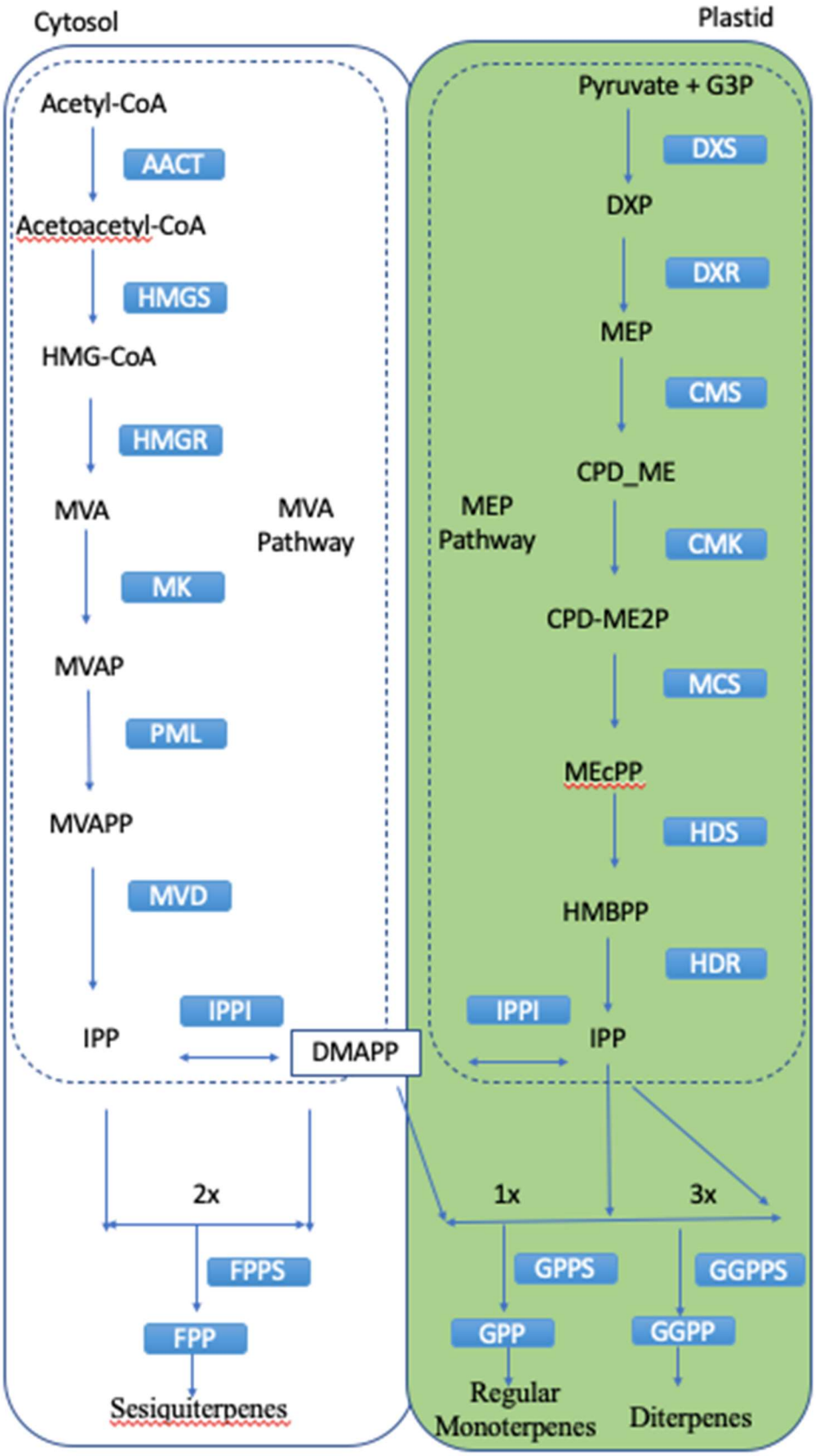

Cytosol
Plastid
Acetyl-CoA
AACT
Acetoacetyl-CoA
HMGS
HMG-CoA
HMGR
MVA
MVA Pathway
MK
MVAP
PML
MVAPP
MVD
IPP
IPPI
DMAPP
2x
FPPS
FPP
Sesiquiterpenes
Pyruvate + G3P
DXS
DXP
DXR
MEP
CMS
CPD_ME
MEP Pathway
CMK
CPD-ME2P
MCS
MEcPP
HDS
HMBPP
HDR
IPPI
IPP
1x
3x
GPPS
GGPPS
GPP
GGPP
Regular Monoterpenes
Diterpenes

Figure 1. A schematic diagram of the terpene biosynthetic pathways.

Biosynthesis of IPP and DMAPP through the plastidial-localized MEP pathway (right) and the MVA pathway (left) in cytosol. The intermediate products and enzymes in the MEP pathway include *Dxp*, 1-deoxyxylulose-5-phosphate; *Dxs*, *Dxp* synthase; *Dxr*, *Dxp* reductoisomerase; *Mep*, 2-C-methylerythritol-4-phosphate; *Cdp-Me*, 4-diphosphocytidyl-2-C-methylerythritol; *Mct*, *Cdp-Me* synthase; *Cdp-Me2p*, 4-diphosphocytidyl-2-C-methyl-D-erythritol 2-phosphate; *Cmk*, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; *MEcPP*, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate; *Mds*, *MEcPP* synthase; *Hmbpp*, 4-hydroxy-3-methylbut-2-enyl diphosphate; *Hds*, *Hmbpp* synthase; *Hdr*, *Hmbpp* reductase; *Ipp*, Isopentenyl diphosphate; *Ippi*, *Ipp* isomerase; *Dmapp*, Dimethylallyl diphosphate. The MVA pathway products and enzymes include *Hmg*, 3-hydroxy-3-methylglutaryl-CoA; *Hmgs*, *Hmg* synthase; *Hmg*, 3-hydroxy-3-methylglutaryl-CoA; *Hmgr*, *Hmg* reductase; *Mva*, mevalonate; *Mk*, *Mva* kinase; *Mvap*, phosphomevalonate; *Pmk*, *Mvap* kinase. 1X, 2X and 3X refer to the number of *Ipp* units which fuse with *Dmapp* to form monoterpenes, sesquiterpenes and diterpenes. Both mono and diterpenes are formed in the plastid, whereas the sesquiterpenes are synthesized in the cytosol. The solid arrows indicate downstream processes, and the dotted lines mark the boundaries of the cytosol and the plastid.

I.2.2.3 Terpene synthases (TPSs)

TPSs are a diverse class of enzymes which catalyze the biosynthesis of specific terpenes from linear precursors. For example, monoterpenes (C_{10}), sesquiterpenes (C_{15}), and diterpenes (C_{20}) are derived from their corresponding linear precursors GPP, FPP, and GGPP through the catalytic activities of monoterpene synthases, sesquiterpenes synthases, and diterpene synthases, respectively.

Plant TPSs are characterized by conserved structural features, such as DDxxD, NSE/DTE, and R(R)X8W motifs (Zhou and Pichersky, 2020). TPS enzymes fall into two classes, type I and type II, based on structure and catalytic mechanisms. Type I TPSs contain the aspartate-rich DDxxD/E motif at their C-terminal domain, called 'α domain' that binds the metal cofactor (Mg^{2+} or Mn^{2+}) for interacting with the prenyl diphosphate substrates. Type II TPSs contain a DxDD motif in the 'β domain' near the N-terminus, with the second aspartate essential for the protonation-initiated cyclisation of GGPP to form copalyl diphosphate (CPP) or other cyclic diterpene diphosphates (Zerbe and Bohlmann, 2015). The single functional TPS gene in *Physcomitrella patens* is bifunctional, encoding an enzyme with both type I and type II active domains (Hayashi et al., 2006).

The TPS family is phylogenetically classified into seven subfamilies, as TPS-a to h. TPS-a is an angiosperm-specific clade that typically contains the largest number of TPS genes in a plant genome, and enzymes in this clade usually are sesquiterpene or diterpene synthases (Chen et al., 2011). The model plant *Arabidopsis* has 32 TPS genes, of which 68% (22) are TPS-a genes, with 18 of these encoding diterpene synthases that are targeted to the chloroplasts and many of the remaining not characterized (Aubourg et al., 2002). The tomato genome has ~44 TPS genes, of which 12 are TPS-a genes (Falara et al., 2011). The TPS-b subfamily is again angiosperm specific and encodes monoterpene synthases with a R(R)X8W motif that catalyzes the isomerization cyclization reaction. TPS-c enzymes catalyze copalyl diphospate synthase and belong to the ancestral clade (Yu et al., 2020). The gymnosperm-specific TPS-d subfamily performs several functions, as diterpene, monoterpene, and sesquiterpene synthases. TPS-e/f subfamilies encode copalyl-diphosphate/kaurene synthases, which are critical enzymes in gibberellin biosynthesis. Another angiosperm-specific TPS-g subfamily encodes monoterpene

synthases without the R(R)X8W motif. An interesting feature of TPS-h subfamily is that it is only observed in *Selaginella moellendorffii* (Yu et al., 2020). Key lavender TPS genes, such as 9-epi-caryophyllene synthase (Sarker et al., 2013), τ-cadinol synthase (Jullien et al., 2014) and β-phellandrene synthase (Landmann et al., 2007), have been identified and characterized.

Many genes are involved in the biosynthesis of lavender EOs, among these, prenyltransferases and several terpene synthases have been characterized from lavenders (Landmann et al., 2007; Lane et al., 2010; Demissie et al., 2012; Sarker et al., 2013). Altogether, at this writing, there are ~72 nucleotide sequences (complete and partial) available in the NCBI GenBank databases from *Lavandula angustifolia* and seven more coding sequences (CDS) for these genes from *Lavandula intermedia* (Adal et al., 2019). Out of the 72 sequences, only 18 were relevant to the EO biosynthesis (Table 1). The remaining 43 sequences (partial sequences) are for genes with other key functions, including 11 for transcription factors. Therefore, many EO genes in lavender are yet to be identified.

Table 1. A list of complete and partial *L. angustifolia* mRNA sequences for EO genes available in GenBank at NCBI.

	L. angustifolia mRNA sequences	Complete/Partial cds
1	trans-alpha-bergamotene synthase	complete
2	linalool synthase	complete
3	limonene synthase	complete
4	subsp. *angustifolia* cultivar Diva bornyl diphosphate synthase	complete
5	cultivar Diva farnesyl diphosphate synthase (*Fpps*)	complete
6	cultivar Diva hydroxymethylglutaryl-CoA reductase (*Hmgr*)	complete
7	cultivar Diva hydroxymethylglutaryl-CoA synthase (*Hmgs*)	complete
8	cultivar Diva 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (*Hdr*)	complete
9	cultivar Diva 4-hydroxy-3-methylbut-2-enyl diphosphate synthase (*Hds*)	complete
10	cultivar Diva 1-deoxy-D-xylulose 5-phosphate reductoisomerase (*Dxr*)	complete
11	cultivar Diva 1-deoxy-D-xylulose 5-phosphate synthase (*Dxs2*)	complete
12	cultivar Diva deoxy-xylulose synthase (*Dxs1*)	complete
13	cultivar Diva germacrene-D synthase	complete
14	cultivar Diva B-caryophyllene synthase	complete
15	cultivar Diva cadinol synthase	complete
16	cultivar Lady 1,8-cineole synthase	complete
17	cultivar Lady monoterpene synthase-like protein (TPS-I)	complete
18	cultivar Lady beta-phellandrene synthase	complete

I.3 Important aspects of plant genomes

The development of genomic resources for non-model plants, which include draft *de novo* genome assembly, annotation and other genetic features has been expedited by the relatively recent advent of next-generation sequencing technologies. These resources provide a wealth of information regarding key features of plant genomes, such as genome size, polyploidy, different classes of repeat elements, gene families including plant defense genes, and many others.

I.3.1 Methods for estimating plant genome size

Plant sizes range dramatically from very small (e.g., *Wolffia globosa* which is 0.1mm in size) to very large (e.g., Eucalyptus which can grow up to 100 m in height). Similar to their physical appearance, apparently although not tied to their physical sizes, plant genomes are inherently complex and highly variable with their sizes ranging from 63 Mbp (*Genlisea aurea*) (Michael, 2014) to more than 20 Gbp (*Pinus tadea*) (Hamilton and Buell, 2014). Knowing the genome sizes across different species can provide important insights regarding the evolution history (e.g., whole genome duplication) and taxonomical relationship of organisms and is useful for designing experimental protocols in genetics and genomics research. As one specific example, having a good estimation of the genome size can provide useful guidance in designing a proper strategy for sequencing the genome.

The estimation of the total amount of DNA in cell nuclei (i.e., the genome size) had been done prior to the start of modern and molecular genetics (Doležel and Bartoš, 2005). In 1950, to avoid the confusion between chromosome number and the total amount of DNA in a single haploid cell, the term, C-value, was coined for denoting the amount of DNA in picogram (pg) (Swift, 1950).

Several different types of methods have been developed for determining the C-value of a species. One of the first-generation methods was reassociation kinetics or Cot analysis first developed by Britten and colleagues (Britten et al., 1974). It is based on the principles of DNA renaturation kinetics: 1) given enough time, all denatured DNA in a sample will be able to renature/reanneal; 2) the rate of renaturation is directly proportional to the copy number of the sequences present in the genome; 3) the more repetitive sequences, the less the time required for renaturation. Overall, the renaturation of the DNA is dependent on DNA concentration,

reassociation temperature, cation concentration, and viscosity (usually not a factor if the DNA is free of contaminants). The analysis of the reassociation kinetics is usually displayed as a plot where the y-axis is the percentage of the single-stranded DNA and the x-axis is a log-scale of the Cot value, which is defined as "Concentration of the DNA (moles/liter) X the time of the renaturation in seconds X buffer factors (counts for the effect of cation)". The Cot profile can be used to understand DNA size and complexity, particularly, the proportion of repetitive sequence in a genome (Peterson et al., 2002).

The reassociation kinetics method is time consuming with the results being not very reliable, and single nuclei counting methods were later introduced. This type of method works by measuring DNA content for individual nuclei based on UV absorption, which is proportional to the DNA content in the nuclei. The method is also called Feulgen densitometry (Hardie et al., 2002). The subsequent development of scanning micro-spectrophotometry provides better accuracy by being able to measure many very small objects and using the average as more accurate results. However, there were some disadvantages in using the above-mentioned methods, foremost of which was the preparation of the suspensions of cells. Plant cells with their rigid cell walls and irregular shapes are not conducive to the above-mentioned methods, and moreover, it is time-consuming and still does not provide very accurate results. Galbraith and co-workers produced a simple method to overcome these difficulties by using small slices of fresh tissues sitting in an isolation buffer (Galbraith et al., 1983). This and other improvements, such as making suspensions of microscopic particles and flow dynamics, made the process of flow cytometry (FCM) successful as the gold standard method in determining genome sizes (Doležel and Bartoš, 2005). A database maintained by the Royal Botanical garden, Kew, UK, last updated in April 2019, as of this writing, provides C-values for 12,273 plants covering angiosperms,

gymnosperms, algae, and bryophytes (Pellicer and Leitch, 2020). Furthermore, the applications of flow cytometry have also led to advancements in estimation of ploidy, cell cycle kinetics and other applications (Doležel and Bartoš, 2005).

To overcome the limitation of flow cytometry for requiring a reference genome, which might not be always available, Wilhelm and co-workers developed a simple real-time PCR method (Wilhelm et al., 2003). The method works with genomic DNA, even with low quality DNA samples if they do not have RNA and UV-absorbing contaminants, and thus is much simpler to perform. The method involves designing a nested PCR for one or two single copy genes. For each single copy gene, a PCR product with the outer primers is obtained and the absolute amount is determined based on UV absorbance. Since the sequence of the PCR product is known, the exact copy number of DNA for a given amount of DNA can be calculated by dividing the amount of DNA in pg with the molecular weight of the PCR fragment in double stranded DNA. A serial dilution of the PCR DNA product (generated using the outer primer) is then made, representing different copy numbers of the DNA ranging from 10^4 to 10^{10} copies and used as the template for quantitative real-time PCR (qPCR) with the inner primer pair to make a standard curve relating the copy number to the delta ct value in qPCR. With the genomic DNA at a given amount running side-by-side with the DNA standards, the total copy number of the genomic DNA can be determined by finding closest reference DNA concentration that gives the same delta ct value. The C-value of the genome (pg/copy) can then be calculated by dividing the total weight of the genomic DNA with the copy number, specifically as $C = m \times N^{-1}$, where m is the mass of the input genomic DNA in pg, and N is the copy number of the genome. Subsequently, the genome size in base pair (bp) can be calculated as $\Gamma = C \times N_A \times M_{Bp}^{-1}$, where C is the C-value, N_A is the Avogadro's number (6.022×10^{23} mol^{-1}), and M_{Bp} is the mean molar

mass of a DNA base pair (650 g mol^{-1}) (Wilhelm et al., 2003). The only requirement of this method is the availability of the sequence for one single copy gene from the test organism, which is not difficult to obtain nowadays. Therefore, it offers a flexible alternative for genome size determination to the FCM method.

A couple of novel approaches have emerged that utilize the Next Generation Sequencing (NGS)-based genome sequences for estimating genome sizes (Pflug et al., 2020). The first of such is a computational approach that analyzes the K-mer distribution among the raw genome sequence reads. K-mers are unique subsequences of specified length *k* in a larger sequence (Figure A.F.1), and K-mer distribution examines the relationship between the sequence coverage of the genome (copy number) and diversity (the number of unique entries) of K-mers at variable lengths in a genome. It is usually presented as a frequency plot, in which the *x*- and y-axis represents different K-mer sizes and the number of unique K-mers present in the genome sequences, respectively. For a genome with negligible number of repeat elements, sequencing errors or even coverage bias, the distribution of the K-mers generated from the reads of the genome displays a pattern of approximate Poisson distribution with the peak centered on the average sequence depth of the genome (Liu et al., 2013). When estimating a plant genome with an elevated level of heterozygosity and a large repeat content in the genome, there will be an intermediate peak, as K-mers from heterozygous regions will have half the coverage of a homozygous regions. The genome size can then be estimated by analyzing the plot to identify the peak (the K-mer size) in the plot and its corresponding genomic k-mer number, which gives the estimated genome size in bp (Chikhi and Medvedev, 2014; Li and Harkess, 2018). The pioneering work of Waterman in 2003 laid the basis for estimating genome size and repeat profile using k-mers, on which the Mixed-Poisson model and EM algorithm was developed but

without a program to use it (Li and Waterman, 2003). Shan and colleagues in 2009 developed and published a program called GSP (genome size prediction) using an extension of the Waterman method by incorporating Bayesian estimation and EM iteration (http://sourceforge.net/projects/gsizepred). However, the GSP program was unable to handle real-world sequencing data and has not been adopted widely.

Since then, many tools for generating k-mer plots have been developed, with the most frequently used ones being the Jellyfish package (Marçais and Kingsford, 2011) and Kmergenie tool (Chikhi and Medvedev, 2014). These methods have been used in the genome size estimation for several organisms such as the giant panda (Li et al., 2010), potato (Xu et al., 2011) and oyster (Wang et al., 2012). Despite the advantage of speed and ease of use, the results from the k-mer estimations have not been consistent and sometimes are significantly higher or lower than the actual genome size.

An alternative approach to inferring genome size from sequence data is to map NGS reads onto a set of putative single-copy genes using a reference-based assembler to determine the average coverage for the set of genes, and use that average as an estimate of coverage for the entire genome (Desvillechabrol et al., 2018). Unlike k-mer based estimates, this method requires a reference sequence for each locus, and its accuracy depends on these loci being truly single-copy. Another simple method is to use the reads to generate a *de novo* genome assembly (in the absence of a reference genome) and estimate the genome size based on the genome assembly length. The disadvantage of such method is that *de novo* genome assemblies may not be sufficiently complete and accurate for the purpose (Sun et al., 2018).

The first estimate of the genome size for lavender was around 5.7 pg using flow cytometry with *Lavandula officinalis* 'Chaix', which is a variety of *L. intermedia* (naturally

occurring hybrid of *L. angustifolia* and *L. latifolia*) (Zonneveld et al., 2005). A couple of years later, another study by Urwin and co-workers determined the genome size of a diploid *Lavandula angustifolia* to be 0.9 (+/- 0.07) pg using flow cytometry (FCM) and the chromosome number to be 50 using tomato and parsley as references (Urwin et al., 2007). The significant discrepancy between the results of the two studies seems to be due to the use of different materials with *Lavandula intermedia* likely having polyploidy genomes from the hybridization, and a C-value of 0.9 (+/- 0.07) pg is likely to be the more accurate value for the genome size of *Lavandula angustifolia.* Nevertheless, it presents a need for validation of this more recent estimate using a different method. A more recent investigation in a different cultivar of lavender (Jingxun 2) shows the genome size to be around 1094.97 Mbp using flow cytometry (Li et al., 2021), suggesting significant variations in genome size may exist among lavender cultivars, also emphasizing the need for validation of the results using multiple methods.

I.3.2 Polyploidy in plant genomes

Polyploidy commonly refers to the fusion of two or more genomes within one nucleus resulting in each cell containing more than two pairs of homologous chromosomes (Jiao and Paterson, 2014). The polyploidization events can lead to duplication, triplication or even higher order of multiplication of a plant genome. Such events have happened during evolution of plant genomes and have contributed to the diversity of the plant kingdom. For example, polyploidy can lead to generation of new species (Vicient and Casacuberta, 2017). Many plant lineages, including monocots (i.e., *Oryza*) and eudicots (e.g., *Arabidopsis*), have had at least one paleopolyploidy (ancient polyploidy) event in their evolutionary history (Jiao and Paterson, 2014). It is estimated that around 80% of all living plants are polyploids (Meyers and Levin, 2006). Many polyploidy genomes have been sequenced, with some of the representatives

including *Glycine max* (soybean) (Schmutz et al., 2010), *Arabidopsis lyrata* (Hu et al., 2011), *Jatropha curcas* (Sato et al., 2011)*, Gossypium arboretum* (Blanc and Wolfe, 2004; Li et al., 2014), wheat (Lukaszewski et al., 2014), strawberry (Shulaev et al., 2011), and coffee (Denoeud et al., 2014; Kyriakidou et al., 2018).

While polyploidy can occur naturally in the environment, it has also been introduced artificially using chemicals such as colchicine and oryzalin in breeding programs. The consequences of polyploidization are complex and can create advantageous or disadvantageous effects in different species. Polyploid plants have more than two paired sets of chromosomes, with many plants have tri- (3n), tetra- (4n), hexa- (6n), and even octo-(8n) sets of chromosomes. Such increase in the ploidy level may arise from a doubling of the same genome within a cell, resulting in an autopolyploid, or from merging two genomes from different species during hybridization, resulting in an allopolyploid (Soltis et al., 2009; Chen, 2010).

Polyploidization plays an important role in generating copy number variations (CNVs), novel genes, and other advantageous characteristics to the plants, including increased heterosis (refers to the hybrid vigour demonstrated by the progeny over its parents), resistance to deleterious mutations and adaptability to stressful environmental conditions (Panchy et al., 2016). For instance, duplicated genome blocks are hotspots of genomic rearrangements as they provide abundant opportunities for non-allelic homologous recombination (NAHR) events. CNVs can be generated by differential loss of duplicated genes, while new functions can be acquired from mutation of the duplicated genes (Simillion et al., 2002). In various plant species, a large proportion of duplicated genes are lost after a polyploidization event, a phenomenon called fractionation or diploidization, while a smaller proportion remains in a duplicated state (Fomeju et al., 2015). It has also been shown in various polyploid plant genomes that the

duplicated regions are involved in controlling complex agronomical traits (Combes et al., 2012). The genes in duplicated regions can undergo pseudogenization (loss of function), neo-functionalization (acquisition of novel function by a duplicated gene), changes in gene expression (Roulin et al., 2013). Mutations that accumulate in duplicated genes can give rise to allelic variants (paralogous genes) in plant genomes (Feldman and Levy, 2009). Beside these advantages, there may be some disadvantages of whole genome duplications (WGDs) in plants, such as unfavourable large-scale changes in gene expression and difficulties in undergoing mitosis and meiosis (Urwin, 2014).

There has been evidence that some high-quality varieties of lavenders, such as *L. angustifolia* and *L. intermedia,* were generated by introducing polyploidy using colchicine, which works by inhibiting mitotic spindles. Such varieties usually have larger flowers, thicker seeds and larger peltate glandular trichomes, as well as increased essential oil yields, compared to those of diploid plants (Urwin et al., 2007; Urwin, 2014).

Apart from whole genome duplication, duplications can also occur on smaller scales, and they include tandem duplication and transposition events, as well as segmental duplication (SD). SDs are defined as blocks of genomic sequence being 1 kb or larger in size and sharing a sequence identity >90% located in more than one site in the genome that are not generated by transposition. SDs can be inter-chromosomal or intra-chromosomal (Eichler, 2001; Feng et al., 2017), and they can impact the genome using similar mechanisms as WGD, but on smaller scales.

Key evolutionary events, such as ploidy and WGD, can be successfully inferred by analyzing the age distribution of homologous gene pairs (Blanc and Wolfe, 2004; Van De Peer et al., 2009). Synonymous nucleotide substitutions of protein-coding genes do not result in amino

acid changes and are neutral and free from natural selection. Therefore, the rate of synonymous substitutions (*Ks*) is proportional to the time lapsed since the duplication event and is thus used to approximate the timing of the occurrence of homologs, much like a molecular clock (Blanc and Wolfe, 2004). Because WGD results in the production of excessive paralogous gene pairs of a particular age, the *Ks* distribution of paralogs displays a peak of high density at a specific *Ks* value from which the timing of WGD is deduced (Blanc and Wolfe, 2004). Speciation timing can be determined by analyzing the *Ks* distribution of homologous gene pairs between two different species (also known as orthologs). The rate of nonsynonymous substitutions (*Ka*) and the *Ka*/*Ks* ratio serves as useful parameters to investigate the molecular evolution of two species that have diverged from each other (Fay and Wu, 2003). As synonymous substitution (*Ks*) occurs more frequently than nonsynonymous substitution (*Ka*), the *Ka*/*Ks* ratios of most orthologous genes are less than a value of 1, indicating that the gene pair is under purifying/negative selection (Fay and Wu, 2003). By contrast, orthologs under adaptive/positive selection exhibit *Ka*/*Ks* values greater than one, providing insights into the molecular evolutionary framework underlying adaptation, divergence, and speciation (Fay and Wu, 2003).

I.3.3 Plant genome characteristics

One of the important qualitative aspects of any genome is the genomic nucleotide composition, which is usually expressed as the proportion of guanine and cytosine bases in the DNA molecule (GC content). Conventionally, the base composition is expressed as the per cent proportion of G+C (i.e., GC content), which is assumed to have the most important influence on the resulting DNA characteristics and function. One important feature of the GC base pair is its higher thermal stability compared with the AT base pair, a feature that arises from the stronger stacking interaction between GC bases via triple hydrogen bond vs two hydrogen bonds between

A-T pairs (Yakovchuk et al., 2006). Apart from this, there is higher mutability of the GC base pairs is due to frequent cytosine methylation and higher energy cost of GC biosynthesis compared to AT base pairs (Šmarda et al., 2014).

The highest GC content in plants is found in grasses (Poaceae family). In monocot plants, a phenomenon called GC bimodality has been observed, which refers to the occurrence of two classes of genes that are distinguished by their GC content. As most plants are prone to whole genome duplications, such events provide redundancy in gene content that can relax selection pressure for individual genes. Nevertheless, the flexibility provided by redundancy after polyploidization may enable the evolution of genomic regions that have varied recombination rates and GC content (Sundararajan et al., 2016). With many non-model plants being sequenced recently, the analysis of GC content in plant genomes has yielded some interesting results in relation to genomic characteristics such as genome size, chromosome structure and environment adaptations (Šmarda et al., 2014).

The gene content in plant genomes is intricate due to various events like genome duplications, paleopolyploidy, and transposon activity, etc. Recent comparisons of model plant genomes, including *Arabidopsis*, rice, grapevine, poplar, papaya, have suggested that the ancestral angiosperm gene count to be around 12,000 to 14,000 (Proost et al., 2011). Gene duplication events are the major drivers for emergence of new genes and gene families (Giussani et al., 2001). One of the striking examples is the increase in the number of starch genes in papaya to 39 (for comparison, *Arabidopsis* has 20). Other gene families, such as secondary metabolism genes, cytochrome P450, and kinase genes, also have highly variable gene copy numbers among different plants (Claros et al., 2012). The analysis of such gene copy number and function data showed that more than half of the plant genes belong to gene families and around 45 % of the

genes have similar functions but having different gene expression patterns (Duarte et al., 2006; Claros et al., 2012). The correct identification and estimation of the presence of large gene families is very challenging and can be used as a key indicator for the accuracy of the genome assembly and annotation.

I.3.4 Repeat elements (REs) in plant genomes

The recent and rapid characterization of plant genomes using genome sequencing has led to the better identification of repeat sequences in plant genomes. The repeat sequences are classified based on their sequence characteristics. One of the common repeat sequences in plant genomes consists of tandem repeats, which includes the microsatellite or simple sequence repeats (Lerat, 2010). The plant genome also contains sequences called transposable elements (TEs), which are distributed across the genome. TEs were first described as controlling elements in the maize genome by Barbara McClintock (McClintock, 1950). The discovery of the repetitive nature of plant genomes was done by Flavell and co-workers in 1974 by measuring the reannealing kinetics of denatured DNA fragments from various plants including maize and wheat. The results showed that, on average, 80% of the plant genomes are repetitive in nature (Flavell, 1986). Later, it was discovered that those repetitive sequences are mostly TEs. Since repetitive sequences represent a large fraction of a genome without an obvious function, it was proposed that they simply act as spacers between genes and were considered "junk DNA" (Ono, 1972), or "selfish DNA" (Doolittle and Sapienza, 1980).

However, it is now widely acknowledged that TEs play an important role in genome dynamics and are the main cause of genome size variations in plants. The term, TE, is used for various genomic elements which have one thing in common: they can move in a genome (mobile elements). The overwhelming number of TEs and high similarity among members in the same

group made it necessary and challenging to categorize and characterize them. Several classification systems arose, as most research groups working on TEs used their own (Finnegan, 1989). Wicker and co-workers proposed a unified classification system for all eukaryotic TEs. It is a consensus of the already existing systems and describes guidelines for naming known and newly identified TEs (Wicker et al., 2007). In this system, two main classes were defined based on the mechanism of transposition: the class I, which transpose by a "copy-and-paste" mechanism using an RNA intermediate and are divided into five orders and 17 super-families; class II, which use a "cut-and-paste" mechanism without an intermediate, and they are subdivided into four orders and 12 superfamilies.

Class I elements transpose via a mRNA intermediate that is used as a template for reverse transcription to make a DNA copy, which is integrated into new locations in the genome. Therefore, every transposition event creates an additional element. In plant genomes, the long terminal repeat (LTR) retrotransposon elements are the most abundant class I members. They resemble retroviruses in their structure and life cycle but are not known to be infectious. Like the life cycle of retroviruses, class I TEs replicate by a cycle of transcription of the integrated copies, as if they were cellular genes, followed by translation of their encoded products and reverse transcription of the RNA into cDNA. They carry two long terminal repeats flanking an internal domain which, in fully functional elements, encode the proteins necessary for their own replication and retrotransposition. The two LTRs from each element are identical at the time of insertion. After insertion, both LTRs accumulate mutations independently and by comparing the two LTRs of a given element, the insertion time can be estimated (SanMiguel et al., 1998). The internal sequences contain two main Open Reading Frames (ORFs) for *Gag* encoding the

structural protein forming the nucleocapsid and Pol for encoding a polyprotein consisting of the reverse transcriptase (RT), RNAseH, and integrase (INT) (Wicker et al., 2007).

Gypsy and *Copia* are the main superfamilies of LTRs in plant genomes. They differ in the order and sequence domains of the encoded ORFs. Despite their abundance, the copy number of individual families differs in the plant genome. Most families are found in low or modest copy numbers. Still, some families successfully colonized their host genome to occupy a large fraction of the genome (Baidouri and Panaud, 2013). For example, in barley, 50% of the genome is made up of only 14 TE families, of which 12 are LTR retrotransposons (Wicker et al., 2009b).

In contrast to the class I elements, the transposition mechanism of class II elements is via a DNA intermediate that does not leave a copy at the original location. The known ten superfamilies of DNA transposon elements are found in virtually all eukaryotes (Feschotte and Pritham, 2007). The most common characteristics of a DNA transposon is the presence of Terminal Inverted Repeats (TIRs) and a transposase, which enables the cut-paste transposition. Eukaryotic DNA transposons can be divided into three major subclasses. The first class is the classical transposons which replicate by cut and paste mechanism. The second subclass utilized a mechanism related to rolling circle replication found in *Helitrons*, and the last subclass is the *Mavericks* whose replication mechanism is not well understood (Feschotte and Pritham, 2007).

Among the sequenced plant genomes, the ratio of TEs ranges from 3% for the minute 82 Mb genome of *Utricularia gibba* L. (Ibarra-Laclette et al., 2013) to 85% for maize (Schnable et al., 2009). Some TEs, such as LTRs, have adapted themselves to escape plant genome control mechanisms to achieve extremely rapid propagation and have contributed to excess DNA content in some plant genomes. The past and ongoing activity of TEs in the plant genomes contributes to

genome evolution by generating inter- and intra-species genomic diversity (Lee and Kim, 2014; Michael, 2014).

Analysis of various plant genomes have shown that TEs are distributed in various genic locations, such as promoters, introns, and exonic/CDS regions. The insertion of TEs can result in changes in protein sequences or changes in gene expression (Vicient and Casacuberta, 2017). The capacity for TEs to invade genomes may depend on both the TEs and the genome, with some TEs being able to escape controls imparted by a particular genome, with some genomes being more permissive to TE proliferation. Moreover, the amplification of TEs is not constant during evolution, and there are relatively dormant periods and periods with a burst of TE proliferations that result in genome expansions (Vicient and Casacuberta, 2017).

Miniature inverted-repeat transposable elements (MITEs) are a special class of non-autonomous TEs with a similar structure to class II transposons, but without transposase activity. MITEs have some common characteristics: short (80-500 bp), terminal inverted repeats (TIR), target site duplication (TSD), high AT content, potential to form a stable secondary structure, and a high copy number. MITEs are present in various plant genomes and have been associated with gene expression and gene duplication events in the genomes (Munoz-Lopez and Garcia-Perez, 2010). For example, MITE-associated genes are shown to have higher levels of expression than those not associated with MITEs in the rice genome (Lu et al., 2012). Importantly, recent genome-wide analyses revealed that MITEs insert preferentially into or near genes and that several families of miRNAs in humans, *Arabidopsis*, rice and *Solanaceae* are shown to be derived from MITEs, suggesting that MITEs play important roles not only in genome evolution but also in gene regulation (Lu et al., 2012).

Before the advent of NGS and automated TE identification tools, manual curation and community efforts were used to identify, characterize and annotate TE elements in model plants such *Arabidopsis*, rice, maize and other eukaryotic organisms (Ou et al., 2019). With the large numbers of reference genome assemblies that have been generated, it is not possible for community effort and manual curation to keep up with the identification and annotation of TEs and this has been replaced by TE identification programs and automation.

There are number of programs such as RepeatMasker (www.repeatmasker.org) (Smit et al. 1996), TRF (Tandem Repeat finder) (Benson 1999), RepeatScout (Price et al. 2005) and pipelines such as McClintock (Nelson et al., 2017) and PiRATE (Berthelier et al., 2018), which are available for identification of TEs in genome assemblies. A list of various TE identification tools and their descriptions are listed in the review paper by Bourque and Goerner- Potvin (Goerner-Potvin and Bourque, 2018).

These TE identification programs generally fall into two main categories based on the different strategies used in the identification and annotation of TE elements. The first category is the repository-based annotation where sequences are queried against known TE consensus sequences or TE motifs and the second is the *de novo* identification and annotation of new TEs. The repository-based annotation is the gold standard in the identification and annotation of TEs. The most widely used tool is the RepeatMasker (www.repeatmasker.org) (Smit et al., 2013) which queries against the RepBase and Dfam databases (Goerner-Potvin and Bourque, 2018). The disadvantages of using repository-based annotation are the depth and accuracy of the database, which are always in question and there are no species-specific data available in most cases. To overcome these disadvantages, *de novo* TE identification is preferred in newly sequenced genomes. Some of the widely used *de novo* identification tools are RepeatScout (Price

et al., 2005), RepeatModeler (Herrmann et al., 2016) and RECON (Bao and Eddy, 2002). These tools use pairwise similarity or consensus seeds as starting points to cluster repetitive sequences and identify them. The advantage of these tools is that they can be used to identity missing TEs from repository-based TE annotations (Ou et al., 2019).

Apart from these, there are general repeat annotation tools, such as TRF (Benson, 1999), which identify high copy number sequences in the genome. These programs have high sensitivity in identifying these repetitive sequences but lack the ability to classify and annotate the identified sequences. There are other programs which rely on learning algorithms and are based on the principle that nucleotide composition of TEs is different from that of the rest of the genome. There are also programs which rely on the characteristics or signatures of the TEs such as TSDs, a poly-A tail, terminal inverted repeats, long terminal repeats, or a hairpin loop. These are called signature-based methods (Girgis, 2015).

I.4 General strategies of genome sequencing

I.4.1 Development of DNA sequencing technologies

Genome sequencing is the process of experimentally determining the Deoxyribonucleic acid (DNA) sequence which can be thought of as long string made up of four characters `A', `C', `G' and `T'. The Sanger sequencing technique was the first method of DNA sequencing, developed by Frederick Sanger (Sanger et al., 1973). It is based on the chain termination principle, which works as sequencing by DNA synthesis, and it uses terminating 2,3-dideoxynucleoside triphosphate (ddNTP) to produce fragments of different lengths. Before its recent replacement by NGS, Sanger sequencing was the only DNA sequencing method for nearly a quarter-century. With most sequencing method, we can only reliably (with small error rate) read only a few hundred base pairs from the ends of a DNA fragment. Therefore, the genome is

first sheared into small fragments and the ends of the fragments are sequenced to produce small sub-sequences of the DNA string known as reads.

NGS, also known as high-throughput sequencing, is a broad term used to describe newer generations of DNA sequencing technologies, which work to sequence millions of nucleotides simultaneously and are thus also called massively parallel sequencing. NGS technologies, first became commercially available in 2004 (Mardis, 2008), have been exponentially increasing sequencing throughputs. While offering tremendous advantages mainly for their high throughput and associated high-cost efficiency, NGS increased the complexity of genome assembly due to the short sequence read length and higher sequencing error rates associated with most platforms. NGS offers applications in genome sequencing and re-sequencing, metagenomics (elucidation and study of nucleotide sequences from environmental samples), transcriptomics (RNA-sequencing) and functional genomics (e.g., ChIP-sequencing and methylome sequencing).

NGS is currently the method of choice for genome sequencing for its high throughput capacity and cost efficiency. This is reflected in the increasing number of genome assemblies from all domains of life. For plants, more than 230 angiosperm genomes have been sequenced (Chen et al., 2018). A recent count of completed plant genomes in the NCBI databases shows around ~1000 plant genomes at various stages of assembly, of which 500 are completed with chromosomal level assemblies and ~580 genomes at the scaffolding stage (a pre-completed stage in genome assembly process discussed below) (https://www.ncbi.nlm.nih.gov/genome). Although the cost of sequencing an organism's genome has been decreasing rapidly over the past few years, whole genome sequencing can still pose a technical and computational challenge.

One widely used NGS platform for genome sequencing is the Illumina platform with the paired-end (PE) and mate pair (MP) sequencing protocols. The Illumina sequencing technology

uses sequencing by synthesis and cyclic reversible DNA extension termination. (Bentley et al., 2008; Li and Harkess, 2018). For generating PE reads, long DNA fragments are sheared into smaller pieces and sequenced from both ends of the same fragment. The two reads are generated are called R1 and R2 and come from the same piece of DNA. Usually, the length of the fragment (<500 bp) is longer than the length of R1+R2, so there is a "gap" between them where the sequences are not determined (Figures A.F.2 and A.F.3.A). Mate-pair reads from Illumina (Li and Harkess, 2018) have also been used widely before the advent of long-read or single-molecule sequencing. Purified genomic DNA is first fragmented to reduce the high molecular weight DNA into smaller fragments of a desired size range. The DNA fragments are then end-repaired and biotin-labeled, placing biotinylated nucleotides at the ends of these fragments. Following biotinylation, DNA fragments of a particular size range required for library preparation are selected from an agarose gel. The length and range of the size-selected DNA determine the gap size and its variance of the mate-pair reads of the final library. The size-selected fragments are circularized by an intramolecular ligation. Any remaining linear molecules are removed by DNA exonuclease treatment (Figure A.F.2).

Compared to other NGS platforms currently on the market, the Illumina sequencing platform offers the best throughput, cost efficiency, and sequencing accuracy. However, it is limited by the short sequence read length (mostly around 150 bp), which imposes great challenges for genome assembly, especially in the case of a large genome with a high content of repetitive sequences. In addition, the use of PCR-based template amplification in the sequencing protocol, similar to other second-generation sequencing technologies, can lead to sequencing artifacts.

NGS technologies are advantageous because, unlike Sanger sequencing, DNA cloning is not required, making the process simpler, with greater adaption for a broad range of biological phenomena, and massive parallelization to reduce the costs. NGS has another advantage which allows a large number of libraries to be pooled and sequenced simultaneously during a single run, for example on an Illumina sequencer (Ekblom and Galindo, 2011; Fanning et al., 2017). However, NGS does suffer from some disadvantages: the short sequence length requires unique assembly algorithms, base calling is less accurate than Sanger sequencing, and the quality of NGS assemblies is lower than Sanger sequencing (Claros et al., 2012).

To ensure that every position of the genome is covered by at least a few reads, over-sampling is necessary. This over-sampling is quantified in terms of "coverage" calculated as the ratio of the total number of base pairs sequenced to the size of the genome. Higher coverage is desirable for completeness of assembly, but this also increases the amount of sequencing data, which, in the presence of errors, complicates the assembly process. The assumption of uniformity of coverage is also not correct: it is known that certain parts of the genome are more difficult to sequence than others (Benjamini and Speed, 2012).

Despite the widespread adoption of the Illumina technology in whole genome sequencing, the lack of long reads was a limiting factor for resolving repeat elements and generating high quality draft assemblies. To overcome the most critical aspect of second generation genome sequencing (which is the short read length generated), long read technologies have been developed, as third generation sequencing (TGS), which currently include Pacific BioScience (PacBio) technologies (Rhoads and Au, 2015) and Oxford nanopore technologies (Lu et al., 2016). The PacBio platform uses the single molecule real time (SMRT) technology, which can produce long reads exceeding 10 kb in length (Eid et al., 2009). Oxford Nanopore

developed a sequencing technology based on nanopores (1 nm in diameter), which allows only a single DNA molecule to thread through a single nanopore at a time, during which a nucleotide-specific change in the electric current is generated across the immersed nanopore as the basis for calling the identity of the individual nucleotides passing through (thus the sequence of the DNA molecule), and it can generate sequence as long as 1 Mbp (Lu et al., 2016), Both TGS platforms use variations of single molecule sequencing in real-time without any prior amplification of the DNA sequences (Li and Harkess, 2018). The reads generated by these sequencing technologies are longer than 10 Kbp, which is very useful in *de novo* genome assembly. The main drawback is the higher error rate as compared to the second generation sequencers (range between 0.1 and 1%), which is around 10 -15% in the case of PacBio reads (Li and Harkess, 2018) and 5 – 15 % error rate for Oxford nanopore sequencing (Rang et al., 2018). PacBio does offer a HiFi option to reduce the error rate by performing multi-rounds of sequencing of the same DNA template to generate a consensus sequence. However, in this case, this further reduces its cost-efficiency and read length. To take advantage of different NGS platforms, an emerging trend is to use the multiple NGS platforms in a genome sequencing project if the cost is of a less concern, for example, combining Illumina and one of the TGS platforms (Yang et al., 2016; Zapata et al., 2016; Jiao et al., 2017; Schmidt et al., 2017; Vining et al., 2017; Li et al., 2021).

I.4.2 *De novo* genome assembly

There are currently two main methods of genome assembly: reference-based genome assembly and de-novo genome assembly. Reference-based assembly uses a reference genome (pre-assembled genome) of an organism (usually a model organism) as a template to guide genome assembly of different individuals from the same species. *De novo* genome assembly is

usually performed in the absence of a reference genome, mostly for the analysis of non-model organisms. Even though it has received enormous attention in the last couple of decades, many challenges remain for *de novo* genome assembly. For this reason and since it is the approach used in this thesis work, the process of *de novo* genome assembly is the focus for this part of literature review. The process of *de novo* genome assembly can be broken down to pre-assembly data processing, contig assembly, scaffolding, and post-assembly fine-tuning.

1.4.2.1. Pre-assembly data processing

Once the sequencing is complete with the sequence read files downloaded from the sequencing centres, analyses should be taken to ascertain the quality of the sequencing data. Poor quality reads, ambiguous base calling, contamination, biases in the data and even technical issues on the sequencing chip are some examples. The read quality is estimated by specialized programs which provide key statistics regarding the quality of the reads. FASTQC (Andrews, 2016) is one of the widely used quality-checking programs for NGS sequencing. This program provides data on per base sequence quality, GC content, sequence length distribution, presence of adaptor sequences (platform specific short oligo sequences of 8-10 bases, required for fragment recognition by the sequencing instrument) and many others. A key feature of adaptor sequences is that each NGS instrument provider uses a specific set of sequences. Once the quality report is generated, adapter and multiplex index sequences can be screened for and removed. One of the characteristics of Illumina sequencing reads is the higher number of sequencing errors towards the 3' end of the reads and non-uniform distribution of reads across the genome. The presence of errors in the sequencing reads can complicate genome assembly by introducing complexities during contig generation and scaffolding processes (Heydari et al.,

2017). The errors introduced during sequencing can be rectified using error correction tools. The main principle behind error correction is the generation of a k-mer coverage spectrum (distribution of a set of decomposed distinct *k*-mer observed in a group of reads) and identifying sequence variants with low coverage and replacing it with similar k-mers at higher coverage. The sequencing error issue can also be dealt with by simply trimming off or discard sequences with low per-base quality.

I.4.2.2 Contig assembly

The first step in genome assembly after error correction is the generation of contigs. A contig is a continuous DNA sequence representing the consensus sequence of a set of overlapping sequence reads (Figure 2). The two main algorithms used by *de novo* genome assemblers for this stage are De Bruijn graph (DBG) (Figure 2B) and OLC (Overlap layout consensus) (Figure 2C) (Ye et al., 2012) algorithms. The DBG algorithm breaks each read into fixed size k-mers (Figure 2B) and constructs the graph where the vertex is a single k-mer and the edge is represented by two adjacent k-mers overlapping by k-1 letters (Figure A.F.1) (Idury and Waterman, 1995). In the context of genome assembly, DBG is a directed graph where each node is a distinct k-mer present in the input fragments, and an edge is present between two k-mers when they share an exact (k−1) overlap (Figure A.F.1) (Compeau et al., 2011). The advantage of the DBG is that there is no requirement for the overlap computation, which can be very memory intensive and time-consuming when dealing with enormous number of NGS short reads. Although the DBG algorithm skips the overlap computation step, the graph construction step is efficient, but the graph generated can be highly complex due to the presence of repeats, which are mostly larger than the k-mer size (Miller et al., 2010; Huang and Liao, 2016). Among the

existing popular genome assemblers, Abyss (Simpson et al., 2009), SOAPdenovo (Luo et al., 2012), velvet (Zerbino and Birney, 2008), SPADes (Bankevich et al., 2012) use the DBG algorithm.

In contrast to DBG algorithm, OLC works by identifying all pairs of overlapping reads to construct a graph with vertices represented by reads and overlaps denoted by edges (Myers, 1995) (Figure 2C). The overlaps are computed by a series of pairwise sequence alignments, which are computationally expensive as they require huge amount of memory to store all the alignment information. However, such kind of graph representation is advantageous as it uses the entire length of the reads and can be useful in resolving repeat regions which are abundant in plant and animal genomes (Miller et al., 2010). Some of the earliest *de novo* genome assembly tools employing OLC algorithms are Newbler (Chaisson and Pevzner, 2008) and MIRA (Chevreux et al., 1999), etc.

Apart from these two common algorithms, there is a derivative of the OLC algorithm which is called a string graph (Myers, 2005; Huang and Liao, 2016). The algorithm uses FM-index (full-text minute space index) in the construction of the assembly graph using the input reads. The FM (Ferragina and Manzini, 2000; Simpson and Durbin, 2010) is a sub-string index based on the Burrows-Wheeler transform (BWT). The BWT is an algorithm for data compression using a suffix array. It is efficient for not requiring additional storage of data and is also reversible (Kufleitner, 2009). The main advantage of using such algorithm in graph generation is that error-prone regions are better handled (Schirmer et al., 2016). Examples of genome assemblers utilizing the string-graph algorithm include Edena (Hernandez et al., 2008), Readjoiner (Gonnella and Kurtz, 2012), and FMJ-assembler (Liang et al., 2014).

SGA (String Graph Assembler) (Simpson and Durbin, 2012) is one of the widely used genome assembler programs based on String graph in combination with the FM-index. The assembler retains many of the key features of the OLC assemblers. In addition, it uses the compressed FM-index which is memory efficient and keeps the reads intact instead of dividing them into k-mers as in DBG algorithm. Fermi (Li, 2012) is another SGA assembler with an improved algorithm generating scaftigs, which are continuous sequences formed by lining up multiple initial contigs which have sequence overlaps.

Most popular genome assemblers (e.g., AllPaths and SOAPdenovo) are pipelines, which consist of modular sets of programs/tools for performing the entire genome assembly process. Apart from these genome assembly pipelines, there are stand-alone tools for specific steps of the genome assembly process, which may outperform the corresponding components in the genome assembly pipelines. Some of the commonly available contig assemblers are CAP3 (Huang and Madan, 1999), velvet (Zerbino and Birney, 2008), FERMI (Li, 2012) and for scaffolders are OPERA (Gao et al., 2016), SSPACE (Boetzer and Pirovano, 2014) and SOPRA (Dayarian et al., 2010).

I.4.2.3 Scaffold assembly

Once the contigs are generated, the next step is the generation of scaffolds (Figure 2D), which are longer sequences consisting of properly ordered and oriented contigs separated by variable gap sequences. The length of gaps may represent estimated distances between contigs based on the estimated sizes of the PE or MP library fragment sizes. Most of the scaffolders use various heuristic solutions to generate the scaffolds and some use combinatorial algorithms which use both heuristics and paired and mate-pair end reads (Figures A.F.2 and A.F.3) to

generate scaffolds (Gao et al., 2011). Some of the scaffolders are part of the genome assembly pipeline like Abyss (Simpson et al., 2009), SOAPdenovo 2 (Luo et al., 2012), SPAdes (Bankevich et al., 2012) and Allpaths (Simpson and Durbin, 2012), while others are stand-alone (not part of the genome assembly pipeline) like OPERA (Gao et al., 2011), SSPACE (Boetzer et al., 2011), WiseScaffolder (Farrant et al., 2015). Some newer scaffolders, such as LINKS (Warren et al., 2015), can utilize the long-reads from the third generation sequencers. OPERA-LG (Gao et al., 2016), an improved version of OPERA, a stand-alone scaffolder, can simultaneously use multiple alignment files such as binary alignment (BAM) files from multiple libraries (paired-end and mate-pair) to better deal with scaffolding of repetitive sequences. A schematic diagram which illustrates the steps in *de novo* genome assembly including contig generation, scaffolding, gap filling is given below (Figure 2).

There have also been some recent collaborative community efforts to provide assembler evaluation, for example, the Assemblathon (Earl et al., 2011; Bradnam et al., 2013) and GAGE (Salzberg et al., 2012). These evaluations indicate that no single assembly software performs universally better than others, suggesting the need of continuing development of better tools for genome assembly along with the quickly changing sequencing technologies.

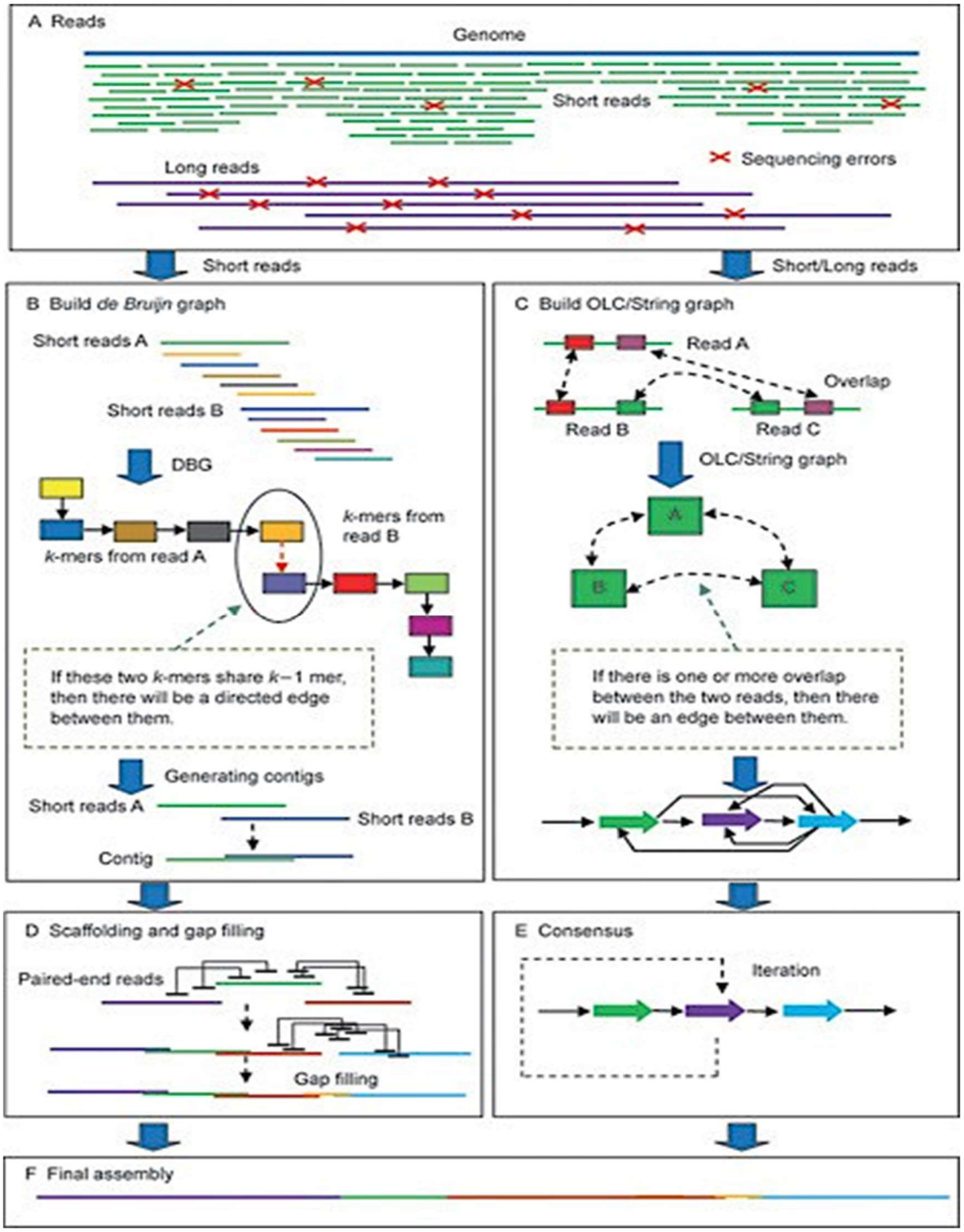


Figure 2. A diagram illustrating the major steps of *de novo* genome assembly.

This figure was adapted from "Current challenges and solutions of *de novo* assembly" by Liao, X., Li, M., Zou, Y. et al. (Liao et al., 2019). A, different types of reads used in the assembly process; B, de Bruijn graphs; C, OLC/String graph; D, the principle of scaffolding and gap-filling using short reads; E, consensus operation; F, final genome assembly generated.

I.4.2.4 Post-genome assembly improvement and quality assessment.

Some tools have been developed for improving newly generated *de novo* genome assembly through gap filling, improving or correcting connections of contigs within scaffolds based on transcriptome data. Programs for gap filling include gapcloser in SOAPdenovo2 (Luo et al., 2012) and GAPPadder, a standalone tool (Chu et al., 2019). Tools, like L_RNA scaffolder (Xue et al., 2013), are designed to improve *de novo* genome assembly based on RNA-seq data or other forms of transcriptome data, which can be generated from the same organism. These tools use long single-end RNA sequencing reads or short PE RNA-sequencing reads for scaffolding contigs by first searching and finding guide transcript exons, which map to different genomic fragments followed by orienting and ordering the genome fragments into longer scaffolds based on their mapped exons belong to the same transcripts (Figure A.F.4).

Once the draft genome assembly is generated and finalized, its quality can be assessed by analyzing multiple metrics, such as number of contigs, N50, the lengths of the largest contig and scaffold, the total length of all sequences and non-gap sequences in the assembly, the number of mismatches /100kb (average number of mismatches per 100 000 aligned bases), and others. The N50 is a commonly used genome assembly quality metric, and it is a length weighted median obtained by first sorting all sequences in order by length and then finding the length of the smallest sequence up to which the sum of all sequence length covers at least 50% of the entire assembly (Figure A.F.5) (Castro and Ng, 2017). The larger the N50, the better the genome assembly. While the N50 value provides information about the degree of the fragmentation in the assembly, it provides no information regarding the correctness of the genome assembly (Narzisi and Mishra, 2011). To overcome the shortfalls of N50, other similar metrics, such as NG50 and NGA50 have been added. NG50 refers to N50 based on the known genome length (G for

genome) (Castro and Ng, 2017). It is useful if there is a reference genome available, in which case, more assessments, such as the number of mis-assemblies, unaligned contigs (number of contigs that have no alignment to the reference sequence), genome fraction (%) and duplication ratio can be performed (Gurevich et al., 2013). In this case, NGA50 is the N50 equivalent based on only the sequences that can be aligned to the reference genome.

Additional levels of quality of assessment can be done based on the presence of core genes, content of repeat elements, etc. There are a number of tools which provide these statistics, with the popular ones including Quality assessment tool (QUAST) (Gurevich et al., 2013), plantagora (Barthelson et al., 2011), and GAGE (Salzberg et al., 2012). The plantagora and GAGE are useful only if there is a reference genome available, while the QUAST tool provides key metrics such as the number of contigs, largest contig, total length of the assembly, GC % (a key genome compositional characteristics) for *de novo* genome assemblies. In the case of the QUAST tool, if a reference genome is provided, additional metrics like the number of mis-assemblies, the number of misassembled contigs and duplication ratio are provided. The definitions and explanations of these metrics are described in the detail in the publication by Gurevich A. and co-workers (Gurevich et al., 2013).

I.4.3 Genome annotation

A genome sequencing project does not stop at the completion of a draft genome sequence. Rather it permits analysis of the genome and functional biology of the organism, first via computational annotation to define the components of genome, including genes and repetitive sequences.

As a standard procedure, a part of the genome annotation is to identify TEs and other repetitive DNA sequences. The most commonly used tool for this task is RepeatMasker (Smit et

al., 2013). Repeatmasker uses RepBase (Jurka et al., 2005; Bao et al., 2015), a curated repeat consensus sequence library, which is frequently updated. If a repeat library is unavailable for a new species, tools like RECON (Bao and Eddy, 2002) or RepeatScout (Price et al., 2005) can be used to generate *ab initio* repeat libraries. The RECON is a part of the RepeatModeler pipeline (Flynn et al., 2020) which is used for *de novo* identification and accurate creation and compilation of consensus of repeat sequences in a genome. The key outputs of RepeatMasker include: 1) a text file containing a list of repeats with their associated information including genome location, repeat element (RE) classification, sequence divergence from the consensus and location in the consensus; 2) a modified or revised version of the input genome sequences, in which the sequences of the repeats are masked either as ambiguous bases, i.e., "N" (hard masked) or in lower case letters (soft masked); 3) a text file containing a summary table of the repeats in the genome. The masking step signals to downstream sequence alignment and gene prediction tools that these regions are repetitive sequences.

The most important part of the genome annotation is gene annotation. The process involves prediction of all genes and their functions in the assembled genome. Specific items of the gene information include the types of genes (protein coding and non-coding) and the precise genomic coordinates or position. For protein coding genes, the annotation includes the specific exons, coding sequence and protein sequences and the function. Different approaches and tools are used for predicting different types of genes, among protein coding genes, rRNA genes, tRNA genes, and other non-coding genes. For protein coding genes, different approaches are available, and they include *de novo* gene prediction, homology-based prediction, and evidence-based prediction. Each of these methods will be described later in this section. Methods of genome annotation can also be organism specific. For example, gene prediction in prokaryotes is quite

different from that in eukaryotes as the DNA sequences in the prokaryotes are less sophisticated due to the lack of introns. Here we focus on the gene prediction in eukaryotes. Among the different types of eukaryotic genes, the prediction of tRNA and rRNA genes is relatively straight forward, and is performed based on utilizing two levels of Hidden Markov Model (HMM) for rRNA and combining three tRNA search methods and Pavesi search algorithm (Pavesp et al., 1997) for tRNA. A commonly used tool for rRNA is RNAmmer (Lagesen et al., 2007) and for tRNA is tRNAScan (Lowe and Eddy, 1996).

Ab initio gene prediction methods predict genes by finding the most likely open reading frames (ORFs) based on the extrinsic features of eukaryotic gene structure, with the most important one being the known exon/intron boundary sequence motifs (Wang et al., 2004). For this reason, *ab initio* gene models mostly lack the prediction of untranslated regions (UTRs) and alternatively spliced variants. The tools include Augustus (Stanke and Waack, 2003), Genemark (Besemer and Borodovsky, 2005) and SNAP(Korf, 2004). Hidden Markov Models (HMM) are commonly used in these tools and can be improved with additional evidence. Newer tools like mGene (Schweikert et al., 2009) and mSplicer (Rätsch et al., 2007) use machine learning techniques for gene prediction. Overall, the effectiveness of *de novo* gene prediction for eukaryotes is still limited with the predicted gene models requiring validation.

The homology-based gene prediction works by finding sequence similarity matches to known genes in other organisms at the DNA level or protein level. Due to the rich information of gene sequences in the databases covering a large number of organisms, the most effective approach and the simplest approach to finding gene function is to find a match to the sequences in many of the reference databases like NCBI (Geer et al., 2009), Pfam (El-Gebali et al., 2019), Uniprot (Bateman et al., 2017), and comprehensive plant genome databases like the Phytozome

repository (Goodstein et al., 2012) and individual plant genome consortium resources. The primary sequence matching tool used widely by the scientific community is the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990).

The third approach is evidence-based gene prediction, and currently it works by aligning the available transcript sequences (RNA-seq or transcriptome data) derived from the same organism to the genome sequences to provide direction identification of exons and splice variants of genes. Before the advent of NGS, generation of transcriptome data relied on the use of expressed sequence tags (ESTs), which is time-consuming and costly. However, with application of NGS in transcriptome analysis, RNA sequencing (RNA-seq) offers a rapid and cost-effective method for generating the transcriptome data for an organism before or along the genome sequencing. So, transcriptome data are now commonly used in gene prediction.

Many packages/pipelines have been developed to automate the genome annotation. While running locally (vs a web server), they allow users to specify a set of gene prediction tools employing different strategies to take the advantages of each approach and use of transcriptome data in different formats. Popular and widely used genome annotation tools include Maker (Cantarel et al., 2008; Holt and Yandell, 2011) and PASHA (Haas et al., 2003). MAKER is a configurable genome annotation pipeline which is able to generate *ab initio* gene prediction, aligns RNA-seq/transcript sequences and proteins to genomes and synthesizes these data into gene annotations (Holt and Yandell, 2011).

These automated gene or genome annotation pipelines use an algorithm called “Chooser algorithm”, which selects the best possible consensus based on the available evidence and combines to generate the final version of the gene structure (Yandell and Ence, 2012). Some of the prominent tools used in annotation pipelines are evidence modeler (Haas et al., 2008),

GLEAN (Elsik et al., 2007) and JIGSAW (Allen and Salzberg, 2005). Pipelines that generate a consensus gene model are available for prokaryotic and eukaryotic genome annotation. Current popular tools to annotate a genome include the automated pipelines MAKER-P specifically developed for plants (Holt and Yandell, 2011) and the generalist BRAKER1 (Hoff et al., 2019). The model parameters are trained using evidence from RNA-sequencing data, expressed sequence tags (ESTs), and annotated gene models in related species, or by using parameters optimized for a model species. The advantage of MAKER pipeline is that it can leverage existing programs and utilize parameters of model organisms or closely related organisms and generates an output which is a consensus, or a best possible gene model based on the evidence provided.

Once the gene structures of protein coding genes have been detected, the next step is to ascribe biological function to the genes in a process known as functional annotation. Surprisingly, performing this task to a degree of high accuracy remains challenging, despite the extensive accumulation of knowledge about gene function in model and crop species. Indeed, there is still a large percentage of genes with orthologs found across multiple species but with function not ascertained. Many methods of function prediction rely on identifying similarity in sequence and/or structure with one or more well-understood proteins. Alternative methods include inferring conservation patterns in members of a functionally uncharacterized family. The most simple method for predicting the function of a protein is to run a BLAST search against reference sequences or use InterProScan (Mitchell et al., 2019) tool which assigns functions to protein sequences. The InterProScan tool (Quevillon et al., 2005) is an integration platform that uses 14 special databases to assign functions to protein sequences using HMMs, position-specific scoring matrices and pattern matches to ascribe function to novel protein sequences. This tool is extensively used and regularly updated to provide a relatively accurate annotation of protein

sequences. The InterProScan annotated protein sequences are regularly uploaded to UniprotKB (Magrane and Consortium, 2011) and forms the biggest repository of annotated protein sequences.

The gene annotation process for eukaryotic genomes is a very CPU intensive and memory/storage space-demanding process. Furthermore, most genome annotation tools require extensive computational resources and bioinformatics expertise to run them properly and processing of the output from the program runs. At the same time, there are web-based genome annotation tools available to facilitate quick annotation of genome sequences for bench scientists and some of the widely used tools include Web Apollo (Lee et al., 2013), Mercator (Lohse et al., 2014), KOBAS (Xie et al., 2011). In addition to these tools there are other tools like GenSAS, which offers a web-based fully functional genome annotation pipeline (Humann et al., 2019). It integrates popular command line annotation tools in a single easy-to-use online interface with a step-by-step guide to complete a genome annotation project. The interface has options for uploading user generated files such as organism specific transcripts, protein and RNA-seq evidence to help and improve the annotation process (Humann et al., 2019). The limitations of such tools are the limited access and long turn-around time as the servers can be extremely busy.

I.4.4 Genome assembly and annotation quality assessment based on gene content

Genome assembly quality metrics mentioned earlier do not provide any information about the presence of genes or if the sequences have been assembled accurately. Furthermore, there is no information about the presence of key genes in the draft genome or the completeness of the genome assembly. Once the genome annotation is completed, the quality and completeness of genome assembly and annotation can be assessed through the presence and

correctness of well-known single-copy genes (a set of 1440 conserved universal single copy genes in plants) called 'Benchmarking Universal Single-Copy Orthologs' (BUSCO) (Simão et al., 2015) or by using a smaller set of genes in the 'Core Eukaryotic Genes Mapping Approach' (CEGMA) (Parra et al., 2007). The BUSCO scores for well-annotated genomes vary between 95–97%, with 3–5% of the 1440 conserved universal and single copy genes being missing or fragmented (Raymond *et al.*, 2018; Springer *et al.*, 2018). However, conserved single-copy genes do not necessarily provide a suitable indicator for annotation quality of genes that have duplication, such as members of gene families. This is particularly relevant for plant genomes in which over 80% of the genes belong to gene families (Guo, 2013). Some other features such as the presence of specific protein domains in the genome can also be used as a reliable indicator of proper annotation. The presence of conserved protein domains indicated as a percentage of a core set, can be used in assessing the quality of the annotation (Dohmen et al., 2016). One of the programs which has incorporated the identification of conserved protein domain arrangement using HMM is DOGMA, which has been used in this study (Dohmen et al., 2016). A disadvantage of this tool is that it cannot be used in the analysis of whole genomes or transcriptomes like BUSCO but can only be used with the annotated protein sequences.

Some other programs are available to provide annotation quality assessment. For example, Genevalidator (Dragan et al., 2016) can be used to provide assessment of the quality of the gene annotation by performing multiple analyses based on comparisons to multiple gene sequences from large databases such NCBI non-redundant (nr) database (Benson et al., 2007, 2017), SwissProt (Bateman et al., 2017) using BLAST. Once the sequence comparison is completed, the BLAST results are subjected to characterization based on sequence signatures such as length, coverage, presence of conserved regions and presence of different genes. The

results of each comparison indicate whether characteristics of the query gene prediction deviate from those of matching sequences (Dragan et al., 2016). Another annotation metric developed by Eilbeck and co-workers, which is now used widely, is the Annotation Edit Distance (AED) score described in detail in their 2009 publication (Eilbeck et al., 2009) (Figure A.F.6). This score is used by the MAKER *de novo* annotation pipeline (Eilbeck et al., 2009; Holt and Yandell, 2011) to identify and determine good quality annotation. AED score, ranging from 0 to 1, determines the level of confidence of the newly annotated genes with 0 or close to 0 being high confident annotation and scores close to 1 being the opposite. Despite being stand-alone, such tools can be incorporated into pipelines and the results can be used as a guide to select or remove newly identified genes (Figure A.F.6). While these tools directly assess the genome annotation quality, much of the gene annotation is impacted by the genome assembly quality, including the connection of contigs, the degree of fragmentation, and the completeness (coverage) of the genome.

I.4.5 Genome features of model and important crop plants

The first plant genome sequenced and annotated was the genome of *Arabidopsis thaliana* (Kaul et al., 2000). The genome project took 10 years and US$100 million to complete, providing important resources for plant genetics research. One advantage of *Arabidopsis* is that it has a small nuclear genome. In its 115.4 Mbp sequence of the 125 Mbp genome, a total of 25,498 protein encoding genes from 11,000 families have been identified (Kaul et al., 2000). The updated *Arabidopsis* genome has 27,655 genes, 5,178 non-coding genes, 905 pseudogenes and 3,901 transposable elements (www.arabidopsis.org).

Rice genome (*Oryza sativa*) was the first crop genome sequenced (Goff et al., 2002; Yu et al., 2002). The genome size is ~430 Mbp containing ~56,000 genes with an average gene size of 2,853 bp at 4.9 exons/gene (Kawahara et al., 2013). Rice has about 30% more genes than *Arabidopsis*, which is largely attributed to gene family expansion. The rice genome has a large number of repetitive sequences which cover 40% of the genome, of which 35 % are TEs (Gill et al., 2010). Subsequent to the publication of the rice genome, the maize (*Zea mays*) genome, with the total sequence length of 2.3 billion bp was sequenced at a cost of US$31 million (Schnable et al., 2009). One key feature of the maize genome is that its 80% is made up of repeat sequences of different types. A recent study has identified ~130,000 copies of intact transposons with the total length of 1,268 Mbp, of which more than 50% are nested retrotransposon copies disrupted by the insertion of other transposable elements (Jiao et al., 2017).

All these three genomes were initially sequenced by the traditional Sanger sequencing method. The *Arabidopsis* genome initiative used large-insert bacterial artificial chromosome (BAC), phage (P1) and transformation-competent artificial chromosome (TAC) libraries (Kaul et al., 2000). Similarly, the rice genome was sequenced using BAC and P1 artificial clones (PAC). The genome assembly used a high-density genetic map, expressed sequence tags (ESTs), yeast artificial chromosomes (YAC), and BAC based physical maps. The total coverage of the BAC clones was approximately ten-fold, leading to a 389 Mbp genome assembly at an error rate of 1/10kb (Goff et al., 2002). Maize was sequenced with BAC's and fosmid clones in conjunction with an integrated physical and genetic map (Schnable et al., 2009).

The close relative of the lavender plant, the mint genome was sequenced in 2017 using a combination of Illumina and PacBio long reads (Vining et al., 2017). The assembled genome size is 353 Mbp with 35, 597 protein-coding genes identified, among which there are 292 disease

resistance genes and nine essential oil genes. The draft genome consists of 190,876 contigs covering 88% (353 Mbp) of the estimated 400 Mbp genome with a calculated GC content at 36.6% and scaffold N50 at 4,474 bp. The scaffolds are assigned to 12 pseudochromosomes using genetic linkage maps. The total coverage of the multiple types of sequencing reads used in the assembly process is ~66X (Vining et al., 2017).

I.5 Research Objectives

Although lavender has been developed as a model system for studying EO production (Lane et al., 2010), and numerous EO genes have been reported along with the accumulation of some transcriptome data (Landmann et al., 2007; Demissie et al., 2011, 2012, 2013; Sarker et al., 2013; Jullien et al., 2014; Adal et al., 2017, 2019; Adal and Mahmoud, 2020; Wells et al., 2020), several questions regarding the regulation of EO metabolism and storage in these plants remain unanswered. The goal of this research is to perform genome sequence analysis of the nuclear genome of *Lavandula angustifolia Maillette*, a cultivar chosen for its significant economic impact. Specifically, the project aims to generate the first draft genome sequence of the *L. angustifolia* (Maillette) and provide annotation of the genes and repetitive elements, followed by a detailed analysis of EO genes. The sequencing and characterization of the lavender genome can lay the foundation in better understanding the biology and genetics of EO biosynthesis and many critical aspects of the plant.

Chapter II. Materials and Methods

II.1 Plant DNA extraction

Plant materials used in this study are from *L. angustifolia* sp. (Maillete) obtained from the Okanagan Lavender and Herb Farm (Kelowna, BC, Canada). Leaf tissue was collected and immediately freeze-dried in liquid nitrogen. Genomic DNA was extracted using Geneaid Genomic DNA extraction kit (plant) (Geneaid Biotech Ltd., Taiwan) as per the manufacturer's instructions. The quantity and quality of the extracted DNA were determined using NanoDrop 2000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and agarose gel (1 %), and the DNA was stored in −20 °C until use.

II.2 Genome sequencing

High quality *Lavandula angustifolia* DNA was sent to Eurofins Genomics (Huntsville, USA) for sequencing using the Illumina HiSeq2000 platform. Sequencing was done to generate short paired-end (PE) reads together with mate-pair (MP) reads. The PE reads were 100 bp x 2 in length with the insert size at ~250 bp, while the MP reads were also 100 bp x 2 for the read length and the three types of insert sizes at approximately 3 kb, 8kb, and 20kb. The PE reads were used to generate the contigs and the MP reads were used for generating the scaffolds. In addition to the PE and MP reads, overlap reads (DNA fragment which is shorter than two times the read length resulting in overlap reads) (A.F.3.B) were also generated, as required by one of the genome assemblers we tested, Allpaths-LG. The overlap reads, done similarly as for the standard PE reads but at insert size ~200 bp, were also provided as input to multiple genome assemblers to complement the PE and MP reads. Table 2 summarizes the number of reads generated with the respective insert size, length and total coverage for each of the read types. A

total coverage of ~150X was obtained with the majority from the PE reads (88X) followed by the overlap reads (~52X).

Table 2. Summary of sequencing libraries, read count and coverage used for *de novo* genome assembly

Library type	Insert size	read length (bp)	read count	Seq (Gbp)	Coverage%
Pair end	300 to 500 bp	100 x 2	752,660,358	75.3	88.5
Overlap	200 bp	100 x 2	282,659,804	28.3	52.1
Mate pair	3kb	100 x 2	23,139,400	2.3	
	8kb	100 x 2	58,548,234	5.9	10.2
	20kb	100 x 2	19,894,354	2.0	
Total			1,136,902,150	113.7	150.8

II.3 Genome size estimation

II.3.1 Genome size estimation using quantitative real time PCR (qRT-PCR)

The qRT-PCR method was based on the work of Wilhelm and co-workers (Wilhelm et al., 2003). Two sets of nested primers were designed for each of two genes, *Dxr* (1- deoxy-D-xylulose 5-phosphate reductoisomerase) and *Hmgs* (Hydroxymethylglutaryl-CoA synthase), which are known to be in single copy in most plant genomes (Carretero-Paulet et al., 2002; Tholl and Lee, 2011), for estimating genome size of *Lavandula angustifolia*. Detailed information of the primers used is provided in Tables 3-6. The outer primers were used for generating PCR fragments in preparing the DNA standards and the inner primers were used for qRT-PCR. As a

validation of the method's accuracy, *Arabidopsis thaliana* was included as a control using the orthologous of the genes.

PCRs for making DNA standards with known copy numbers using the outer primers were performed as using standard end-point PCR. DNA from the PCR reaction were purified using gel purification method (Abraham et al., 2017) and the concentration was determined using Nanodrop instrument. Serially diluted DNA standards at incremental dilutions (1,2 and 4 µg/µl) were used to generate the standard curve required for qRT-PCR analysis.

qRT-PCR with the inner primers was performed on the Corbett Rotor-Gene 6000 thermocycler (Corbett Robotics Inc., San Francisco, CA, USA). Each 15 µL reaction contained 1 µL of the DNA template, 0.75 µL of each primer (forward and reverse) and 7.5 µL of SsoAdvanced Universal SYBR Green Supermix (Bio-Rad, Hercules, CA, USA) and ddH20 was used to make up the total volume to 15 µL. The qPCR reaction was performed at 98 °C for 3 min, followed by 40 cycles of 95 °C for 10 sec and 60 °C for 20 sec. A high-resolution melt curve was generated to verify amplification specificity, and cycle threshold (CT) values were determined by the Rotor-Gene Q series software. The C-value for each of the replicative qTR-PCR runs was determined from the ratio of input DNA concentration (determined by the UV spectrophotometer) and the absolute DNA copies calculated based on the CT value of the test sample in comparison with the DNA standards. The average for the three replicates was used to estimate the C-value for each of the DNA samples. The genome size in base pairs was then calculated using the formula, $\Gamma = C \times N_A \times M_{Bp}^{-1}$, where C is the C-value, N_A is Avogadro's number (6.022×10^{23} mol^{-1}) and M_{Bp} is the mean molar mass of a base pair (650 g mol^{-1}).

Table 3. Primer pairs for Lavender *Dxr* gene used for genome size estimation

Outer Primer**	Sequence (5'->3')	Length (bp)	Tm (°c)	GC%	Product Length (bp)
Forward primer	CAGCACCAAACAAATAGTGAGC	22	58.4	45.45	387
Reverse primer	GCTTGCCAGAAGGTGCTTTG	20	60.3	55	
Inner Primer	**Sequence (5'->3')**	**Length (bp)**	**Tm (°c)**	**GC%**	**Product Length (bp)**
Forward primer	TCGAGATACGGTCGAGAATAGGA	23	59.9	47.83	195
Reverse primer	ACGGGCAATTCTGTCGATGA	20	59.7	50	

** - Primer sequences and product sizes were based on cDNA sequence [JX630151.1].

Table 4. Primer pairs for Lavender *Hmgs* gene used for genome size estimation

Outer Primer**	Sequence (5'->3')	Length (bp)	Tm (°c)	GC%	Product Length (bp)
Forward primer	GCTTAGGGCGAGTCACATGG	20	65.8	60.00	343
Reverse primer	TTGGTAGCTTTCCTCGTTGGTT	22	65.3	45.45	
Inner Primer	**Sequence (5'->3')**	**Length (bp)**	**Tm (°c)**	**GC%**	**Product Length (bp)**
Forward primer	CCCGACCTTGCCAGTGAATA	20	65.0	55.00	141
Reverse primer	GGTATCTGAGACTGAGAACTGCT	23	64.3	47.83	

** - Primer sequences and product sizes were based on cDNA sequence [JX630154.1].

Table 5. Primer pairs for Arabidopsis Dxr gene used for genome size estimation

Outer Primer*	Sequence (5'->3')	Length (bp)	Tm (°c)	GC%	Product Length (bp)
Forward primer	TCAGTTTTGAGCATCTCAATGAAG	24	57.7	37.5	384
Reverse primer	TGGCTTGTTCGGATCACAGAT	21	59.7	47.62	
Inner Primer	**Sequence (5'->3')**	**Length (bp)**	**Tm (°c)**	**GC%**	**Product Length (bp)**
Forward primer	CAGTTTTGAGCATCTCAATGAAGTT	25	58.5	36.00	185
Reverse primer	ACGTGCCCACAAGTCATAGT	20	59.3	50.00	

* - Primer sequences designed from genomic DNA sequences [NC_003076.8].

Table 6. Primer pairs for Arabidopsis *Hmgs* gene used for genome size estimation

Outer Primer*	Sequence (5'->3')	Length (bp)	Tm (°c)	GC%	Product Length (bp)
Forward primer	CGAAGGTGTTGACTCGACCA	20	64.9	55.00	388
Reverse primer	TGCATTGAAGAGTTACCGGG	20	63.1	50.00	
Inner Primer	**Sequence (5'->3')**	**Length (bp)**	**Tm (°c)**	**GC%**	**Product Length (bp)**
Forward primer	AAGAGTCTTTGCATAATTCTTGACT	25	62.0	32.00	200
Reverse primer	CGCTAGCAAGATTGGGCTTG	20	62.0	55.00	

* - Primer sequences designed from genomic DNA sequences [NC_003075.7].

II.3.2 Genome size estimation using a K-mer counting method

For this computational method, the KmerGenie tool (version 1.6741) (Chikhi and Medvedev, 2014) was used to estimate the genome size based on the raw genome sequencing reads. In this case, we used the PE reads. The sequencing reads were provided as input to the KmerGenie tool via a text file labeled "list_files.txt" (Appendix B.1). KmerGenie was executed on SHARCNET and Compute Canada systems with default parameters to estimate the genome size by generating k-mer abundance histograms for a range of k-mer sizes from 21 to 121 bp. For each of the k-mer sizes, KmerGenie estimated the number of distinct k-mers and returned the best k-mer size and estimated genome size in bp. The output generated a plot for the best predicted k-mer value showing a clear fit and a global maximum.

II.4 *de novo* genome assembly

II.4.1 Optimal genome assembly tool selection

FASTQC (Andrews, 2016) (Appendix B.2) was used for verification of the quality of sequencing and for generating summary statistics for the sequencing data. *De novo* genome assembly was performed using the PE, MP and overlap reads, utilizing widely used genome assembly tools to select the optimal protocol for the lavender genome based on the genome assembly quality metrics. The tools tested include de Bruijn graph-based genome assembly tools, Abyss (Simpson et al., 2009) and Allpaths-LG (Gnerre et al., 2011), and the string graph-based tools, SGA (Simpson and Durbin, 2012) and FERMI (Li, 2012). Stand-alone scaffolders, OPERA-LG (Gao et al., 2011) and SSPACE (Boetzer and Pirovano, 2014), were also tested.

Based on the above testing, FERMI v1.1 (Li, 2012) (Appendix B.3) was used to generate contigs and OPERA v2.0 (Gao et al., 2011) tool (Appendix B.4.a, B.4.b) was used for

scaffolding to generate the draft genome assembly. Once the contigs were assembled, the contigs were filtered to retain only those which were greater than or equal to the minimal length of 500 bp. Post-assembly processing was performed using two transcriptome-based tools, L_RNA_scaffolder (Xue et al., 2013) (Appendix B.6) and our in-house developed RNA_Seq_scaffolding tool (Appendix B.7). Additionally, Gapcloser (Luo et al., 2012) (Appendix B.5) was used to close some gaps to generate a final draft version of the lavender genome assembly. The draft genome assembly was checked for the presence of viral or bacterial sequence contamination by running BLAT (Kent, 2002) against a known database of viral and bacterial genomes downloaded from the NCBI database.

II.4.2 Assessment of draft genome assembly quality

QUAST v4.3 (Gurevich et al., 2013) (Appendix B.8) was used to generate key metrics of genome assembly quality including number of contigs, the size of the largest contig, total length of the genome (in bp), N50 value, GC content and # of N's/100 Kb. The software, BUSCO - **B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs (Simão et al., 2015)(Appendix B.9) was used to estimate the completeness of the draft assembly by surveying the number and completeness in covering the highly conserved single-copy orthologues documented in OrthoDB_v9 (Kriventseva et al., 2015). In this case, the plant specific SCOs (single copy orthologues) of the OrthoDB_v9 were used.

II.4.3 GC profile characterization

The GC content profiling for lavender and other selected plant genomes was performed using an in-house Perl script, which calculates GC % of genome sequences in a sliding-window of 1000 bp and a step size of 500 bp. The genome sequences in fasta format for the mint genome

were obtained from (http://langelabtools.wsu.edu/mgr/downloads) and those for *Arabidopsis*, rice and maize were obtained from the Phytozome plant genomics resource database (https://phytozome.jgi.doe.gov/pz/portal.html).

II.5 *de novo* genome annotation

II.5.1 Automated *de novo* genome annotation

The draft assembly was first subjected to automatic annotation using the MAKER_P pipeline (Campbell et al., 2014) using the default settings. It uses Augustus (Stanke et al., 2006) for *ab initio* gene prediction along with the transcriptome data for evidence-based gene prediction, and plant reference protein sequence data for homology-based gene prediction. To reduce the memory and process time, the draft assembly was broken into 200 non-overlapping sequence chunks and executed in parallel on Compute Canada high performance computing facility (HPC) systems. The detailed commands and parameters used in running MAKER pipeline are provided in Appendices B.10.a & b.

II.5.2 Post annotation improvement

The GAG (Genome Annotation Generator) tool (Geib et al., 2018) (Appendix B.11.c) was used to generate a set of protein sequences in fasta format and annotation summary statistics based on the General Feature Format version 3 (GFF3) file from the MAKER pipeline. A custom in-house Perl script (Appendix B.11.a) was used to perform the first level of filtering on the protein sequences generated from the GAG tool, based on an open reading frame at a minimum of 30 a.a. in length.

The filtered proteins generated were used as input to run BLASTP (Appendix B.11.b) (e-value at $1e^{-20}$) analysis against NCBI plant Refprot sequences. A PlantRefprot database was

generated in the working directory and the protein sequences broken into smaller chunks were used as input to run BLASTP (e-value at $1e^{-20}$) (Altschul, 2005) command.

The predicted protein sequences with a match to the PlantRefProt sequences were subject to further validation using the Genevalidator (Dragan et al., 2016) tool (Appendix B.11.d). The tool identifies problematic gene predictions by comparing the input sequences to similar sequences in large public databases such as Swiss-Prot, NCBI nr database using BLAST (e-value at $1e^{-5}$). The resulting HSPs (High-scoring segment pairs) were analyzed for main traits such as length, coverage, presence of conserved regions and different genes, to identify high quality gene-models and protein-coding gene predictions. The selected candidates were retained for further downstream analysis.

The data from the last two steps were processed using another in-house Perl script (Appendix B.11.e) to generate a modified gff3 based on the positions of longest open reading frames (ORFs) and functional annotation based on the best matched protein sequences. The final step in the post-annotation improvement pipeline was to analyze the protein sequences using an in-house Perl script which uses protein sequences, draft genome sequences to process each scaffold and makes necessary adjustment to the gene models to satisfy sequence completeness and start and stop codons requirements, and it outputs the final coding sequence (CDS), protein and genomic sequence files and a gff3 file, which also covers the non-coding genes.

II.5.3 Annotation of non-coding and repeat elements

tRNA genes were identified using the tRNAscan program (Lowe and Eddy, 1996) (Appendix B.12.a), while rRNA genes were identified using RNAmmer (Lagesen et al., 2007) (Appendix B.12.b).

To identify the repeat elements, the two genomes (lavender and mint) were first processed with RepeatModeler tool (Herrmann et al., 2016) (Appendix B.13.a). The resulting *de novo* consensus repeats were used with RepeatMasker (Smit et al 1996) (Appendix B.13.b) running with default parameters to identify all repeat sequences in the genomes based on both the repeat consensus sequences from Repbase (Bao et al., 2015) and the outcome of RepeatModeler. In addition, RepeatProteinMasker, which identifies repeats based on repeat protein sequences, was used to identify additional repeat elements that might be missed by RepeatMasker. Bedtools (Quinlan and Hall, 2010; Quinlan, 2014) was used to merge the outputs of RepeatMasker and RepeatProteinMasker and generate a non-redundant list of repeats in each of the two genomes. Bedtools was also used to relate selected types of transposable elements (LTR elements *Gypsy* and *Copia*) with the annotated genes in the genome based on the genomic coordinates.

Miniature inverted-repeat transposable element identification was performed using the miteFinder II tool (Hu et al., 2018) (Appendix B.16) with default parameters for five plant genomes lavender, mint, rice, maize, and *Arabidopsis*.

II.6 Ploidy estimation

The ConPADE (Contig ploidy and allele dosage estimation) tool (Margarido and Heckerman, 2015) (Appendix B.15) with default parameter setting was used to determine the ploidy levels in the lavender genome. The input was the binary alignment (BAM) files consisting of alignment information of reads against the draft genome assembly. Ploidy was estimated separately for scaffolds containing EO genes and for all scaffolds (the entire genome assembly) (Appendix B.15).

II.6.1 Estimation of polyploidization events in the lavender genome

To estimate the occurrence of polyploidization events in the lavender genome, a pairwise aligner LASTZ was used to perform an all-against-all alignment for lavender CDS sequences and EO gene CDS sequences to generate pairwise alignments in axt format (alignment format for UCSC genome browser). The latter was used as input to the KaKs calculator tool (Zhang et al., 2006), which identifies the synonymous and non-synonymous substitutions in the CDS sequences. The estimation for the *Ka, Ks* and the ratio of *Ka*/*Ks* was done by the Nei and Gojobori method (Nei and Gojobori, 1986). The *Ks* (synonymous substitutions) values for both all CDS and EO CDS genes were plotted show the distribution of the *Ks* values. The *Ks* values were limited to 1.3, as higher *Ks* values are not reliable due to accumulation of substitutions in a non-neutral manner.

II.7 InterProScan and Gene Ontology (GO) analysis

The draft protein sequences identified from the annotation pipeline were used as input for the InterProScan tool (Quevillon et al., 2005) (Appendix B.14) using default parameters. The output in XML format included information such as GO term assignment, Interpro annotation, accession and other sequence characteristics. The XML file was parsed to generate RAW (InterProScan specific tsv format) and gff3 format files, which were used in downstream analysis (e.g., WEGO and protein annotation information).

GO analysis was performed using the online webserver WEGO (Ye et al., 2018) for genes overlapping with *Copia* and *Gypsy* elements to estimate the impact of the LTR elements in the lavender genome. The WEGO webserver uses RAW files generated from InterProScan as one of the input files. The genes impacted by *Copia* and *Gypsy* elements were first identified by position matching between their genomic coordinates and those of the genes with predicted

functions (i.e., with “Uncharacterized” and “Hypothetical” genes removed). The frequency of the GO terms and the related statistical measures were calculated and tabulated to identify the associated group of genes impacted by these two types of LTR elements.

II.8 Analysis of EO pathway genes in lavender and model plants

BLAST (Altschul, 2005) suite of tools was extensively used in the identification of EO pathway genes. In addition, Orthovenn (Wang et al., 2015) (https://orthovenn2.bioinfotoolkits.net/home), an orthologous clustering program, was utilized to identify terpene synthases (TPS) and isoprenoid biosynthetic genes in lavender and other model plant genomes. The mint genome was left out of this analysis due to the lack of publicly available genome annotation data (i.e., protein sequences).

II.8.1 Identification and comparison of TPS genes in lavender and model plants

The webserver TERZYME was used in the identification, classification and analysis of TPS genes in the lavender genome and other model plants. The input to the webserver (http://www.nipgr.res.in/terzyme.html) was protein sequences (in fasta format) for a genome. Since mint genome protein sequences were not available, it was excluded from the analysis. The webserver runs searches using Hidden Markov Model (HMM) to predict matching TPSs from the input sequences that are classified by their putative biochemical functions and by their gene family.

II.9 Computational analysis

Data analysis and figure plotting were performed using a combination of Linux shell scripts, R and Microsoft Excel. Most of the genome assembly, annotation and sequence analyses

were performed on SHARCNET (https://www.sharcnet.ca/) and Compute Canada high-performance computing facilities (http://computecanada.ca).

Chapter III. Results

III.1 The haploid Lavandula angustifolia genome size was estimated to be ~870Mbp

In light of the prior highly variable reports regarding the genome size of lavender ranging from 772 to 5,574 Mbp (Urwin et al., 2007; Urwin, 2014), we performed our own estimation of the genome size for *Lavandula angustifolia* (Maillette) using two very different methods. The first method is an experimental approach employing a quantitative PCR protocol (Wilhelm et al., 2003) to determine the weight of a haploid genome (C-value). It is achieved by obtaining the absolute number of genome copies in a DNA sample with known weight based on a standard curve using a known lavender DNA sequence at known copy numbers. As shown in Figure 3A, very similar results were obtained using two different single-copy genes as being 0.969 pg and 0.953 pg for *Dxr* and *Hmgs*, respectively. This gives an average of C-value (1C) for *Lavandula angustifolia* (Maillette) at 0.96 pg, which converts to 873Mbp in sequence length. The accuracy of the method was tested by determining the genome size of *Arabidopsis thaliana*, which gave an estimated C-value of 0.17 pg (157 Mbp in size) (Appendix/Table A1), matching its reported genome size at 0.17 pg data in the Plant DNA C-values Database based on the majority support of prior studies (Kew, 2008; Michael, 2014). The genome size of 873 Mbp for *Lavandula angustifolia* (Maillette) is closely matched by a computationally predicted size of 869 Mbp based on raw genome sequence reads using KmerGenie tool (Chikhi and Medvedev, 2014). This tool tests a range of k-mer values to generate a plot of k-mer size vs. the copy number to provide an optimum k-mer value and estimation of the genome size based on a clear concave curve and distinct global maximum in the plot (Figure 3B). These numbers also agree very closely with the size of our draft genome assembly at ~870 Mbp (Malli et al., 2019). Agreeing with the result of Urwin et al. (2007), which reports the genome size of *Lavandula angustifolia* to be between 772

to 880 Mbp using a flow-cytometry method, we believe that our estimation of the genome size at ~870 Mbp, supported by multiple methods, provides the best estimation for the genome size of *Lavandula angustifolia*, at least specifically for the Maillette cultivar, thus resolved the outstanding dispute on its genome size.

A.

Gene	C-value (pg)	Size (bp)
DXR	0.970	8.76E+08
HMGS	0.953	8.70E+08
Average	0.967	8.73E+08

B.

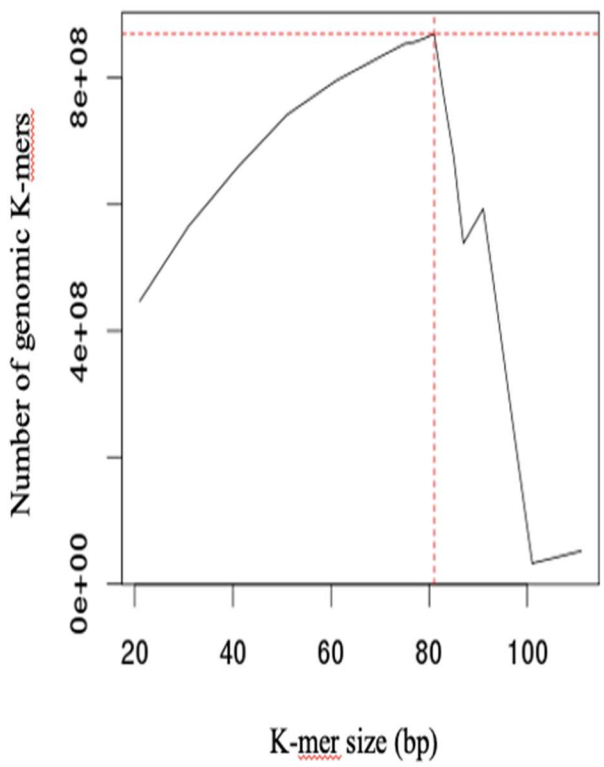


Figure 3. Determination of lavender genome size

A. A table showing lavender genome size estimations using the qPCR method (Wilhelm et al., 2003) based on two single copy genes, DXR (1- deoxy-D-xylulose 5-phosphate reductoisomerase) and HMGS (Hydroxymethylglutaryl-CoA synthase). B. Genome size prediction using KmerGenie (Chikhi and Medvedev, 2014) based on the k-mer count of the raw genome sequencing reads, showing a value of 81 bp which converts to a genome size of 869,617,739 bp or ~870 Mbp.

III.2 *De novo* genome assembly of lavender

The genome of *L. angustifolia* (Maillette) was sequenced to a total of ~150 x coverage (Table 2) using the Illumina HiSeq 2000 platform combining the use of pair end and pair mate libraries, all sequenced at 100 bp x 2 setting. Additional sequencing reads such as overlap reads (A.F.3.B) were also used as input for the *de novo* genome assembly.

III.2.1 Contig assembly

The initial contig assembly was generated after extensive testing of stand-alone contig assemblers and of *de novo* genome assembly pipelines, which use one of the two main genome assembly algorithms, the de Bruijn and String Graph algorithms. Among the four different tools tested, two are String Graph based assemblers (String Graph Assembler (SGA) and Fermi), and two are de Bruijn based pipelines (Abyss and Allpaths-LG). The string graph-based assemblers performed better as compared to the de Bruijn based assemblers by generating close to two times the total length of the contigs (Fermi vs. Allpaths-LG) (Table 7). The cut-off for retaining contigs generated by contig assembly tools (to be used for scaffolding) was 50% of the estimated genome length (879,915,704 bp). The total length of the contigs generated by Fermi and SGA was around 67% and 55% of the above estimated genome length, respectively, compared to 49% for Abyss and 30% for Allpaths-LG. Although the N50 was better in Allpaths-LG, the length of the non-gap sequences was less than half of that from Fermi and SGA. With the same amount of input genomic sequences (~150X), including the overlap libraries (Figure A.F.3.B) required by Allpaths-LG, the Fermi assembler was able to utilize more genomic reads to generate the largest total contig length, which is more than double of that for Allpaths-LG (Table 7). Furthermore, the string graph-based programs (Fermi and SGA) provide contigs without gaps (#Ns/100 Kb),

while contigs from the de Bruijn graph-based pipelines (Abyss and Allpaths-LG) do contain gaps (Table 7). For this reason, the contigs generated by Fermi assembler was used in scaffolding and subsequently in generating the final draft genome assembly.

Table 7. Contig assembly quality comparisons across different assemblers

	SGA	**Fermi**	**Abyss**	**Allpaths-LG**
# contigs	334,714	332,999	246,605	87,387
Largest contig (kb)	26,954	45,181	61,761	59,650
Total length (bp)	485,753,506	589,672,940	432,183,514	268,578,043
N50	1945	2,686	2462	4,361
# Ns per 100 kb	0	0	124.45	69.03
non-gap (bp)	242,876,845	294,837,520	216,070,484	134,292,133

III.2.2 Scaffolding

The scaffolding process to generate the lavender draft genome assembly followed a process similar to the contig assembly generation by testing multiple stand-alone scaffolders, including OPERA (Gao et al., 2011), SSPACE (Boetzer et al., 2011) and pipelines including Abyss (Simpson et al., 2009), Allpaths-LG (Gnerre et al., 2011), SPADes (Bankevich et al., 2012) and SOAPDenovo (Luo et al., 2012). The minimal length required for retaining an assembled scaffold was 500 bp and the total length of the assembled scaffolds is closer to the estimated genome length of the assembly (879,915,704 bp). Based on the initial test of the scaffolding tools, stand-alone scaffolders like SSPACE and pipeline scaffolders from Abyss, SPADes and Allpaths-LG did not perform well (data not provided) with the total length of

scaffolds being lower than the cut-off (90% of the estimated length of the genome), and this led to our selection of OPERA (Optimal Paired-End Read Assembler) as the primary scaffolding tool to generate the initial lavender draft assembly.

The scaffolds generated using Fermi in combination with OPERA tested at different k-mer values (59 and 79 bp) generated the best draft assembly with the total length at 874 Mbp (OPERA with K-mer at 79 bp). Even though the number of scaffolds generated by SGA+OPERA (65,933) were fewer and the largest scaffold length is longer (908,388 kb) compared to the scaffolds from Fermi+OPERA (83,170 with the largest at 826,790 Kbp), the Fermi+OPERA generated the longest non-gap sequences at 437 Mbp and had the lowest number of Ns/100 Kb (32,557), which is close to 7000 Ns lower than SGA+OPERA (~39,500) (Table 8). The Abyss+OPERA protocol generated a smaller number of scaffolds (~66,000), but the resulting draft assembly was around 220 Mbp shorter as compared to Fermi+OPERA (with the K-mer value of 79) assembly (874 Mbp). The N50 for Abyss+OPERA was much lower (69,422 bp) compared to that of SGA+OPERA (98,925) and FERMI+OPERA (96,478) (Table 8).

For Allpaths-LG, since the contigs generated were not comparable to the string graph-based assemblers and Abyss (Table 7), it was not included in the scaffolding analysis. The draft scaffolds which were generated using FERMI+OPERA-k79 had better results in key parameters like the total length of the scaffolds, higher N50 values, lower number of Ns/100 kb and longer non-gap sequences (Table 8), when compared to SGA+OPERA and Abyss+OPERA. For this reason, FERMI+OPERA-k79 was used to generate the final draft genome.

Table 8. Scaffolding quality comparison using various tools

	# contigs	Largest scaffold (kb)	Total length (bp)	N50 (bp)	# Ns per 100 kb	non-gap (bp)
SGA+OPERA	65,933	908,388	802,944,424	98,925	39,503	401,483,297
Abyss_k61+OPERA_k59	66,259	596,388	653,521,560	69,422	34,838	326,795,053
Fermi+OPERA_k59*	85,714	808,828	881,318,095	95,234	33,093	440,669,479
Fermi+OPERA_k79**	83,170	826,790	874,275,510	96,478	32,557	437,196,158

*- refers to the scaffolder OPERA run with the K-mer size of 59 and
** - refers to the scaffolder OPERA run with the K-mer size of 79

III.2.3 Genome assembly improvements using various tools and genomic data

The draft assembly generated using Fermi+OPERA_k79 was subjected to a few post-assembly processes to increase the assembly quality. Specifically, we used a tool called L_RNA scaffolder (Xue et al., 2013), which uses the transcriptome reads to order, orient and combine genomic fragments or contigs into larger scaffolds. The resulting genome assembly is larger in length and has fewer scaffolds due to the merging of genomic sequences (Table 9). This assembly was subjected to a further process using an in-house tool, RDNA_scaffolder, which also uses transcriptome data to merge fragmented genome scaffolds and fill gaps, resulting in an increase of the total length of the assembled genome, larger scaffolds, improvements in the N50 statistics, reduced gaps and length of the non-gap sequences. The resulting assembly was further improved by running GapCloser (Luo et al., 2015), which resulted in a better assembly with longer genome assembly length and non-gap sequences. The final draft assembly generated using FERMI (Li, 2012) for contig assembly and OPERA (Gao et al., 2011, 2016) for

scaffolding contains 869,786,077 bp in 84,291 scaffolds with N50 being 96,735 bp and the total length of the non-gap sequences (sequences with no gaps) being 688,040,719 bp or 79.1 % of the final draft genome (Table 9).

Table 9. Genome assembly statistics at different stages of the assembly

Metrics	**Fermi + OPERA scaffolder**	**Fermi+OPERA+LRNA/ RDNA_scaffolder**	**Final draft***
Number of scaffolds	83,170	84,291	84,291
Largest scaffold (bp)	826,790	944,633	942,053
Total sequence length (bp)	874,275,510	875,514,707	869,786,077
N50 (bps)	96,478	97,529	96,735
# Ns (bp/100 kb)	32,557	32,506	20,878
Non-gap (bp)	589,389,676	590,665,403	688,040,719
GC%	38.05	38.06	38.01
BUSCO coverage (%)	85.97%	90.40%	91.08%

*: gap-filled sequences using GapCloser tool.

III.3. Characterization of the lavender genome

III.3.1 Gene annotation

Using an automated and highly configurable *de novo* genome annotation pipeline MAKER (Campbell et al., 2014) with an in-house lavender transcriptome assembly data generated in our lab using Trinity (Haas et al., 2013) with the RNA-Seq data from Dr. Soheil Mahmoud's lab and additional datasets from NCBI Sequence Read Archive (SRA) database, as well as NCBI plant reference protein sequences to guide gene prediction for Augustus (Stanke et

al. 2006), we identified 218,000 initial gene models contained in gff3 files. These initial gene models were subjected to Genome Annotation Generator GAG (Geib et al., 2018) to generate a draft annotation gff3 file containing functional annotation information and also translated protein sequences. The resulting protein sequences were then filtered by using a custom Perl script (Process_orf.pl) (Appendix B.11.a) based on a minimal sequence length of 30 a.a. This led to 191,951 protein sequences, which were then subjected to BLASTP search against NCBI non-redundant protein sequences database. Among these, 92,450 sequences had a match to an existing protein sequence in the NCBI GenPept database. To check the validity of these protein sequences, a sequence evaluating tool called Genevalidator was used, which further reduced the number of reliable protein coding sequences to 63,672. A custom Perl script (Generate_final_files.pl) (Appendix B.11.e) was then used to process the whole lavender genome, one scaffold at a time, to trim and filter the protein sequences based on alignment to other reference protein sequences and generate the cDNA and protein sequences and a gff3 annotation file for a final list of 60,819 protein coding sequences (Table 10). This script makes adjustments to the start of the protein sequence and determines the sequence completeness based on the alignment with known protein sequences. In addition, 2,129 tRNA and rRNA (1512 tRNA genes and 617 rRNA genes) genes were identified using tRNAScan (Lowe and Eddy, 1996) and RNAmmer (Lagesen et al., 2007), respectively, and their annotation information was added to the above gff3 file. Altogether, a total of 62,948 genes consisting of 60,819 protein coding genes, 1,512 tRNA genes, and 617 rRNA genes, were identified for the lavender draft genome (Table 10).

The total regions of repeat elements (45.5%) and gaps (20.9%) cover close to ~66% of the lavender genome (Figure 4). In comparison, the lavender genome has a higher (~3 times)

density of protein coding sequences (CDS, 9.2%) than the maize genome (CDS, 3.5%), likely due to the higher repeat content in the latter, but it is more than four times lower than that of *Arabidopsis* (CDS, 36.2%) (Table 11). The lavender genome has the lowest non-CDS exons (0.5%) while *Arabidopsis thaliana* has the highest (9.8%), likely due to lack of sufficient annotation of the untranslated regions (UTRs) for the lavender genes.

Table 10. Number of genes at different stages of gene annotation in the lavender draft genome

***De novo* Annotation Process**	**No. of Genes**
Maker Annotation Pipeline (*De novo* annotation software)	218,091
Process_ORF.pl (script for filtering ORF sequences)	191,951
GAG and BLASTP Filtering (Genome annotation generator)	92,450
Genevalidator (checking protein sequences)	63,672
Generate_FinalFiles.pl (generating final annotation files)	60,819
Non-coding genes identification (tRNA & rRNA genes)	2,129
Total	62,948

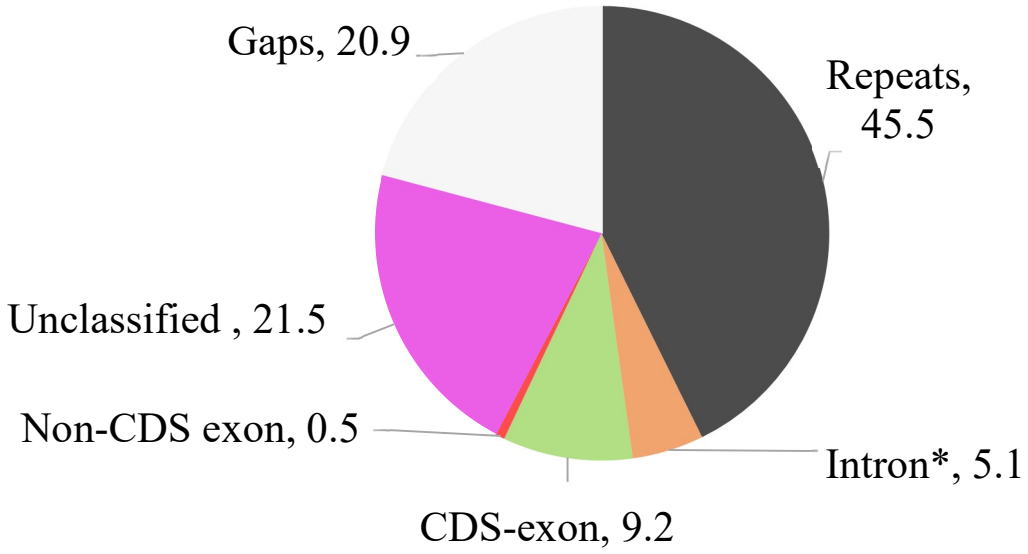


Figure 4. Pie-chart showing the genome composition.

Categories denoted by "*" exclude overlapping repeats and gaps.

Table 11. Comparison of genomic features for lavender with two model plants

	L. ang		***A. tha***		***Z. may***	
Genomic region	**Mbp**	**%**	**Mbp**	**%**	**Mbp**	**%**
CDS-exons	80.3	9.2	43.3	36.2	72.9	3.5
Non-CDS exons	4.4	0.5	11.8	9.8	53.1	2.5
Introns*	44.2	5.1	28.2	23.5	100.2	4.8
Repeats	372.4	45.5	24.9	20.8	1729.2	82.2
Unclassified	186.8	21.5	11.3	9.4	118.2	5.6
Gaps	181.7	20.9	0.2	0.2	30.7	1.5
Total	869.8	100	119.7	100	2104.4	100

, excluding overlapping repeats and gaps; L. ang*, *Lavandula angustifolia*; *A. tha*, *Arabidopsis thaliana*; *Z.may*, *Zea mays*

As a summary shown in Table 12, the final lavender draft genome assembly is 869 Mbp in total length in 84,291 scaffolds, with the N50 being 96,735 bp and with 688 Mbp or 79% being non-gap sequences and 181 Mbp as gap regions. The genome completeness was measured to be ~92% by BUSCO, covering 1323 (1292 complete and 31 fragmented) SCOs. The GC content of the genome was measured to be at 38.1%, and the genome contains 60,819 protein coding genes and 2129 non-coding (1512 tRNA and 617 rRNA) genes. Known repeat elements contribute to 45.5 % of the total genome assembly or ~57% of the non-gap genome sequences.

To examine the biological characteristics of lavender genome, we analyzed the 60,819 lavender protein coding genes using InterProScan. Approximately 80% of the annotated protein sequences (~49,200) received one or more IPR code (a code assigned by the InterProScan program to protein sequences) assignments (Table A2). A similar analysis was done to check the numbers of genes being assigned a specific IPR code (Table 13). Ranking by the number of genes involved from high to low, the first on the list was the IPR code IPR013103 for "Reverse transcriptase, RNA-dependent DNA polymerase" with more than ~9100 genes. Furthermore, more than 30% of the top ten IPR codes were related to transposases and reverse transcriptase, indicating that the lavender genome has a large number of genes derived from retrotransposons.

Table 12. Summary statistics for the lavender genome assembly

Metrics	Values
Total genome sequence length (bp) #	869,786,077
Gap length (bp)	181,745,358
Non-gap sequence (bp)	688,040,719
Number of scaffolds	84,291
N50 (bp)	96,735
Largest scaffold (bp)	942,053
GC%	38.1
Number of protein-coding genes	60,819
Number of non-coding genes (tRNA/rRNA)	2,129
Total number of genes	62,822
BUSCO coverage (%)	91.8
Genome covered by genes	26.8%
Genome covered by CDS	9.2%
Genome coverage by annotated repeats	45.5%

Table 13. InterProScan analysis of the lavender genes

IPR Code	IRP Description	Gene Count
IPR013103	**Reverse transcriptase, RNA-dependent DNA polymerase***	**9114**
IPR012337	Ribonuclease H-like domain	7734
IPR001584	**Integrase, catalytic core**	**6191**
IPR025724	**GAG-pre-integrase domain**	**3635**
IPR001878	Zinc finger, CCHC-type	3102
IPR027417	P-loop containing nucleoside triphosphate hydrolase	1752
IPR011009	Protein kinase-like domain	1551
IPR000719	Protein kinase domain	1471
IPR005162	**Retrotransposon gag domain**	**1377**
IPR000477	**Reverse transcriptase domain**	**1211**
IPR008271	Serine/threonine-protein kinase, active site	1093
IPR032675	Leucine-rich repeat domain, L domain-like	930
IPR017441	Protein kinase, ATP binding site	881
IPR013083	Zinc finger, RING/FYVE/PHD-type	868
IPR011990	Tetratricopeptide-like helical domain	767
IPR021109	Aspartic peptidase domain	688
IPR013320	Concanavalin A-like lectin/glucanase domain	686
IPR009057	Homeodomain-like	651
IPR002885	Pentatricopeptide repeat	646

*, categories related to retrotransposons are indicated in bold font.

III.3.2 Genome assembly completeness assessment based on gene content

Once the initial draft assembly was generated, an assessment of the draft assembly was done to assess the completeness of the assembled genome. The assembly quality statistics provided by tools such as QUAST (Gurevich et al., 2013) give out information like N50, total length, number of scaffolds and other key details (Table 9), but lack a critical parameter, which is the gene content in the newly assembled genome. This information is determined by quantifying the presence of conserved single copy orthologues (SCO) in genomes using BUSCO v2 (Simão et al., 2015) in comparison with four other published plant genomes including *Mentha longifolia*, which was the closest genome to lavender at the time of analysis (Figure 5). The total number of *Plant* (embryophyta) SCOs tested was 1440 from the OrthoDB database v9. The key parameters analyzed include the numbers of complete single-copy, complete duplicate, fragmented and missing SCOs present in the genome.

As shown in Figure 5, the lavender genome has a completeness level of ~92%, covering 1292 of the 1440 SCOs with 596 being complete single-copy, 696 being duplicated, 118 being fragmented SCOs (F) and missing 30 SCOs (M). Based on SCO coverage, the completeness of our lavender genome draft assembly is much better than that of the closely related genome of *Mentha longifolia*, which has the highest number of missing (M, 409) and fragmented (F,183) SCOs and the lowest number of complete SCOs (C, 848). Lavender genome is comparable to that of the maize genome (~92% with 31 F and 115 M), which was generated and improved using much more resources. As expected, the model plant genomes, *Arabidopsis thaliana* (98%) and *Oryza sativa* (98%) have a higher level of genome completeness, likely attributed to the much more extensive data and resources used (Chang et al., 2016).

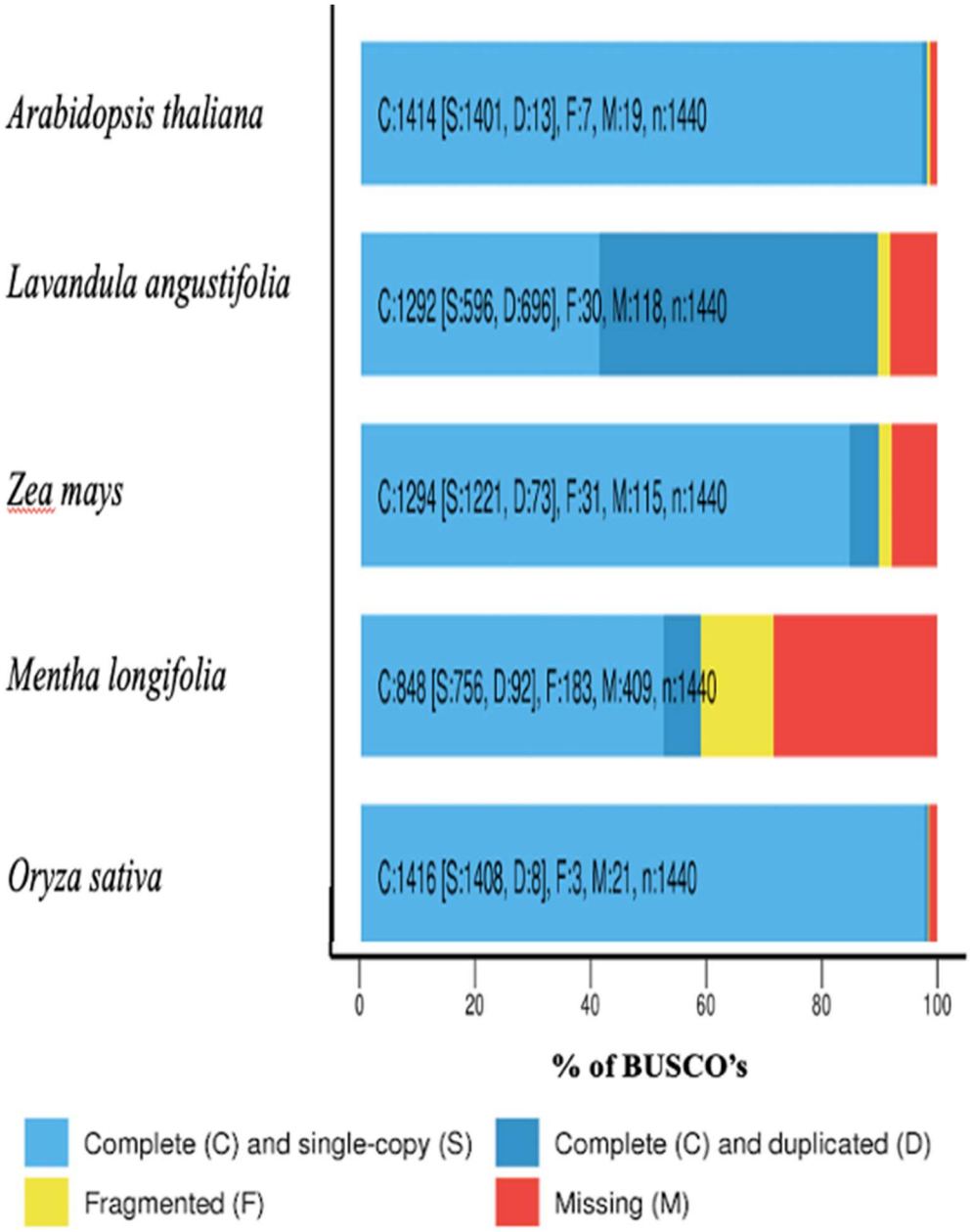


Figure 5. Assessment of genome completeness of the lavender draft genome in comparison with other five published plant genomes using the Benchmark Universal Single Copy Orthologues (BUSCO).

We further analyzed the genome completeness using a tool called DOGMA (Domain-based General Measure for transcriptome and proteome quality Assessment) (Dohmen et al.,

2016) based on the presence of three different conserved domain arrangements (CDAs) in the genome against expected CDAs. Although lavender protein sequences were identified solely based on the *de novo* genome annotation, the quality of the lavender protein sequences is almost 82 % complete and is only 9 to 10 % lower than rice and maize (Table 14). These two model plant genomes (rice and maize) have been studied and analyzed and improved multiple times with additional sequencing and transcriptome data. As expected, the *Arabidopsis* genome, as the first plant genome to be sequenced and a model plant with a small genome that has been studied extensively, has the best completeness score at 99.49% (Table 14).

Table 14. Comparison of lavender protein sequences against conserved domain arrangements (CDA) in model plant proteomes using DOGMA.

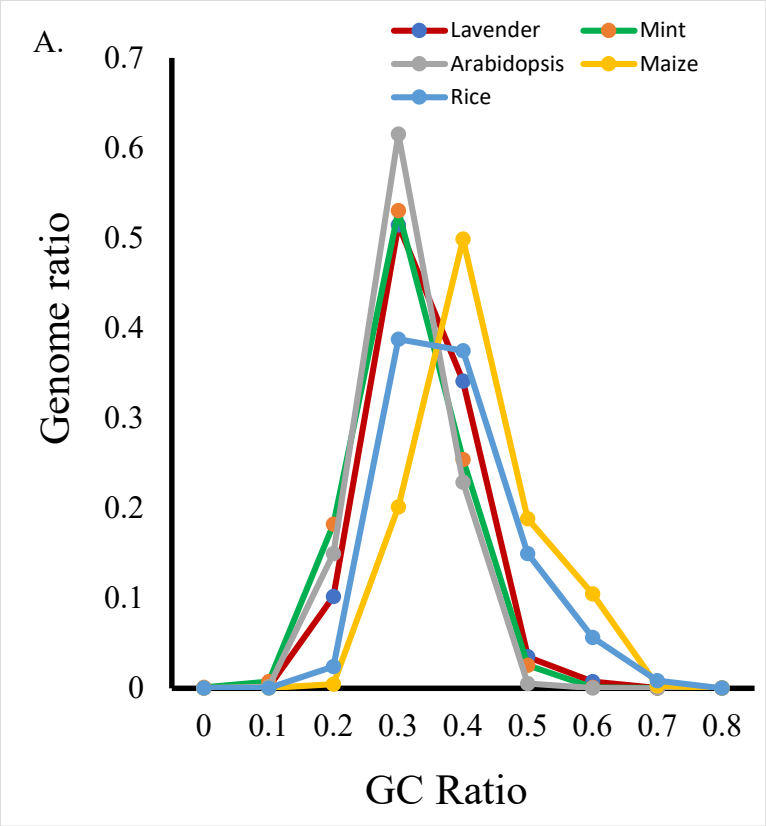
	Plant genome	***L.ang****	***A.th***	***O.sat***	***Z.may***
CDA size 1	Found	704	834	787	796
	Expected		835		
CDA size 2	Found	310	389	341	342
	Expected		392		
CDA size 3	Found	109	136	106	114
	Expected		139		
Total	Found	1,123	1,359	1,234	1,252
	Expected		1,366		
	completeness (%)	**82.21**	**99.49**	**90.34**	**91.65**

**L. ang, Lavandula angustifolia*; *A. tha, Arabidopsis thaliana; O. sat, Oryza sativa*; *Z. may, Zea mays*.

III.3.3 GC content of the lavender genome

The GC profile of the lavender genome was analyzed and compared with that of mint (*Mentha longifolia*) and three other model plant genomes (both dicot and monocot sp.). The GC % of maize is the highest (47.4 GC%), followed by rice (43.6 GC%), and then by lavender and mint (~ 38% GC %), with the *Arabidopsis* genome having the lowest GC content (36.0% GC) (Figure 6B). As expected, among all the genomes compared, lavender (38.1 GC %) has a GC profile similar to that of the mint genome (38.0 GC%). The GC distribution profile of lavender and mint is similar at the peak, but some minor differences are also observed, with the lavender genome having slightly higher portion at 40% GC content and lower at 20% in comparison with the mint genome (Figure 6A). Some differences are evident in the GC% distribution profile of the four genomes with the rice and maize genome having a very different profile as compared to all other plants (*Arabidopsis*, lavender, and mint) (Figure 6A) with the maize genome showing GC peak at ~40% and the rice genome having a flatter peak between 30 -40%. The *Arabidopsis* has the same peak at 30% GC with lavender and mint genome but higher genome percentage (>60%) (Figure 6A).

A.



B.

Plant Name	GC%
Arabidopsis	36.0
Rice	43.6
Maize	47.4
Lavender	38.1
Mint	38.0

Figure 6. Comparative analysis of GC content for lavender and four other plant genomes. A. GC profile of the lavender genome in comparison with four other genomes. B. The GC% values for various plant genomes.

III.4 Lavender genome has distinctive features optimized for essential oil production

III.4.1 Comparative analysis of essential oil pathway genes

Since lavender is valued for its high capacity for producing EOs, we performed detailed analysis of genes involved in the EO pathways. In this regard, we focused on genes involved in two main EO-related pathways, the methylerythritol pathway and the mevalonate pathway, as well as prenyltransferases to survey the copy number of genes in these pathways in comparison with other plants (Table 15).

One copy each of the *Dxr* and *Mct* genes, both involved in the MEP pathway, is found for all five plant genomes analyzed here except for maize, which has multiple copies. Interestingly, for all other genes in this pathway (*Dxs*, *Dpmdk*, *Mcs*, *Hds*, and *Hdr*), lavender has more copies than all other genomes. For *Dxs*, both lavender and mint have a much larger copy number (13 for lavender and 12 for mint) than all non-mint plants (the other three model plants having no more than two copies of *Dxs* gene), likely reflecting the importance of this gene in EO production. Furthermore, lavender is the only plant among the list to have multiple copies of *Hds* (4 copies) and *Hdr* (7 copies) with all other four plants including mint having only one copy, likely related to the unique aspect of EO production in lavender that is different from mint. For genes in the MVA pathway, the two mint-family plants do not seem to have a higher copy number than the three non-mint-family plants. Nevertheless, lavender does have a slightly higher copy number for *Mpdc* and *Ipp* than do the other four plants. The fact that, in most cases, lavender has a slightly larger copy number for these genes than mint may be a reflection of better genome sequence completeness. It is interesting to notice that the maize genome has the highest copy number for the *Hmgr* gene with nine copies, while lavender has six copies as the second highest. For farnesyl pyrophosphate synthase (*Fpps*), which is a precursor of sesquiterpenes, lavender has four copies, while all others have only one or two copies. An interesting observation is the presence of nine copies of *Ggpps* in *Arabidopsis* genome, while lavender has eight copies with the remaining three genomes having no more than five copies (Table 15). In summary, the fact that the lavender genome uniquely has a larger copy number for many of these EO genes may the critical contributing factor to its uniquely high efficiency in EO production.

Table 15. A comparison of gene copy numbers for the MEP, MVA pathways, and prenyltransferases in lavender and other plant genomes.

Gene ID	*L. ang*	*M. long*	*A. tha*	*Z. may*	*O. sat*
Methylerythritol pathway (MEP)					
Dxs	13	12	1	2	2
Dxr	1	1	1	3	1
Mct	1	1	1	2	1
Dpmdk	2	1	1	1	1
Mcs	3	1	1	2	2
Hds	4	1	1	1	1
Hdr	7	1	1	1	1
Mevalonate pathway (MVA)					
Aact	3	2	2	3	4
Hmgs	2	2	1	4	3
Hmgr	6	3	2	9	2
Mk	4	2	1	5	1
Pmk	2	2	1	1	1
Mpdc	3	1	2	1	1
Ipp	4	1	2	3	2
Prenyltransferase					
Gpps.SSU	1	1	0	0	0
Gpps.LSU	2	2	0	0	1
Lpps	1	1	0	0	0
Fpps	4	1	2	1	1
Ggpps	8	5	9	3	2
Sqs	1	2	2	2	2

L. ang, *Lavandula angustifolia*; *M.long*, *Mentha longifolia; A. tha*, *Arabidopsis thaliana*; *Z.may*, *Zea mays; O.sat, Oryza sativa*

III.4.2 Comparative analysis of TPS genes

The TPS gene family is responsible for generating a large class of plant secondary metabolites derived from 5-carbon isoprenoids produced by two biosynthetic pathways, MVA and MEP pathways (Figure 1). Based on function and product classification, lavender TPS genes can be classified into three main classes, which are monoterpene synthases, diterpene synthases and sesquiterpene synthases. An initial identification of these three classes of genes in the lavender draft genome resulted in a total of 56 putative genes, among which 24 are monoterpenes, 16 are diterpenes, and the remaining 16 are sesquiterpenes (Table 16). The total number of TPS genes in lavender is more or less similar to those of maize and rice and higher than that of *A. thaliana*, likely due to the smaller genome of the latter. However, lavender is unique in having the largest number of genes for monoterpenes 24 vs. 11 in *A. thaliana*, which is the second largest. In this regard, *A. thaliana* stands out by having the smallest number of genes for diterpenes, while rice stands out by having just a single copy of the gene for monoterpenes (Table 16).

Table 16. Distribution of TPS functionally classified genes in various plant genomes

TPS gene class	***L. ang****	***A. tha***	***O. sat***	***Z. may***
Monoterpene	24	11	1	10
Diterpene	16	4	27	17
Sesquiterpene	16	25	39	29
Total	56	40	67	56

**L. ang*, *Lavandula angustifolia*; *M. long*, *A. tha*, *Arabidopsis thaliana*; *Z. may*, *Zea mays; O. sat, Oryza sativa*

According to a recent study (Chen et al., 2011), the identification of the TPS genes can be extended and classified into eight sub-families, TPS-a to TPS-h based on the sequence properties and functional characteristics, which have been studied and recognized in a wide range of plants (Kaul et al., 2000; Yu et al., 2002; Jaillon et al., 2007). To perform a similar analysis, we analyzed the lavender draft genome using the TERZYME TPS gene finder (Priya et al., 2018) and identified a total of 2,898 TPS genes, which is far more than the other model plants (Table 17). The mint genome was not analyzed due to the lack of protein sequences which is required as input for identifying TPS genes. The number of putative TPS-a genes identified in lavender is 2,862, which is ~100 times higher than the sum found for *Arabidopsis*, rice and maize. Even among the TPS-b genes, lavender had the highest number (17) in comparison to other model plants (Table 17). Among the three model plants, the rice genome had the highest total of 59 TPS gene family with TPS-a and TPS-f having higher numbers than *Arabidopsis* and maize genome (Table 17).

Table 17. Distribution of TPS genes based on subfamilies classification in various plant genomes

TPS* gene family	*L.ang*	*A.tha*	*O.sat*	*Z.may*
TPS-a	2862	29	36	23
TPS-b	17	6	1	6
TPS-c	7	2	5	6
TPS-f	10	2	15	7
TPS-g	2	1	2	4
Total	2898	40	59	46

*Terpene Synthase; *L. ang*, *Lavandula angustifolia*; *A. tha*, *Arabidopsis thaliana*; *Z.may*, *Zea mays; O.sat, Oryza sativa*

III.4.3 Comparison of orthologous genes among different plant genomes

We also performed a search for orthologous sequences/clusters search as a way of validating the genome annotation using the Orthovenn tool (Wang et al., 2015). For a total of 60,819 genes in lavender, orthologous sequences can be found in the *Arabidopsis* (1,242), rice (106), maize (112) genomes (Figure 7), and these genes form a total of 8,571 clusters involving 71,763 genes from the four species. Additionally, ~700 genes are shared between lavender, *Arabidopsis*, rice and maize genomes: 473 genes are shared among lavender, rice, and maize genomes (we were unable to include mint genome for this analysis due to lack of protein sequences). A search for orthologous sequences containing terpene synthase genes resulted in 20 orthologous clusters divided into characterized and putative TPS genes. We also checked the conservation of all known TPS genes found based on the presence and absence of their orthologous sequences in these genomes (Table 18). *S*-linalool synthase gene is present in all

these genomes, mostly with multiple copies. Interestingly, the *R*-linalool synthase is absent in rice and maize but has two copies in lavender (Table 18). Among the putative TPS genes, the genes for ent-copalyl diphosphate synthase and solanesyl diphosphate synthase are shared among all of the genomes. However, the other three genes (cis-abienol synthase, gamma-cadinene synthase and (+)-epi-alpha-bisabolol synthase) seem to be uniquely present in lavender (Table 18). These TPS genes may also contribute to the unique aspects of EO production in lavender.

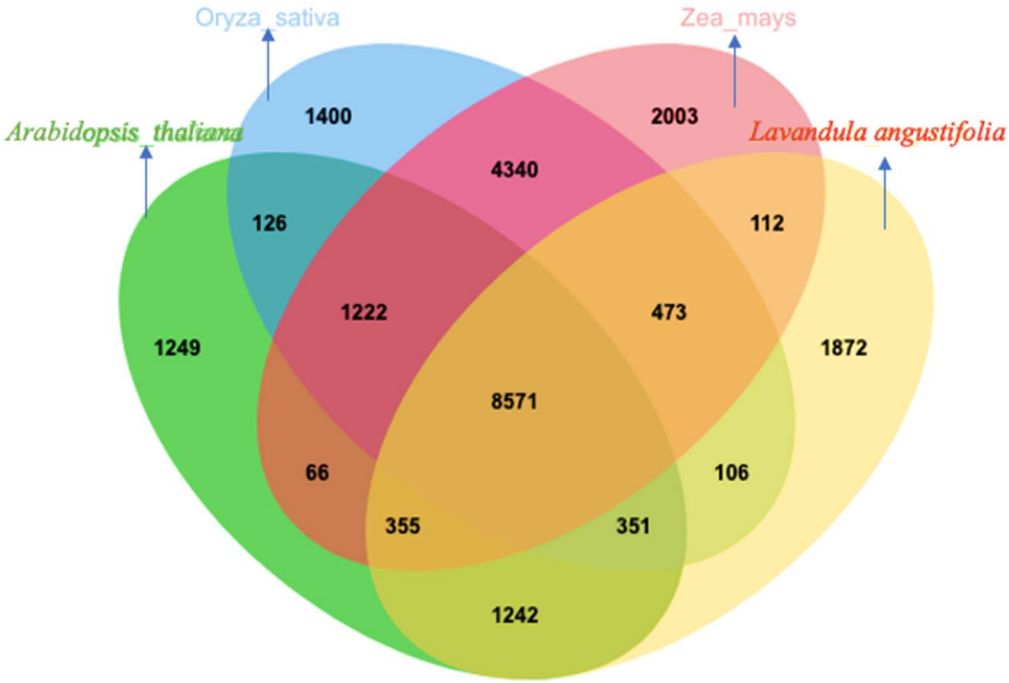


Figure 7. Analysis of genes sharing among lavender and other plant genomes.

Analysis of key orthologous genes using the Orthovenn 2 (Wang et al., 2015) platform which performs identification and annotation of orthologous sequences among multiple species generating an informative Venn diagram as output for further analysis. The required input are the protein sequences of the species to be analyzed.

Table 18. Monoterpene synthase, sesquiterpenes synthase, and acetyltransferase genes responsible for producing mono- and sesquiterpene essential oil constituents in lavender and other model plants.

TPS gene/product Name	*L. ang*[*]	*A. tha*	*O. sat*	*Z. may*
R-Linalool synthase	2	1	0	0
S-Linalool synthase	2	1	2	2
Limonene synthase	4	0	0	0
1,8-Cineole synthase	15	5	0	0
Boronyl diphosphate synthase	3	0	0	0
Beta-phellandrene synthase	5	0	0	0
τ-cadinol synthase	0	0	0	0
Beta-caryophyllene synthase	6	1	1	0
Germacrene synthase	3	0	0	0
Trans-alpha bergamotene synthase	2	0	0	0
3-carene synthase	0	0	0	0
Borneol dehydrogenase	0	0	0	0
9-epi--caryophyllene synthase	0	0	0	0
Alcohol acetyl transferases 1	0	0	0	0
Alcohol acetyl transferases 2	3	0	0	0
Putative TPS genes				
Ent-copalyl diphosphate synthase	7	1	1	2
cis-abienol synthase	3	0	0	0
(+)-epi-alpha-bisabolol synthase	2	0	0	0
gamma-cadinene synthase	4	0	0	0
Solanesyl diphosphate synthase	5	2	1	2

[*]*L. ang*, *Lavandula angustifolia*; *M. long*, *Mentha longifolia; A. tha*, *Arabidopsis thaliana*; *O. sat, Oryza sativa; Z. may*, *Zea mays*

III.4.4 Lavender has a history of genome polyploidization favouring essential oil genes

III.4.4.1 Confirmation of genome polyploidization in the lavender genome

Since a history of polyploidization during the evolution of lavender was suggested for the lavender genome (Urwin, 2014), we performed analysis of the genome for detection of possible historical genome polyploidization events. For this, we analyzed the draft genome assembly using the ConPADE tool (Margarido and Heckerman, 2015), a probabilistic method that estimates the ploidy of any given contig/scaffold based on its allele proportions. The analysis was performed for the entire genome and for the scaffolds containing known EO genes as a subset of the genome. As shown in Figure 8, for the entire genome, 28%, 18%, 11%, and 43% of the scaffolds represent levels 1, 2, 3, and 4 of polyploidization. Level 1 represents haploid regions due to lack of genetic variation (homozygous) or incorrect collapse of the highly homologous regions during genome assembly or low sequence coverage. Level 2 represents diploid regions, while levels 3 and 4 represent polyploidy genome regions which can be a reflection of repetitive regions due to the presence of long transposable elements or truly duplicated genomic regions. The fact that more than half (54%) of scaffolds were classified as polyploid by this method cannot be entirely explained by the presence of repetitive elements as the total repeat region is less 45% of the genome. This means that at least part of these polyploidy regions represents true polyploidy regions. Very interestingly, 95% of the EO scaffolds are classified as level 4 (tetraploidy), making a striking contrast with the ploidy classification profile for the whole genome (Figure 8), indicating that the gene duplication strongly favours EO genes. This agrees with the observation that many EO genes in the lavender genome have a much larger copy number than their counter parts in other plant genomes (Malli et al., 2019). This bias can be a result of two mechanisms, i.e., a whole genome duplication

followed by selective gene loss, which favoured EO genes retention, or by local gene duplication favouring EO genes or a combination of the two.

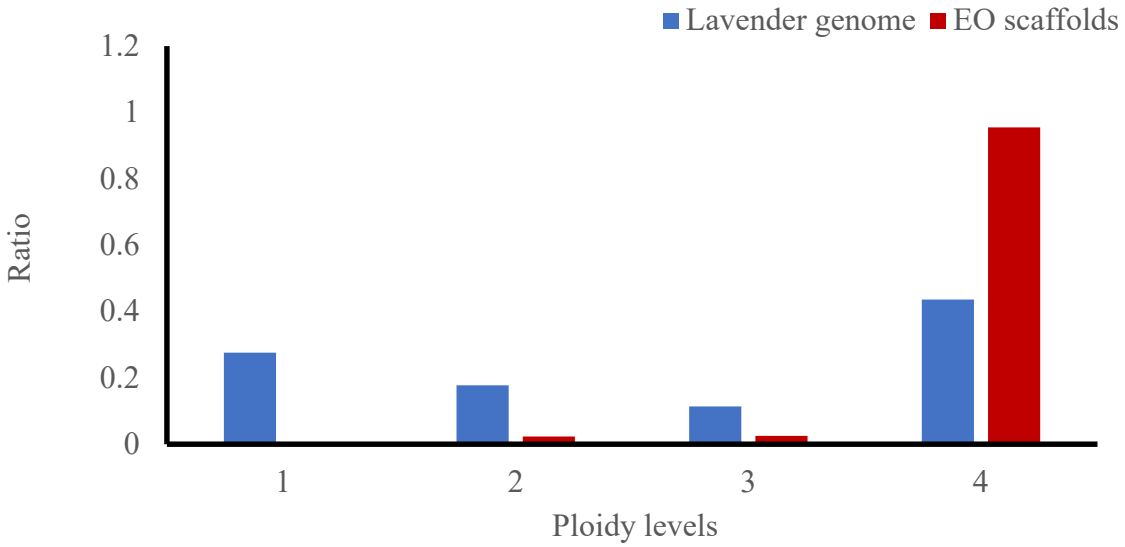


Figure 8. Estimation of ploidy levels in the lavender draft genome.

Ploidy estimation in lavender genome was performed with ConPADE (Margarido and Heckerman, 2015), showing the proportion of the whole genome and EO scaffolds at different levels of polyploidy.

III.4.4.2 Comparative analysis of polyploidization events in lavender genome.

Knowing that a significant proportion of the lavender is predicted to be polyploid, it would be interesting to determine the age profile for the genome duplication events and compare this side-by-side with the genome of mint, since lavender is as a close relative of the mint family (Li et al., 2017). Such analysis may indicate how much duplication is shared in the common ancestor of the two species and how much is unique to lavender. To estimate the age profile of the duplication events, synonymous nucleotide substitutions per synonymous site (*Ks*)

divergence values were calculated for the duplicate gene pairs in lavender and mint using the lastz tool (Marschinke and Strömberg, 2008).

As shown in Figure 9, the *Ks* plots for the lavender and mint plants show an approximate L-shape with a high initial peak at the low *Ks* value representing very recent gene duplications, and lower secondary peaks at higher *Ks* values resulted from older/ancient burst of gene duplication (e.g., polyploidization), and the flat tail at higher end of the *Ks* value representing pairs of duplicated genes under selective constraint (Blanc and Wolfe, 2004).

The *Ks* plot showed clear secondary peaks around *Ks* value at 0.2, and 0.5 for both genomes (Figure 9). The peak height for *Ks* = 0.5 in mint was close to double of that for lavender (0.1 vs. 0.2), while the peak at *Ks* = 0.2 in mint is slightly lower than that of lavender, and the initial peak is much lower than that in the lavender. Based on the *Ks* value of the two peaks, the estimated timing of polyploidization events in lavender and mint genomes was between 16.6 (*Ks* = 0.2) to 40.6 Mya (*Ks* = 0.5). Furthermore, it seems to indicate that lavender genome is more active than the mint genome by having a much higher level of recent gene duplications, as well as greater gene loss from the ancient genome duplication (*Ks*=0.5).

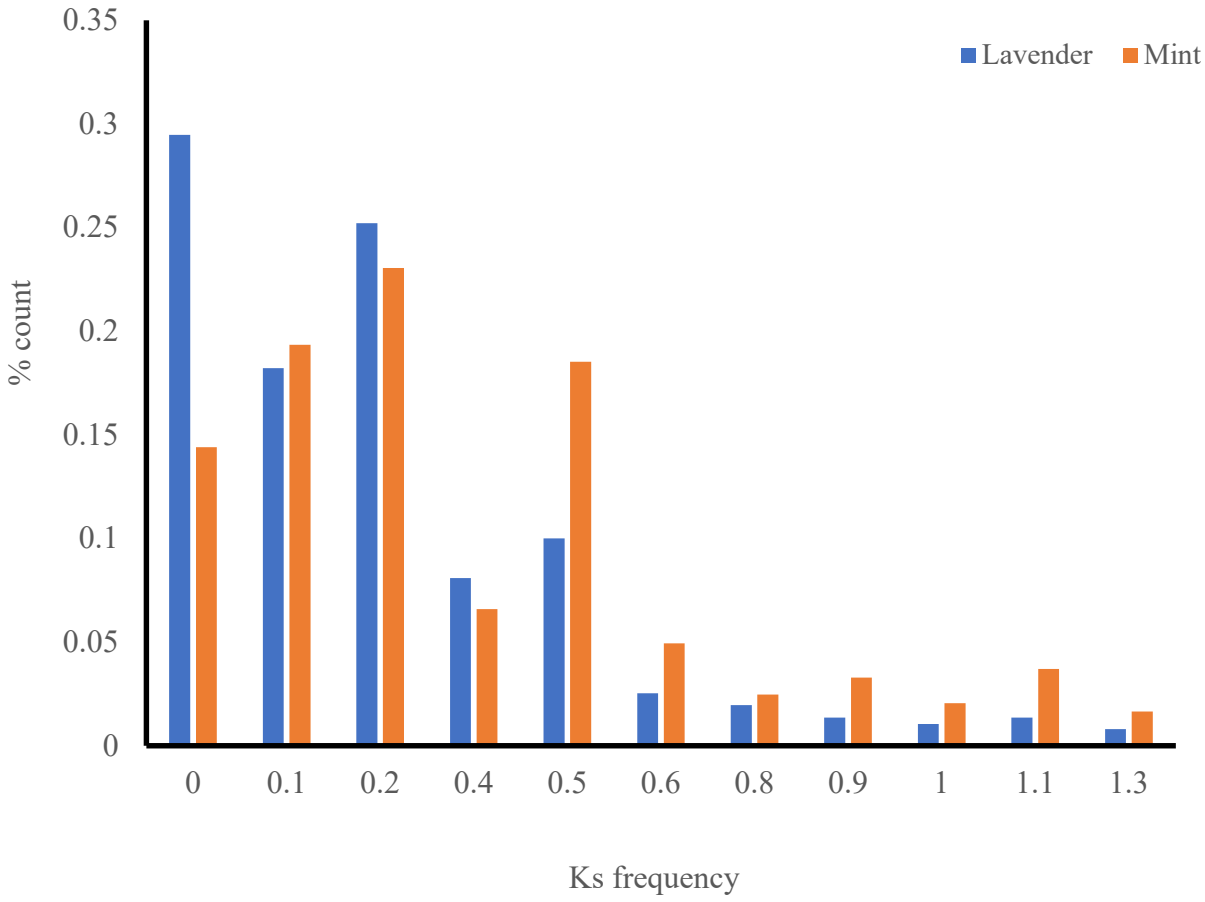


Figure 9. Comparison of polyploidization events in lavender and mint genome. Estimating recent polyploidization events using synonymous nucleotide substitutions per synonymous site (*Ks*) divergence between duplicate gene pairs based on the coding sequences.

We also compared the *Ks* values of scaffolds containing EO genes with that of the whole genome in lavender to see if there was any bias in gene duplication towards EO genes. As shown in Figure 10, while the older secondary peak for the entire genome is at *Ks*=0.5, the older secondary peak for EO scaffolds is at *Ks*= 0.4 (translates to ~32.2 Mya). The same is seen for the earlier secondary peak with that for EO scaffolds having a lower *Ks* value than the whole genome (*Ks* at 0.1 vs. 0.2). This *Ks* pattern for EO scaffolds may suggest that the EO genes had a

higher probability of retention from each of these two genome duplication events. Furthermore, at all higher *Ks* value from 0.6 up to *Ks*=1.3, EO scaffolds always show a higher percentage than the entire genome, indicating EO genes were subject to a higher probability of retention since early genome duplication occurred prior to genome duplication represented by the peak at *Ks*=0.5. Altogether, our data suggest that during the evolution of lavender, the genome has experienced at least two polyploidization events, after which gene retention favoured the EO genes. This supports and provides a mechanistic explanation for the higher level of overall gene duplication for EO genes shown in the below section (Figure 10). In addition, the fact that the initial peak for EO scaffolds is much lower than for the whole genome may suggest that EO genes have been more stabilized in the genome by showing lower levels of recent gene duplications.

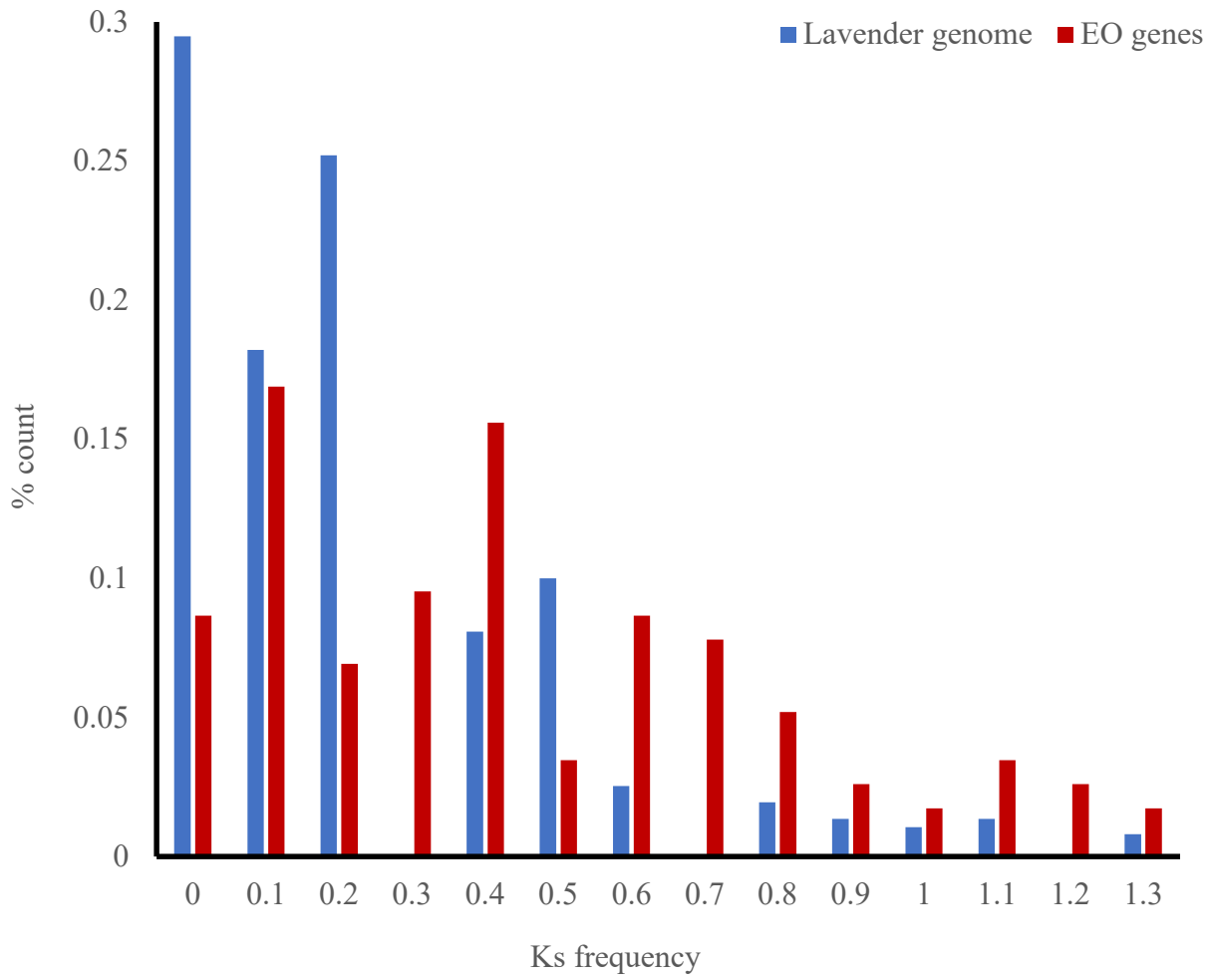


Figure 10. Age profiling of gene duplications in lavender draft genome and EO gene containing scaffolds.

Estimating recent polyploidization events using synonymous nucleotide substitutions per synonymous site (*Ks*) divergence between duplicate gene pairs based on the coding sequences. EO, essential oil.

III.5 Identification and characterization of lavender transposable elements

III.5.1 Comparative analysis reveals a unique LTR profile in lavender

The repeat composition is an important aspect of a plant genome's characteristics. We annotated the lavender genome repeat sequences using RepeatMasker (Smit et al., 2013), RepeatModeler (http://www.repeatmasker.org/RepeatModeler/), and RepeatProteinMasker (http://repeatmasker.org). As shown in Table 19, a total of ~372 Mbp repeat sequences were identified, which constitute 45.5% of the draft genome including the gap sequences, or ~57% of the non-gap sequences. Among the known repeat types, LTRs are the dominant type with over 194,000 copies contributing ~165 Mbp or 19% of the genome. DNA transposons with over 128,000 copies contributing ~74Mbp or ~3.7% of the genome represent the second most common repeats. Other minor types are Short interspersed nuclear elements (SINEs), Long interspersed nuclear elements (LINEs), simple repeats and other low complexity sequences, each type contributing to no more than 0.5% of the genome. In addition to these known types, there is a total of ~143Mbp (~19%) repeat sequences that are uncharacterized repeat sequences.

In comparison, the repeat composition of the mint genome is 32.55% (~115 Mbp) (32.6% for non-gap sequences), significantly lower than that of the lavender genome. The difference is mainly attributed to the lower content of the LTRs, being less than half of that for lavender even by genome percentage or 1/3 by copy number (Table 19). Among the repeat types, the content of DNA retrotransposons is also slightly lower (2.86% vs. 3.73%) and the percentage of the unclassified repeats is similar to that in lavender (20% vs. 18.85%).

Table 19. Comparison of repeat composition between lavender and mint genomes

Plant genome	***Lavandula angustifolia***			***Mentha longifolia***		
RE type	RE count	RE length (bp)	genome (%*)	RE counts	RE length (bp)	genome (%*)
SINES	397	50,260	0.01	45	4,977	0
LINES	5,476	2,417,175	0.28	7,108	825,619	0.23
LTR elements	194,380	164,495,826	18.91	66,997	25,659,648	7.23
DNA elements	216,471	74,076,904	3.73	31,395	10,167,267	2.86
Unclassified	527,949	143,516,873	18.85	405,753	70,998,842	20.00
Small RNA	1,772	301,641	0.03	3,946	439,334	0.12
Satellites	770	235,272	0.03	7	643	0
Simple Repeats	148,041	7,178,014	0.83	155,247	6,220,878	1.75
Low complexity	25,218	1,350,401	0.16	29,334	1,514,900	0.43
Total	1,120,474	393,622,366	45.25	699,832	115,832,108	32.55

*, based on total genome sequence including gaps; RE, repeat elements.

We performed more detailed analysis of the LTRs by examining the relative abundance between the two major LTR subtypes, *Gypsy* and *Copia,* and compared the data in lavender with that in four other genomes. Very interestingly, as shown in Figure 11A, the LTRs in lavender are mainly composed of *Copia* elements. The largest amount of *Copia* elements is in lavender among all 5 genomes analyzed both by copy number and by percentage in the genome. More specifically, the copy number of *Copia* LTRs in lavender is ~150,000, higher than that number (~80,000) for the maize genome, which is more than four times larger in size (Figure 11A). By genome percentage, the content of *Copia* in lavender (16.30%) is even higher than that in the

other four genomes (Figure 11B) for being 5, 4, 2.8, 1.1 times higher than rice (3.25%), mint (3.63%), *Arabidopsis* (5.7%) and maize (14.6%), respectively. This is in a clear contrast with the situation of *Gypsy*, which is lower than *Copia* in lavender but higher than *Copia* in all other four genomes.

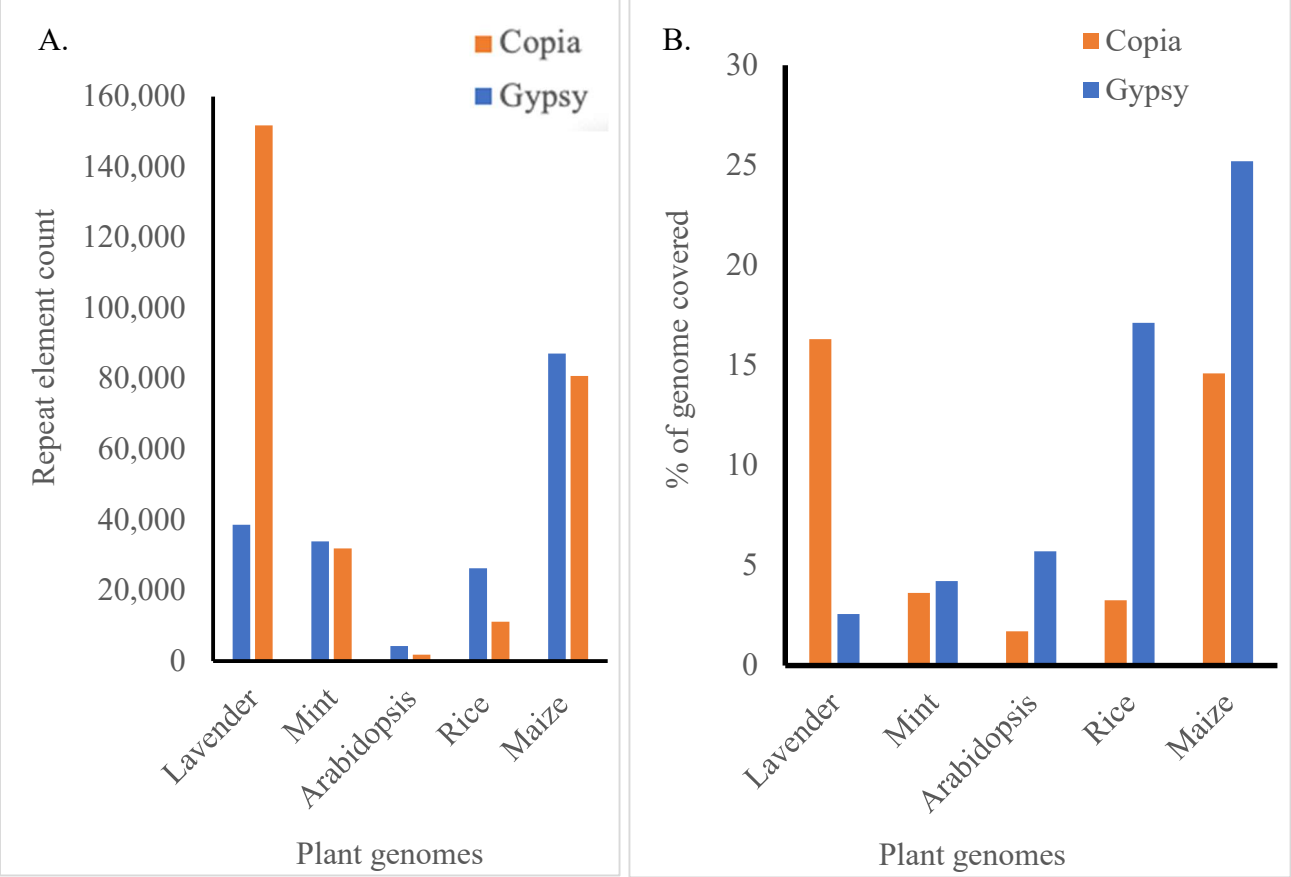


Figure 11. Comparison of LTR composition of lavender genome with four plant genomes. A. LTR counts *(Copia* and *Gypsy*) in lavender and other model plants and B. The percentage composition of *Copia* and *Gypsy* in lavender and other model plants.

Lavender also seems to have a high content of MITEs (Miniature Inverted Tandem Repeats), which are a subtype of DNA transposons. MITEs are class II non-autonomous DNA elements, lacking their own transposases, which are characterized by their small size, terminal

inverted repeats TIRs and high copy numbers (Feng, 2003). By copy number, lavender's MITE content is slightly lower than that in the maize genome (which is the highest) (Figure 12 A), but by genome percentage, lavender and rice have the third highest (4.7%), among the five genomes analyzed (Figure 12B). In *Arabidopsis* and mint genomes, the proportion of MITE is higher even though the count is lower, due to their small genome sizes (157 Mbp, and ~400 Mbp, respectively).

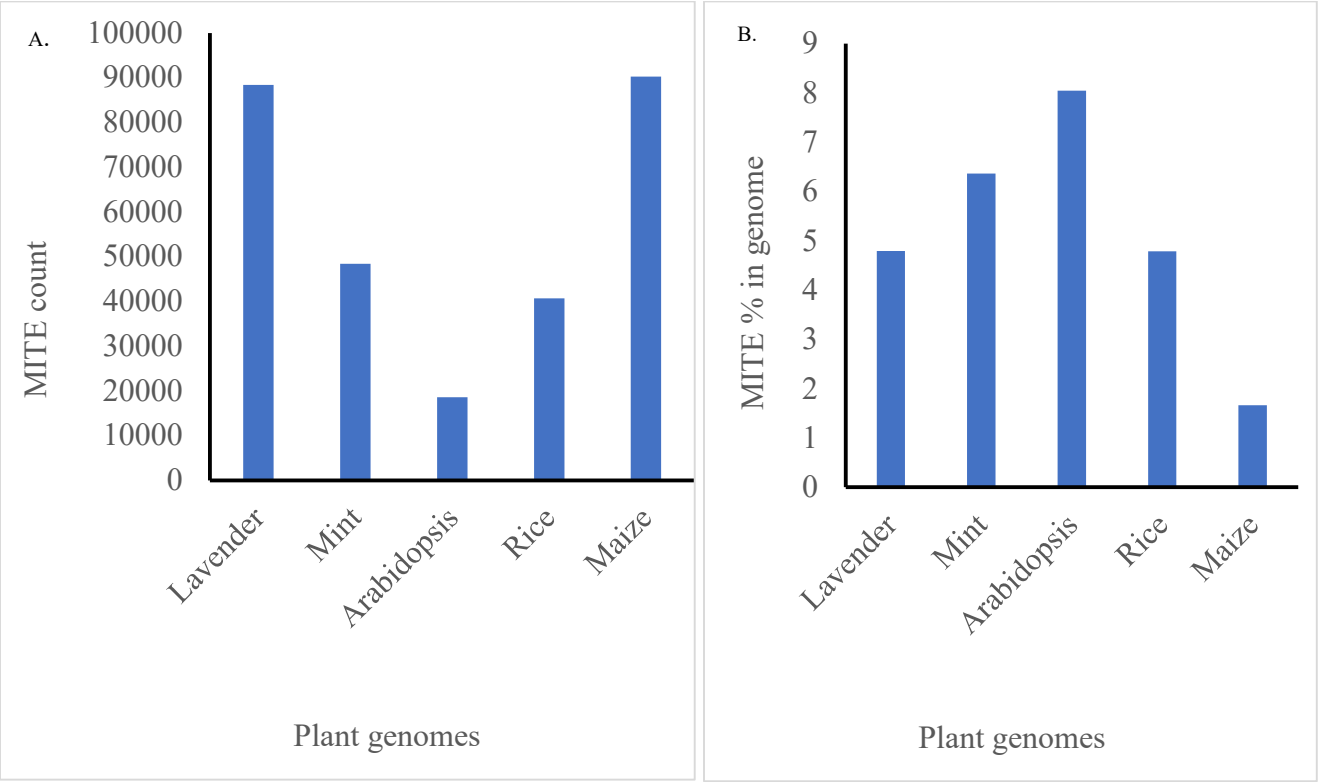


Figure 12. Comparison of MITE content in lavender and other plant genomes.

A. Comparison by copy number of MITE elements estimated by MiteFinder II (Shang et al. 2018) in lavender and model plants . B. Comparison and distribution of MITE elements by genome percentage in lavender and model plants.

We further analyzed age profiles of the *Gypsy* and *Copia* elements in lavender genome in comparison with that for the mint genome. For this, we surveyed the composition of sequence

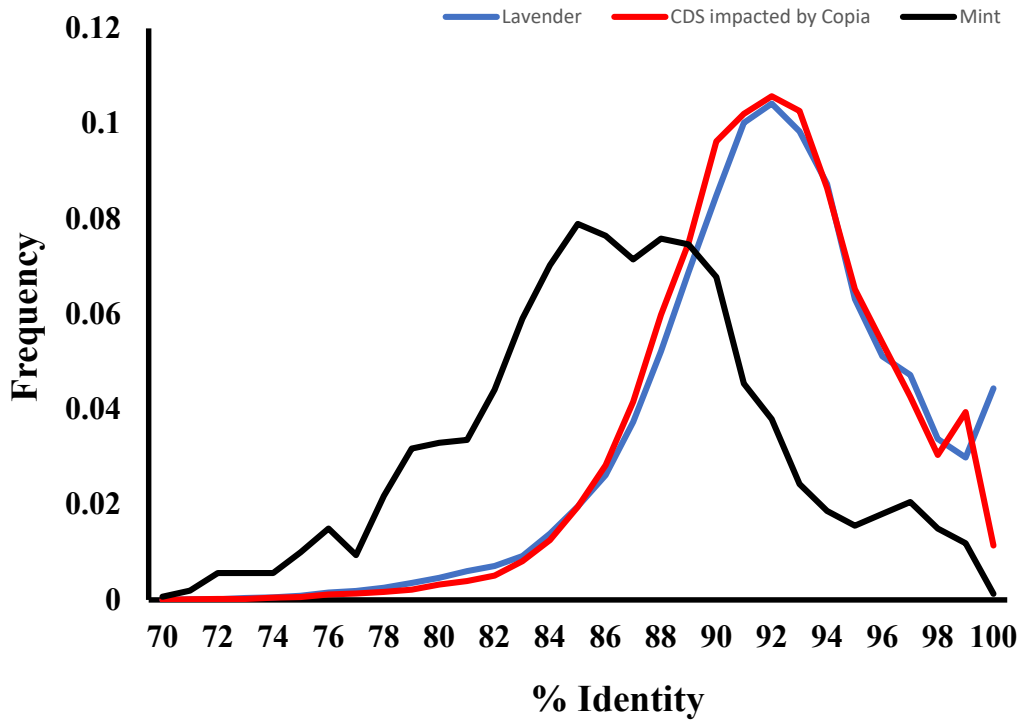
similarity based on the best non-self-match from an all-against-all sequence alignment within each group of the LTRs in each genome. The sequence similarity serves as a measure of the relative ages of the LTR elements in a genome: the lower the sequence similarity, the older its insertion time in the genome. The age profile also provides depiction of the proliferation rate of a repeat group over the course of the genome evolution.

As shown in Figure 13, for both *Gypsy* and *Copia*, there is a striking difference between lavender and mint genomes with the lavender genome having a later start of the emergence of the LTRs than in the mint genome. Specifically, as shown in Figure 13A, the activity level for *Copia* elements peaked from 84% to 89% similarity in the mint genome, while it peaked at 92% similarity level in the lavender genome. The peak activity of *Copia* elements is higher in the lavender genome than in the mint genome, explaining their higher copy numbers in the lavender genome. Similarly, as shown in Figure 13B, for *Gypsy* elements, the activity peaked from 86% to 89% similarity in the mint genome, while it peaked at 92% similarity level in the lavender genome. In this case, the peak height is similar, agreeing with the similar content of *Gypsy* elements in the two genomes.

Another interesting difference is that, in the mint genome, both *Gypsy* and *Copia* elements showed a rapid drop in their activity with the current activity being close to zero (i.e., very low percentage of elements at 100% similarity). However, in the lavender genome, both types of the LTR elements showed a recent increase of activity as indicated by the uptick of the curve from 99% to 100% similarity. We also examined the LTR sequences involved in CDS sequences in the lavender genome. Overall, the age profile is similar to their genome profile. However, some differences are also seen, particularly at the young age end. For both *Copia* and *Gypsy* elements, their contribution to CDS sequence showed a decreasing trend among younger

elements, despite their increasing activity in the genome (Figures. 13A and B), suggestion the very young LTR elements have less impact on gene function.

A



B

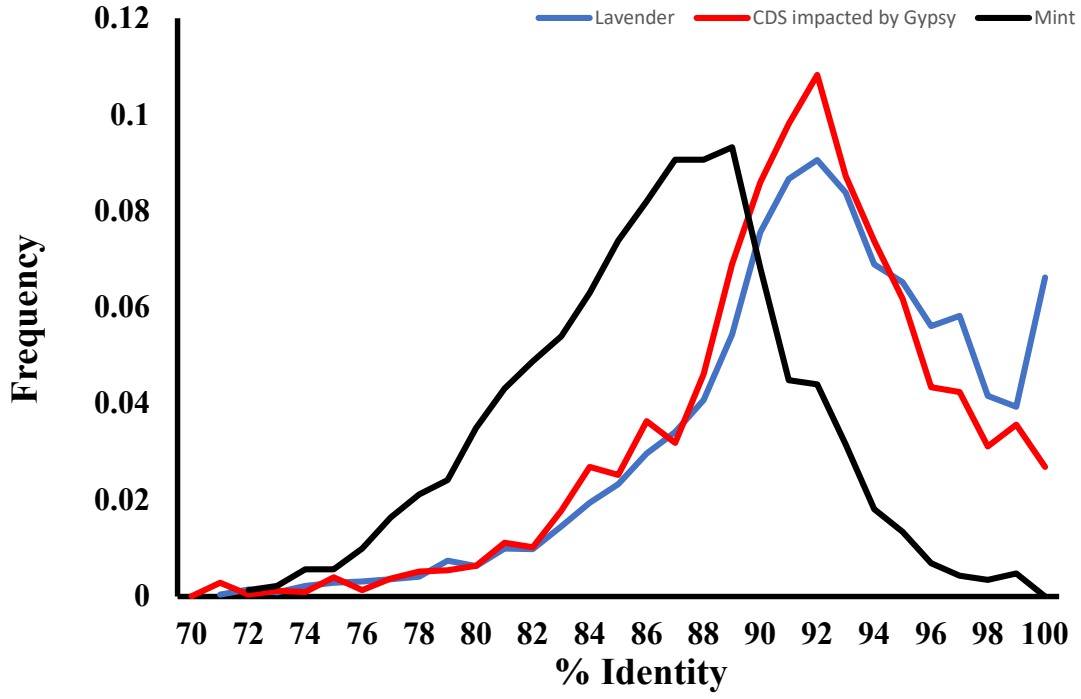


Figure 13. Comparison of age profiles of *Copia* and *Gypsy* elements in lavender and mint genome.

A. Line plots showing the sequence divergence profile of *Copia* elements in the lavender and mint genomes, with the *Copia* elements showing signs of recent activity based on increasing frequency at 100% identity. B. Line plots showing the sequence divergence profile of *Gypsy* elements in the lavender and mint genome, with *Gypsy* elements in the lavender genome showing signs of recent activity based on increasing frequency at 100% identity.

III.5.2 *Copia* and *Gypsy* elements make significant contribution to protein coding genes in lavender genome

To assess the functional impact of the LTRs in the lavender genome, we examined their distribution in the genome in gene context. Specifically, we analyzed the position overlap of the LTR elements with the positions of the annotated genes and obtained the percentage of *Copia* and *Gypsy* sequences in genic categories classified as promoters (200 bp upstream of transcription starting sites), CDS, UTRs (5' and 3' UTRs), and introns in comparison with the genome average. Very interestingly, as shown in Tables 20 and 21, for both *Copia* and *Gypsy* elements, a biased distribution towards the CDS regions and a bias against introns and UTR regions are seen, while the content of these two elements in the promoter regions is lower but close to the genome average. Specifically, *Copia* elements contribute to 30.4% of the CDS with 57,906 copies, while *Gypsy* elements contribute to 6.1% (10,159 copies), and together they constitute 36.5%, or more than 1/3 of the CDS sequences, with 68,065 copies. This is more than 50% higher than the genome average (24.5%). By copy number or total sequence length, *Copia* elements make a much larger contribution to CDS than *Gypsy* elements (Table 20), apparently due to the much larger copy number of *Copia* than that of *Gypsy* elements in the genome, since

the ratio between *Copia* and *Gypsy* in each genic category is more or less around 6:1, which is similar to that for the whole genome. This suggests that *Copia* and *Gypsy* elements have a similar distribution pattern by gene context in the lavender genome, including a strong bias towards the CDS regions.

Table 20. Contribution of LTRs to genes in the lavender genome.

Category/LTR (%*)	*Copia*	*Gypsy*	Total
Genome	21.15	3.36	24.5
CDS	30.41	6.10	36.5
UTRs	12.63	2.81	15.4
Promoter	19.10	2.20	21.3
Introns	9.80	1.83	12.0

*, percentage calculated based on non-gap sequences

Table 21. Impact of *Copia* and *Gypsy* elements in different regions of lavender protein coding genes.

		All coding genes			EO genes		
Genic region	Parameters	*Copia*	*Gypsy*	Total	*Copia*	*Gypsy*	Total
Promoter	LTR count	20,588	3,168	23,756	4	NA	4
	Prom count	25,577	4,192	29,769	4	NA	4
	Length (bp)*	7,915,267	913,490	8,828,757	1073	NA	1073
CDS	LTR count	57,906	10,159	68,065	1	2	3
	CDS count	64,648	14,077	78,725	1	2	3
	Length (bp)*	23,936,250	4,797,936	28,734,186	78	336	414
Intron	LTR count	39,437	6,943	46,380	4	4	8
	Intron count	43,831	7,802	51,633	4	4	8
	Length (bp)*	4,898,795	916,035	5,814,830	836	626	1462
5'-UTR	LTR count	2,018	553	2,571	NA	NA	NA
	UTR count	2,149	592	2,741	NA	NA	NA
	Length (bp)*	218,516	54,652	273,168	NA	NA	NA
3'-UTR	LTR count	2,474	509	2,983	NA	1	1
	UTR count	2,611	544	3,155	NA	1	1
	Length (bp)*	316,318	64,435	380,753	NA	189	189

*, LTR length in the specified genic region; EO, Essential Oil; CDS, Coding sequences; UTR, Untranslated regions.

Among genes involving *Copia* and *Gypsy* elements, eight genes have LTRs in the intronic regions, three genes have LTRs in the CDS regions, one gene has a *Gypsy* in the 3'-UTR, and four genes have LTRs in their promoters (Table 21). The three genes with LTRs in CDS include the key MVA pathway gene AACT (Acetoacetyl-coenzyme A thiolase) with two *Gypsy* elements contributing to two CDS exons (Figure 14 A), the TPS (Terpene synthase) gene with one *Copia* element contributing to one of the CDS exons (Figure 14 B), and a (Z)-gamma-bisabolene synthase gene entirely derived from a *Copia* element (Figure 14 C). As an interesting example, two small genes were derived from one *Copia* element (Figure 14 D), although in this case, the authenticity of these two genes requires validation.

GO term analysis of the genes involving *Copia* and *Gypsy* elements in CDS (Figure 15) revealed that these genes are associated with a wide variety of functions ranging from nucleic acid binding (GO:0003676) to oxidation-reduction process (GO:0055114). The GO term with the largest number of genes (2154) involving *Copia* elements in CDS is GO:0003676 (nucleic acid binding), while that for *Gypsy* elements is GO:0006468 (protein phosphorylation) (Figure 15). The nucleic acid binding function (GO:0003676) is also associated with the IPR classification IPR013103 (reverse transcriptase enzyme), which is associated with more than 9000 genes in the lavender genome (Table 13). Top GO categories for genes impacted by *Gypsy* elements include protein phosphorylation (GO:0006468 with1032 genes), protein kinase activity (GO:0004672 with 1027 genes) and ATP binding (GO:0005524 with 867 genes). Our data indicate that *Copia* and *Gypsy* elements contribute to different gene functions.

Altogether, our results reveal a unique aspect of the lavender genome's repeat profile in that it displays an unusually high content of *Copia* elements and a recent increase of LTR is predicted for both *Gypsy* and *Copia* elements. These LTR elements showed an active

participation in gene function by showing a strong distribution bias towards the CDS regions, including for genes directly involved EO pathways, suggesting their unique contribution in making lavender as a special EO-producing species.

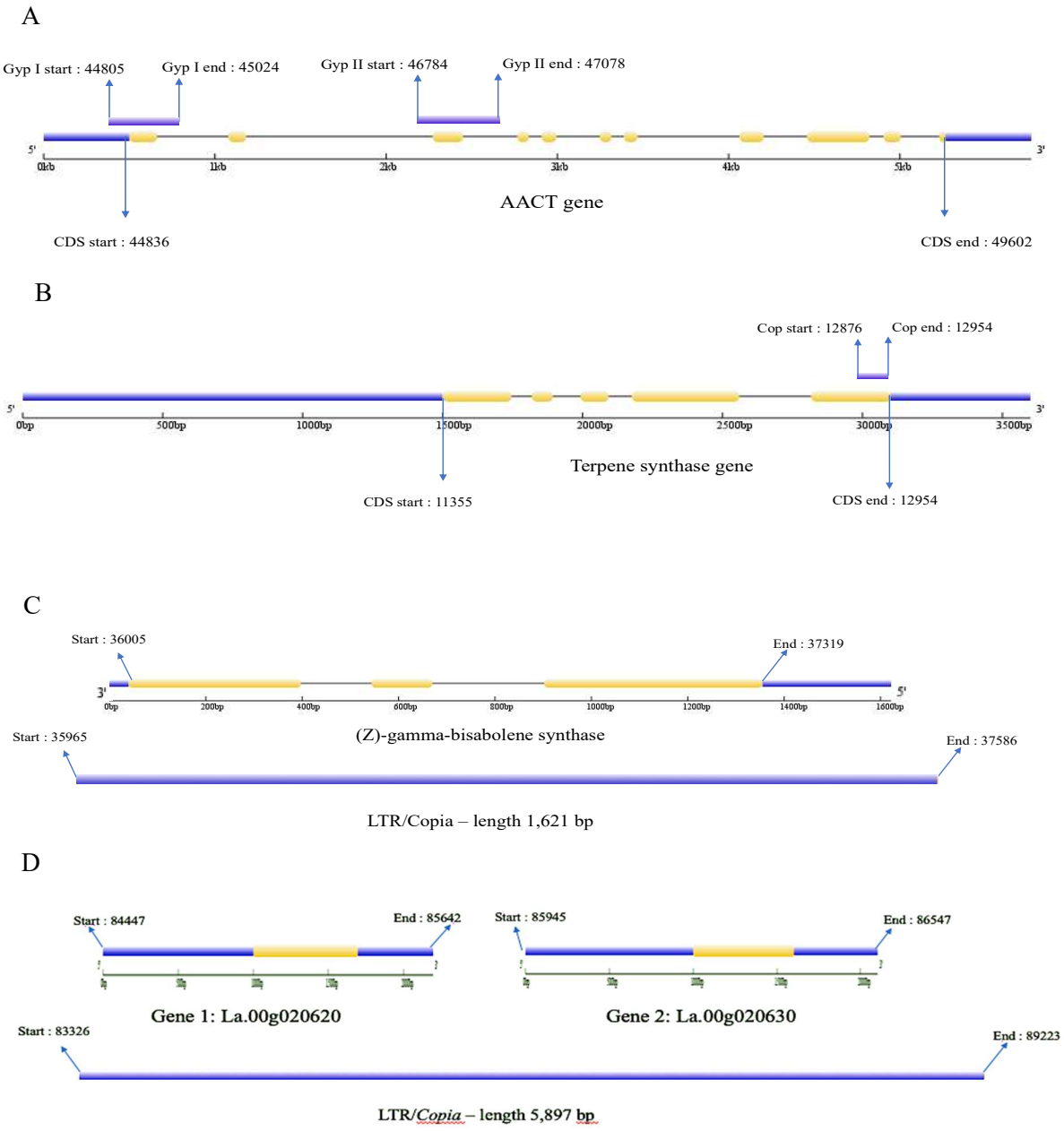


Figure 14. Examples of genes with CDS contributed by *Copia* and *Gypsy* elements.

A. Overlap of two *Gypsy* elements with two CDS exons of Acetoacetyl-coenzyme A thiolase (*A*act) gene. B. Overlap of a *Copia* element with a CDS exon of the Terpene synthase (*Tps*) gene. C. A sesquiterpene biosynthesis gene, (Z)-Gamma-bisabolene synthase, was entirely derived from a *Copia* element. D. Multiple genes derived from a single *Copia* element. The figures were

generated using the Gene Structure Display Server (GSDS) (Hu et al., 2015) with fasta nucleotide sequences (CDS and genome sequences) as input and default parameters. CDS exons are indicated in yellow color, while UTRs are indicated in blue color. LTR sequences are represented as blue bars either above or below the associated genes.

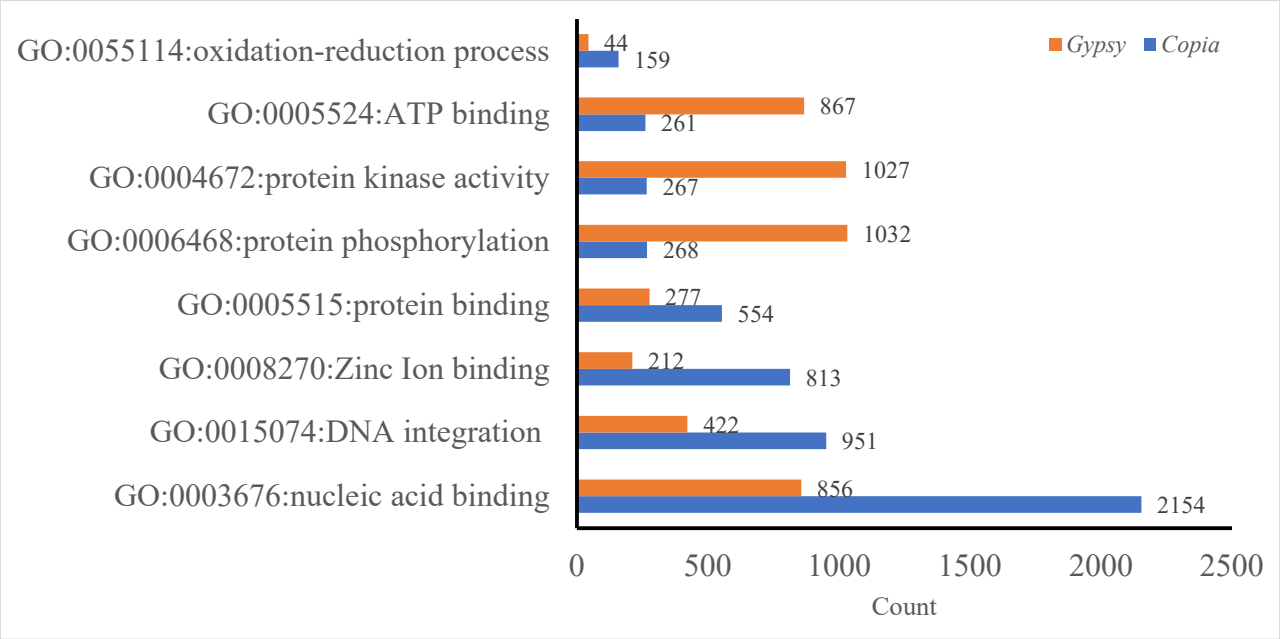


Figure 15. Top gene ontology (GO) terms associated with genes impacted by LTR elements. Functional GO analysis, distribution profile and comparison of genes impacted by *Gypsy* and *Copia* elements**. The GO annotation and plotting were done using WEGO (Shi et al. 2018).**

Chapter IV. Discussion

To better understand the unique biological aspects of lavender as an essential oil producing plant, we performed *de novo* genome sequencing of *Lavandula angustifolia* (Maillette), generated the first lavender genome assembly and determined the genome size of *Lavandula angustifolia* to be around 830 Mbp. Through genome annotation and comparative analysis with closely related and model plant genomes, we identified a number of unique features of the lavender genome, which reveals the genome's adaption for essential oil production. We discuss below several important aspects of our results in context of the existing literature.

IV.1 The genome size of Lavandula angustifolia (Maillette).

Regarding the genome size of lavender, there have been two previous reports, one suggesting a very large size at 5.7 pg for its 1C-value (Zonneveld et al. 2005) and another reporting 1C-value between 0.79 pg and 0.9 pg (Urwin et al. 2007). In both reports, flow cytometry was the method used was to estimate the genome size of lavender. The tomato cultivar 'Grosse Lisse' was used as reference (1C-value of 1.0 pg) in the study by Urwin and co-workers (Urwin et al., 2007). Another report (Urwin, 2014) estimated the 1C-values of several lavender species (*L. angustifolia, L. latifolia, L. lanata,* and *L. x intermedia*) and their varieties and induced polyploids using flow cytometry; 1C-values were reported for diploid plants ranging from 0.823 pg to 1.39 pg. The *Lavandula angustifolia* (Maillette) genome size was not estimated in any of these reports. Since the genome sequencing was done with this cultivar, we decided to estimate the genome size of *Lavandula angustifolia* (Maillette) using a qPCR method developed by Wilhelm et al (Wilhelm et al., 2003) and a computational method, KmerGenie (Chikhi and Medvedev, 2014), based on the genome sequencing data. The estimated genome size (1C-value) is 0.96 pg (Figures 3A and B) based on the qPCR method, which converts to ~870 Mbp in

genome length, while the genome size estimated from KmerGenie is ~870 Mbp, perfectly matching the qPCR results (Malli et al., 2019). It is important to note that although the *Hmgs* gene was shown to have two copies in the genome (Table 15), it turns out that the primers designed for qPCR assay for this gene (Table 4) would only be able to amplify one of the copies due to sufficient sequence divergence between the gene copies. Therefore, our qPCR assay worked as if the *Hmgs* was a single copy gene. Our estimate of the genome size 870Mb is slightly higher but close to earlier genome size estimates in the range of 0.82 to 0.90 pg for a few other cultivars of *Lavandula angustifolia*, including 'Riverina Eunice' at 0.86 pg, 'Hidcote' at 0.82 pg, and 'C7/103' at 0.90 pg (Urwin, 2014). A recent report has determined the genome size of a different lavender cultivar (Jingxun 2 from China) to be around 1094.97 Mbp using both the traditional (flow cytometry) and k-mer based genome size estimation methods (Li et al., 2021). Therefore, we believe that 870 Mbp (or a 1C value of 0.96 pg) is a reliable estimate of the genome size for *Lavandula angustifolia* (Maillette). This adds to the coverage of genome size determination for cultivars of *Lavandula angustifolia*, and it places the genome size of this particular cultivar (Maillette) at the larger end of the genome size spectrum among the diploid cultivars of *Lavandula angustifolia.* It appears that there is quite a bit variation of genome size for *Lavandula angustifolia* cultivars, likely reflecting a complicating breeding history of these cultivars.

The agreement between the genome size estimations from the qPCR method and the computational method also indicates the reliability of the qPCR method, which may be easier and quicker to perform than the flow cytometry method. The requirement for the qPCR method is the availability of the DNA sequences for a few genes known to be in single copy. The *Dxr* gene used by the qPCR method to estimate the lavender genome size, was verified as a single

copy gene by our draft genome assembly. For situations where genome sequencing data has been generated, the computational methods can also be a convenient option. However, the K-mer based computational methods can be subject to biases and inaccuracies (Ranallo-Benavidez et al., 2020). Therefore, the use of a different method, such as qPCR or flow cytometry, should always be recommended for validation purposes.

IV.2 The lavender genome assembly and its quality

Despite their importance as essential oil crops worldwide, there has been a lack of genomic sequences for lavender species and the sequences available in NCBI are few in number (Table 1). Since 2010, there has been a steady development of genomic resources for lavender starting with EST data developed by Mahmoud and co-workers (Lane et al., 2010), followed by the identification and characterization of key lavender EO genes (Demissie et al., 2011, 2012, 2013; Erland and Mahmoud, 2014; Sarker and Mahmoud, 2015; Adal and Mahmoud, 2020). To bridge the gap in knowledge for lavender genomic resources as the main objective of this thesis research, we performed *de novo* genome sequencing of *Lavandula angustifolia* (Maillette) using the Illumina NGS platform and generated a high quality of draft genome assembly. We also performed genome annotation with emphasis on key EO genes and repeat profiles.

In generating the draft genome of *L. angustifolia* (Maillette), the Illumina HiSeq 2000 platform was used, combining the use of pair-end and mate-pair libraries, all sequenced at 100 bp X 2 setting, the most efficient genome sequencing approach at the time of the project. We generated a total of ~ 150 X coverage, which is considered more than sufficient.

For *de novo* genome assembly, we tested a multitude of tools, including popular genome assembly pipelines, such as Abyss (Simpson et al., 2009), SOAPdenovo (Luo et al., 2012),

SPAdes (Bankevich et al., 2012), Allpaths-LG (MacCallum et al., 2009), and stand-alone contig generators, such as Fermi (Li, 2012), SGA (Simpson and Durbin, 2012), Velvet (Zerbino and Birney, 2008), and scaffolders such as OPERA (Gao et al., 2011), SSPACE (Boetzer et al., 2011), SOPRA (Dayarian et al., 2010). Since each genome is different in character with variations in genomic components, such repeat content, ploidy status, size of the genome, and the number of genes, there is no best universal approach for *de novo* genome assembly. We compared the quality of genome assemblies from these tests and decided on the use of FERMI (Li, 2012) for contig assembly and OPERA (Gao et al., 2011, 2016) for scaffolding. This was followed by post-assembly improvements using tools such as L-RNA scaffolder (Xue et al., 2013), which utilized RNA-Seq data to improve the draft assembly, and gapcloser (Luo et al., 2012) tool which utilized mate-pair reads to close the gaps in the scaffolds. The resulting lavender draft genome assembly consists of 84,291 scaffolds with a total sequence length of 869,786,077 bp, among which the non-gap sequence (excluding bases in "N" as undetermined sequences) length is 688,040,719 bp or ~79% of the draft genome assembly (Table 12). Despite not being as good as the reference genomes for well-studied model or important crop genomes, such as *Arabidopsis* (Kaul et al., 2000), maize (Schnable et al., 2009), rice (Yu et al., 2002), and grape (Martin et al., 2010), this is high quality genome assembly and better than that of the mint genome assembly, which was generated using more resources including PacBio long-read sequencing and transcriptome data (Vining et al., 2017).

The annotation of the draft genome was done to identify functional elements and critical genes involved in the essential oil biosynthesis and key gene elements like the positions of introns, promoters and UTRs of protein coding genes. The final tally of protein coding genes in the lavender genome was 60, 819 genes (Table 12), which is higher as compared to the

Lamiaceae relatives, such as mint (35,597 genes), sesame (Wang et al., 2014) (27,148). Another plant, *Ocimum sanctum* "Holy basil" from *Lamiales*, was found to have 53,480 genes (Rastogi et al., 2015). The difference in the genome sizes of the mint (353 Mbp) and *O. sanctum* (386 Mbp) and lavender (~870 Mbp) might be one of the reasons for the increase in the gene count in lavender. Apart from the protein coding genes, 1,512 tRNA genes and 617 rRNA genes were identified, bringing the total to 62,948 genes (Table 12). Analysis of the number of tRNA genes in various model plant genomes shows that the number of tRNA genes in lavender is highest among the model plants *Arabidopsis* (580), rice (538) and maize (771) (Chan and Lowe, 2009, 2016) included in our comparison. Once again this might be due to lavender's relatively large genome size.

The *de novo* genome annotation allowed a BUSCO (Simão et al., 2015) analysis of the lavender draft genome to assess the genome assembly quality based on the coverage of known conserved genes. The results indicated that BUSCO coverage in the lavender draft genome is similar to that of the maize genome (Figure 5) in terms of missing and fragmented SCOs (Figure 5). The BUSCO results of lavender and maize were almost similar, only differing in the complete and duplicated SCO numbers, with lavender being the highest (696) and maize only having 73 (Figure 5). The maize genome sequence was generated by a large research consortium integrating a large amount of genetic data (Bennetzen et al., 2001; Haberer et al., 2005). The generation of the mint draft genome also integrated PacBio sequencing reads in addition to the Illumina sequencing data (Vining et al., 2017). The draft genome (Maillette) is comparable to a new lavender genome (cultivar Jingxun 2) (Li et al., 2021) generated using significantly more resources including PacBio long-read sequencing, Illumina and Hi-C technology. The BUSCO scores of 91.4% and 91.8% were obtained for Jingxun 2 and Maillette, respectively. Therefore,

with our lavender genome assembly built purely based on Illumina short-read sequencing data, achieving a comparable quality to the maize genome and the new lavender genome by gene coverage is an excellent outcome.

The assessment of annotation quality was also done by aligning the annotated protein sequences against those from model plants, including *Arabidopsis*, rice and maize. Annotated lavender protein sequences showed matches to 83% proteins in *Arabidopsis* and 95% of protein sequences in rice and 94% matches to maize genomes (Table A.3), serving as another indicator for the good quality of our lavender draft genome assembly.

IV.3 The GC content of lavender genome

The GC content, one of the important compositional features of the genome, is ~38% for the lavender genome. Studies have shown that there is correlation between the genome size and GC content (Šmarda et al., 2014) and that the GC content in the genome is also linked to various environmental factors such as temperature fluctuations (Veleba et al., 2017). The GC composition is varied in plants ranging from 36% in *Arabidopsis* to more than 40% in rice (43.6%) and maize (47.2%) (Singh et al., 2016). The 38% GC content of lavender genome (Maillette) places it in the lower GC range, even though the genome size is larger than those of *Arabidopsis* and rice. Despite the significant difference in genome size, there is no difference in the GC composition between the two cultivars of lavender, as Jingxun 2 was shown to have a GC content of 38.5% (Li et al., 2021). One of the reasons for higher GC content in a genome may be due to higher content of LTR retrotransposons. The best example is that of the maize genome, which has a GC-rich Huck element (GC content of 62%) contributing to almost 10% of the genomes (Šmarda et al., 2014). The other reason for the increase in GC content in monocot

genome is the presence of holocentric chromosomes (chromosomes lacking centromere and kinetochore spreads over the full length of the chromosome). The organization of these chromosomes results in lower recombination rates and reduced frequency of repair in heterologous sites, resulting in the preferential introduction of GC bases in the genome (Melters et al., 2012). There are other factors such methylation which may impact the GC content, and studies in parasitoid wasps have shown that methylation may contribute to low GC content in insect genomes, leading to the increase in AT-rich sequence in the genome (Dennis et al., 2020).

IV.4 The highly duplicated nature of the lavender genome

It is estimated that up to 80% of all living plants are polyploids, and many plant lineages including monocots (i.e., Oryza) and eudicots (Arabidopsis) have at least one paleo-polyploidy event in their evolutionary history (Meyers and Levin, 2006). With the availability of large scale genomics data, it is now possible to analyze polyploidization events in various plant genomes (Soltis et al., 2009; Jiao and Paterson, 2014). *Glycine max* (soybean) was the first polyploid genome to be published (Schmutz et al., 2010). There are 47 polyploid plant genomes which has been sequenced and published (Kyriakidou et al., 2018) and most important among them are wheat (Choulet et al., 2010; Marcussen et al., 2014), cotton (Li et al., 2015), potato (Xu et al., 2011) and other agriculturally important and critical crops and plants.

Using synonymous substitutions to estimate divergence times between duplicate genes has helped researchers to identify peaks, corresponding to bursts of duplication, inferred to be the result of ancestral WGD or polyploidization (Blanc and Wolfe, 2004; Roth and Liberles, 2006). The distribution of synonymous substitutions per site (*Ks*) among gene pairs within a genome when plotted and visualized is referred to as "*Ks* plot" (Cui et al., 2006). WGDs or polyploidization events generate additional normally distributed peaks in the *Ks* plots. By

counting the number of synonymous substitutions at these peaks the age of the ancient WGDs can be estimated (Tiley et al., 2018). A *Ks* distribution profile where genome duplication events occur at random show an approximate L-shaped distribution. When plotting a specific genome duplication event, there will be a secondary peak near the initial peak (corresponding to most recently duplicated genes) which then drops off after deletion of duplicated genes that are not under selective constraints (Figures 10 and 11). Larger *Ks* values (e.g., >0.75) are associated with increasingly large error. To minimize the associated error while retaining a reasonably sized data set, researchers have used only *Ks* values <2.0 for these kind of analysis (Eckardt, 2004). One of the interesting and earliest observations after quality check of the lavender draft genome using BUSCO (Simão et al., 2015) was the presence of 696 complete duplicated copies of SCOs (Figure 4). Sequence analysis using BLASTn (Altschul et al., 1990) combined with the BUSCO score analysis revealed that at least 547 (78%) of the 696 complete and duplicated SCOs represented real duplication events in the genome, distinguishing them from possible assembly errors due to genetic heterogeneity between homologous chromosomes. Indeed, possible duplication events resulting in polyploidy have been reported for *L. angustifolia* (Upson and Andrews, 2005). In this context, our data indicate that the cultivar used in this study (Maillette) is most likely a polyploid line.

The estimated timing of polyploidization events in lavender genome have also been calculated and found to have occurred between 16 to 41 Mya and similar to that of the mint genome. A previous study had estimated the polyploidization event in the genome of tomato (Sato et al., 2012) and potato (Xu et al., 2011) to be around 71 Mya , which shows that the lavender and mint genome experienced recent polyploidization events. It should be noted that the *Ks* plots can be difficult to interpret, and the accuracy of WGD identification from Ks plots is not

certain and subject to interpretation and discussion (Tiley et al., 2018). The analysis of the recently assembled lavender cultivar Jingxun 2 indicated two WGD events at 29.6 and 6.86 Mya (Li et al., 2021) which is similar to the estimated WGD events of the Maillette cultivar (Malli et al., 2019).

In summary, ploidy estimation of the lavender genome shows the diploid genome has a history of polyploidization, biased towards the regions with essential oil genes (Figure 8), likely responsible for the larger gene copy numbers observed for certain essential oil genes, as further discussed in the next section. These advantageous changes in the genome may have provided the fitness to the lavender plant, in terms of increased gene expression and higher enzymatic activity (Chen, 2010), to survive in a competitive environment (Renny-Byfield and Wendel, 2014) and for its unique EO production capacity.

IV.5 The essential oil genes in the lavender genome

The generation of the lavender draft genome has made it possible to identify key EOs biosynthesis and TPS genes. The sequences for the genes involved in the two key pathways, cytosolic MVA and plastidic MEP pathways have been identified along with their copy numbers and reported (Malli et al., 2019). The lavender draft genome was shown to contain all of the genes known to be involved in the MVA and MEP pathways (Table 14), as well as genes encoding prenyltransferases involved in isoprenoid metabolism, which are pathways and processes related to EO production. Further, the assembly contains sequences for several putative (uncharacterized) TPS genes that can potentially control isoprenoid metabolism in lavenders (Table 17).

All the genes encoding enzymes involved in the MEP pathway, except for the *D*xr and *M*ct genes, were shown to be present in multiple copies ranging from 2 to 13 (Table 15). Among these,

the 1-deoxy-D-xylulose 5-phosphate synthase (*Dxs*) gene (13 copies) and 4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase (*Hdr*) gene (7 copies) play a key role in the biosynthesis of the EO intermediaries. The significance of the *Dxs* gene is that it involves the first step of the MEP pathway, while the *Hdr* is involved in the last step. Various studies suggest that *Dxs* is a rate-limiting enzyme in the biosynthesis of terpenoids (Simpson et al., 2016; Zhang et al., 2018). The expression of *Dxs* leads to increase in the terpenoids and carotenoids in *Arabidopsis* and tomato plants, while overexpression or suppression of *Dxs* alters the levels of specific isoprenoids in *Arabidopsis* tomato and potato (Muñoz-Bertomeu et al., 2006; Zhang et al., 2018). It is likely that having a larger copy number of the first and last genes of the MEP pathway permits the plant to be more efficient in the production of terpene compounds (essential oils, resins, etc.) as evidenced in other plants, such as pine (Kim et al., 2009). Recent studies have indicated that the metabolite exchange depends on a more complex regulation of the MEP and MVA pathways by light and by metabolic and developmental factors, which is in contrast to the previously held notion of a simple unidirectional transport of terpene intermediate products from plastids to the cytosol (Tholl and Lee, 2011).

In addition to the MEP and MVA pathway genes, a total of 56 lavender TPS genes (Table 16) were identified in the lavender genome using HMM-based gene identification method and tool called Terzyme (Priya et al., 2018). Once the preliminary identification of the TPS genes were made, they were further classified into monoterpene, diterpene and sesquiterpene based on function and product classification. The classification of the TPS genes into sub-families based on sequence properties and functional characteristics led to the identification of ~2000 TPS-a genes in lavender (Table 17), more than in genomes that are much larger in size, such as maize. Further analysis is required to functionally characterize and validate these TPS candidates.

In addition to the lavender genome, we investigated the presence of orthologous TPS genes in model plants such as *Arabidopsis*, rice and maize (Figure 7) using a web-based tool

called Orthovenn (Wang et al., 2015; Xu et al., 2019). The tool was able to identify various clusters corresponding to characterized and uncharacterized TPS genes in lavender and model plants. Multiple copies of key genes were found, such as linalool synthase, 1,8-cineole synthase, borneol synthase in lavender and other model plants. Notably, the model plants included in our analysis, obviously lacking specialized genes required for EO biosynthesis, have no or low copy numbers of these genes (Table 18). The maize genome was missing most of the genes except for the linalool synthase genes (Table 18), while the rice genome has only the linalool and Beta-caryophyllene synthase genes. Further analysis is required to characterize the candidate orthologous genes which are shared between lavender and model plants and validate the presence and absence of key genes in *Arabidopsis*, rice, and maize genome.

Altogether, it indicates that this draft genome assembly is of very high quality with respect to the number and type of gene sequences it contains, as well as the proportion of annotated sequences, providing a good resource for systematic analysis of EO genes and other isoprenoid related genes, including novel ones in lavenders. Furthermore, comparative analysis results demonstrate the lavender genome as a genome specialized for EO production by having high copy number of key genes in the MVA and MEP pathways and TPS genes.

IV.6 The unique aspects of transposable element profile in the lavender genome

One of the important characteristics of a genome is its repeat composition profile, especially the profile of the transposable elements regarding their total percentages in the genome, the relative ratio of different TE types, their activity profiles, and genome distribution in context of genes. Therefore, we performed repeat annotation for the lavender draft genome assembly using RepeatModeler (Smit and Hubley, 2018; Flynn et al., 2020) and RepeatMasker

(Smit et al., 2013). The results revealed a total of ~393 Mbp of repeat sequences contributing ~45% of the draft genome or 57% of the non-gap genome. This is significantly higher than 35.9% repeat content for the mint genome, primarily due to the higher content of LTR elements (18.9% in lavender vs. 7.2% in mint) (Table 19).

More detailed analysis of the LTR elements in comparison with four other genomes, mint, *Arabidopsis*, rice, and maize, revealed an unusual pattern of LTRs of the lavender genome. While the copy number of *Gypsy* elements is larger than that of the *Copia* elements in all other four genomes, the lavender genome has more than three times the number of *Copia* elements than *Gypsy* elements (Figures 11 A and B). Such a pattern is only found in a few other genomes, including grape (Martin et al., 2010) , flaxseed (González and Deyholos, 2012), and a few others (González and Deyholos, 2012).

The activity of TEs is known to be variable among the TE types and is also dependent on the control of their activity by the host genome in genome. The amplification of TEs is not constant during evolution and there may be evolutionary periods where TEs are relatively dormant or highly active (Grover and Wendel, 2010). In this regard, the LTR elements are known to play a major role in rapid genome expansion or shrinking in response to environmental conditions. The shrinking of the genome may also be due to the LTR sequence structure in having the long terminal repeats, which share high sequence similarity that allows recombination-based quick removal of the larger internal sequences. The rate of such LTR removal processes is known to be triggered by external factors, such as environmental stresses (Bennetzen et al., 2005; Estep et al., 2013; Gaubert et al., 2017), thus significant differences in the LTR removal rate can result from the difference in the plants' environmental conditions. For example, exposure to extreme environments, such as high radiation and/or DNA breaking agents

or chemicals, would result in deletion of LTR elements by the action of activated error-prone repair mechanisms (Bennetzen et al., 2005). For this reason, we performed the activity profiling of *Gypsy* and *Copia* LTR elements based on their age composition in the genome. Our results indicated that, the peak times of both the *Gypsy* and *Copia* elements are significantly younger than those of the mint genome (Figure 13), suggesting a longer sustained transposition activity, likely explaining their larger copy numbers in lavender than in the mint genome. Furthermore, a new peak of most recent activity is also seen for both types of LTRs in the lavender genome, with *Gypsy* being higher than *Copia*, while little or no such activity is detectable in the mint genome, suggesting that an uptrend of recent and ongoing transposition activity is present in the lavender genome (Figure 13). It is very likely that the higher activity of *Copia* elements than that of the *Gypsy* elements in the early peak and recent peak is responsible for the higher copy number of *Copia* elements than the *Gypsy* elements. Other than the higher rates of new insertions, the differential rates of removal could also be a contributing factor for copy number differences for these LTR elements as discussed earlier.

Other than impacting the genome size, TEs are known to participate in gene function via a number of mechanisms, including, but are not limited to, disrupting gene function by inserting into the coding sequence of protein coding region, alternation of RNA splicing, creation of new genes, and alteration of gene regulation (Grover and Wendel, 2010; Negi et al., 2016; Ali et al., 2021). There is a direct correlation between the age profiles of *Copia* and *Gypsy* elements in the genome and in the CDS regions (Figure 13), indicating that the level of involvement in gene function is directly linked to the level of the transposition activity of these TE elements in the genome during evolution. It is interesting to notice that *Copia* elements impact ~30% of CDS

regions, which is five times higher than that of the *Gypsy* elements (Table 20). This indicates that even among similar types of TEs, their impact on gene functions can be very different.

Another interesting feature of the lavender genome is the presence of a large number of MITE elements (Figure 12 A) with ~88,000 copies contributing to ~41 Mbp of genome sequences, which is double that of the mint genome (~48,000 MITE elements, ~22 Mbp) and comparable to the maize genome (~90,000 MITE elements), which is several times large in genome size. MITEs are known to participate in gene function by inserting themselves into genic regions in plants. One of the examples of MITE interference in genic region is the discovery of 33 MITE sequences in a 225KB region of the maize genome flanking a key gene *Adh1*(Feng, 2003). MITE elements are well-studied in the rice genome, and it was found that a MITE element when inserted into the promoter region of the *Ubiquitin2* gene resulted in a 20% increase in gene expression (Chen et al., 2012). Another study showed that there is no preferential insertion of MITE elements near genes in the rice genome, which is in contrast to the widely held belief that MITE elements insertional preference to genic regions (Naito et al., 2006).

In summary, our results indicate that TEs, especially the *Gypsy* and *Copia* LTR elements make a significant contribution to the uniqueness of the lavender genome, and very likely the unique phenotype of the species, including the EO production.

IV.7 Conclusions and future work

In conclusion, using a *de novo* genome sequencing approach, we have successfully generated and annotated, at a reasonable quality as measured by the genome and gene coverage and quality of assembly, the first lavender genome. We determined the genome of *Lavandula angustifolia* (Maillette) to have a 1C value of ~0.96 pg or 870 Mbp in sequence size of the

haploid genome. The detailed comparative genomic analysis revealed the lavender genome to be adapted for EO production by having a large copy number of many key EO pathway genes and genes unique to EO-producing plants. Our analysis also revealed that while being currently diploid, the genome of lavender had a history of polyploidization events favouring of EO production. Furthermore, the lavender genome exhibits a unique profile of transposable elements, revealing an active role of LTR elements in the evolution of the species and its adaption for highly efficient EO production.

Priority for future work may involve: 1) sequencing of more lavender cultivars and species for better understanding of the genetic diversity and evolution and for development of genetic markers and tests for true-to-typing of lavender cultivars; 2) improvement of the reference genome by incorporating sequence data from a long read platforms, such as PacBio's SMRT or Oxford Nanopore, and optical mapping technologies to map the genome sequences to their chromosomal locations; 3) gene expression profiling with a focus on key EO biosynthetic genes for diverse cultivars and species under various conditions; 4) the functional characterization of putative TPS genes. The lavender draft genome, and new data from future research in above areas will serve as significant genomic resources for the lavender research communities and related industries.

V. References

Abraham, O.S.J., T.S. Miguel, H.C. Inocencio, and C.C. Blondy. 2017. A quick and effective in-house method of DNA purification from agarose gel, suitable for sequencing. 3 Biotech 7:180. doi:10.1007/s13205-017-0851-1

Adal, A.M., and S.S. Mahmoud. 2020. Short-chain isoprenyl diphosphate synthases of lavender (Lavandula). Plant Mol. Biol. 102:517–535. doi:10.1007/s11103-020-00962-8

Adal, A.M., L.S. Sarker, A.D. Lemke, and S.S. Mahmoud. 2017. Isolation and functional characterization of a methyl jasmonate-responsive 3-carene synthase from Lavandula x intermedia. Plant Mol. Biol. 93:641–657. doi:10.1007/s11103-017-0588-6

Adal, A.M., L.S. Sarker, R.P.N. Malli, P. Liang, and S.S. Mahmoud. 2019. RNA-Seq in the discovery of a sparsely expressed scent-determining monoterpene synthase in lavender (Lavandula). Planta 249:271–290. doi:10.1007/s00425-018-2935-5

Agostino, M. 2012. Introduction to the BLAST Suite and BLASTN. Pract. Bioinforma. 26

Akhtar, T.A., Y. Matsuba, I. Schauvinhold, G. Yu, H.A. Lees, S.E. Klein, and E. Pichersky. 2013. The tomato cis-prenyltransferase gene family. Plant J. 73:640–652. doi:10.1111/tpj.12063

Ali, A., K. Han, and P. Liang. 2021. Role of transposable elements in gene regulation in the human genome. Life 11:1–23. doi:10.3390/life11020118

Allen, J.E., and S.L. Salzberg. 2005. JIGSAW: Integration of multiple sources of evidence for gene prediction. Bioinformatics 21:3596–3603. doi:10.1093/bioinformatics/bti609

Altschul, S.F. 2005. BLAST Algorithm

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410. doi:10.1016/S0022-2836(05)80360-2

Andrews, S. 2016. FastQC: A Quality Control Tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Aubourg, S., A. Lecharny, and J. Bohlmann. 2002. Genomic analysis of the terpenoid synthase (AtTPS) gene family of Arabidopsis thaliana. Mol. Genet. Genomics 267:730–745. doi:10.1007/s00438-002-0709-y

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

Bankevich, A., S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, and P.A. Pevzner. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19:455–477. doi:10.1089/cmb.2012.0021

Bao, W., K.K. Kojima, and O. Kohany. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6:11. doi:10.1186/s13100-015-0041-9

Bao, Z., and S.R. Eddy. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 12:1269–1276. doi:10.1101/gr.88502

Barthelson, R., A.J. McFarlin, S.D. Rounsley, and S. Young. 2011. Plantagora: Modeling whole genome sequencing and assembly of plant genomes. PLoS One 6:1–8. doi:10.1371/journal.pone.0028436

Bateman, A., M.J. Martin, C. O'Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M.

Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-AJee, A. Cowley, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, L.G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. ArgoudPuy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A.L. Veuthey, C.H. Wu, C.N. Arighi, L. Arminski, C. Chen, Y. Chen, J.S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D.A. Natale, K. Ross, C.R. Vinayaka, Q. Wang, Y. Wang, L.S. Yeh, and J. Zhang. 2017. UniProt: The universal protein knowledgebase. Nucleic Acids Res. 45:D158–D169. doi:10.1093/nar/gkw1099

Benjamini, Y., and T.P. Speed. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res.. doi:10.1093/nar/gks001

Bennetzen, J.L., V.L. Chandler, and P. Schnable. 2001. National science foundation-sponsored workshop report. Maize genome sequencing project. Plant Physiol. 127:1572–1578. doi:10.1104/pp.010817

Bennetzen, J.L., J. Ma, and K.M. Devos. 2005. Mechanisms of recent genome size variation in flowering plants. Ann. Bot. 95:127–132. doi:10.1093/aob/mci008

Benson, D.A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and E.W. Sayers. 2017. GenBank. Nucleic Acids Res. 45:D37–D42. doi:10.1093/nar/gkw1070

Benson, D.A., I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler. 2007. GenBank. Nucleic Acids Res. 35:D21–D25. doi:10.1093/nar/gkl986

Benson, G. 1999. Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res. 27:573–580. doi:10.1093/nar/27.2.573

Bentley, D.R., S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P. Hall, D.J. Evers, C.L. Barnes, H.R. Bignell, J.M. Boutell, J. Bryant, R.J. Carter, R. Keira Cheetham, A.J. Cox, D.J. Ellis, M.R. Flatbush, N.A. Gormley, S.J. Humphray, L.J. Irving, M.S. Karbelashvili, S.M. Kirk, H. Li, X. Liu, K.S. Maisinger, L.J. Murray, B. Obradovic, T. Ost, M.L. Parkinson, M.R. Pratt, I.M.J. Rasolonjatovo, M.T. Reed, R. Rigatti, C. Rodighiero, M.T. Ross, A. Sabot, S. V. Sankar, A. Scally, G.P. Schroth, M.E. Smith, V.P. Smith, A. Spiridou, P.E. Torrance, S.S. Tzonev, E.H. Vermaas, K. Walter, X. Wu, L. Zhang, M.D. Alam, C. Anastasi, I.C. Aniebo, D.M.D. Bailey, I.R. Bancarz, S. Banerjee, S.G. Barbour, P.A. Baybayan, V.A. Benoit, K.F. Benson, C. Bevis, P.J. Black, A. Boodhun, J.S. Brennan, J.A. Bridgham, R.C. Brown, A.A. Brown, D.H. Buermann, A.A. Bundu, J.C. Burrows, N.P. Carter, N. Castillo, M.C.E. Catenazzi, S. Chang, R. Neil Cooley, N.R. Crake, O.O. Dada, K.D. Diakoumakos, B. Dominguez-Fernandez, D.J. Earnshaw, U.C. Egbujor, D.W. Elmore, S.S. Etchin, M.R. Ewan, M. Fedurco, L.J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K.J. Gietzen, C.P. Goddard, G.S. Golda, P.A. Granieri, D.E. Green, D.L. Gustafson, N.F. Hansen, K. Harnish, C.D. Haudenschild, N.I. Heyer, M.M. Hims, J.T.

Ho, A.M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M.Q. Johnson, T. James, T.A. Huw Jones, G.D. Kang, T.H. Kerelska, A.D. Kersey, I. Khrebtukova, A.P. Kindwall, Z. Kingsbury, P.I. Kokko-Gonzales, A. Kumar, M.A. Laurent, C.T. Lawley, S.E. Lee, X. Lee, A.K. Liao, J.A. Loch, M. Lok, S. Luo, R.M. Mammen, J.W. Martin, P.G. McCauley, P. McNitt, P. Mehta, K.W. Moon, J.W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S.M. Novo, M.J. O'Neill, M.A. Osborne, A. Osnowski, O. Ostadan, L.L. Paraschos, L. Pickering, A.C. Pike, A.C. Pike, D. Chris Pinkard, D.P. Pliskin, J. Podhasky, V.J. Quijano, C. Raczy, V.H. Rae, S.R. Rawlings, A. Chiva Rodriguez, P.M. Roe, J. Rogers, M.C. Rogert Bacigalupo, N. Romanov, A. Romieu, R.K. Roth, N.J. Rourke, S.T. Ruediger, E. Rusman, R.M. Sanches-Kuiper, M.R. Schenker, J.M. Seoane, R.J. Shaw, M.K. Shiver, S.W. Short, N.L. Sizto, J.P. Sluis, M.A. Smith, J. Ernest Sohna Sohna, E.J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C.L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S.M. Virk, S. Wakelin, G.C. Walcott, J. Wang, G.J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J.C. Mullikin, M.E. Hurles, N.J. McCooke, J.S. West, F.L. Oaks, P.L. Lundberg, D. Klenerman, R. Durbin, and A.J. Smith. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59. doi:10.1038/nature07517

Berthelier, J., N. Casse, N. Daccord, V. Jamilloux, B. Saint-Jean, and G. Carrier. 2018. A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga Tisochrysis lutea. BMC Genomics 19:1–14. doi:10.1186/s12864-018-4763-1

Besemer, J., and M. Borodovsky. 2005. GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. 33:W451–W454. doi:10.1093/nar/gki487

Białon, M., T. Krzysko-Łupicka, E. Nowakowska-Bogdan, and P.P. Wieczorek. 2019. Chemical composition of two different lavender essential oils and their effect on facial skin microbiota. Molecules 24:3270. doi:10.3390/molecules24183270

Blanc, G., and K.H. Wolfe. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667–1678. doi:10.1105/tpc.021345

Boetzer, M., C. V. Henkel, H.J. Jansen, D. Butler, and W. Pirovano. 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578–579. doi:10.1093/bioinformatics/btq683

Boetzer, M., and W. Pirovano. 2014. SSPACE-LongRead: Scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics 15:211. doi:10.1186/1471-2105-15-211

Bradnam, K.R., J.N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J.A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W.C. Chou, J. Corbeil, C. Del Fabbro, R.R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N.A. Fonseca, G. Ganapathy, R.A. Gibbs, S. Gnerre, É. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J.B. Hiatt, I.Y. Ho, J. Howard, M. Hunt, S.D. Jackman, D.B. Jaffe, E.D. Jarvis, H. Jiang, S. Kazakov, P.J. Kersey, J.O. Kitzman, J.R. Knight, S. Koren, T.W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. MacCallum, M.D. MacManes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T.D. Otto, B. Paten, O.S. Paulo, A.M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F.J. Ribeiro, S. Richards, D.S. Rokhsar, J.G. Ruby, S. Scalabrin, M.C. Schatz, D.C. Schwartz, A. Sergushichev, T. Sharpe, T.I. Shaw, J. Shendure, Y. Shi, J.T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B.M. Vieira, J. Wang, K.C. Worley, S. Yin, S.M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I.F. Korf. 2013.

Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. Gigascience 2:10. doi:10.1186/2047-217X-2-10

Britten, R.J., D.E. Graham, and B.R. Neufeld. 1974. [29] Analysis of Repeating DNA Sequences by Reassociation. Methods Enzymol. 29:363–418. doi:10.1016/0076-6879(74)29033-5

Campbell, M.S., M.Y. Law, C. Holt, J.C. Stein, G.D. Moghe, D.E. Hufnagel, J. Lei, R. Achawanantakun, D. Jiao, C.J. Lawrence, D. Ware, S.H. Shiu, K.L. Childs, Y. Sun, N. Jiang, and M. Yandell. 2014. MAKER-P: A Tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 164:513–524. doi:10.1104/pp.113.230144

Cantarel, B.L., I. Korf, S.M.C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A.S. Alvarado, and M. Yandell. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18:188–196. doi:10.1101/gr.6743907

Carretero-Paulet, L., I. Ahumada, N. Cunillera, M. Rodríguez-Concepción, A. Ferrer, A. Boronat, and N. Campos. 2002. Expression and molecular analysis of the Arabidopsis DXR gene encoding 1-deoxy-D-xylulose 5-phosphate reductoisomerase, the first committed enzyme of the 2-C-methyl-D-erythritol 4-phosphate pathway. Plant Physiol. 129:1581–1591. doi:10.1104/pp.003798

Castro, C.J., and T.F.F. Ng. 2017. U50: A New Metric for Measuring Assembly Output Based on Non-Overlapping, Target-Specific Contigs. J. Comput. Biol. 24:1071–1080. doi:10.1089/cmb.2017.0013

Chaisson, M.J., and P.A. Pevzner. 2008. Short read fragment assembly of bacterial genomes. Genome Res. 18:324–330. doi:10.1101/gr.7088808

Chan, P.P., and T.M. Lowe. 2009. GtRNAdb: A database of transfer RNA genes detected in

genomic sequence. Nucleic Acids Res. 37:D93–D97. doi:10.1093/nar/gkn787

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

Chang, C., J.L. Bowman, and E.M. Meyerowitz. 2016. Field Guide to Plant Model Systems. Cell 167:325–339. doi:10.1016/j.cell.2016.08.031

Chang, T.H., F.L. Hsieh, T.P. Ko, K.H. Teng, P.H. Liang, and A.H.J. Wang. 2010. Structure of a heterotetrameric geranyl pyrophosphate synthase from mint (Mentha piperita) reveals intersubunit regulation. Plant Cell 22:454–467. doi:10.1105/tpc.109.071738

Chatzivasileiou, A.O., V. Ward, S.M.B. Edgar, and G. Stephanopoulos. 2019. Two-step pathway for isoprenoid synthesis. Proc. Natl. Acad. Sci. U. S. A. 116:506–511. doi:10.1073/pnas.1812935116

Chen, F., W. Dong, J. Zhang, X. Guo, J. Chen, Z. Wang, Z. Lin, H. Tang, and L. Zhang. 2018. The sequenced angiosperm genomes and genome databases. Front. Plant Sci. 9:418. doi:10.3389/fpls.2018.00418

Chen, F., D. Tholl, J. Bohlmann, and E. Pichersky. 2011. The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. Plant J. 66:212–229. doi:10.1111/j.1365-313X.2011.04520.x

Chen, J., C. Lu, Y. Zhang, and H. Kuang. 2012. Miniature inverted-repeat transposable elements (MITEs) in rice were originated and amplified predominantly after the divergence of Oryza and Brachypodium and contributed considerable diversity to the species. Mob. Genet. Elements 2:127–132. doi:10.4161/mge.20773

Chen, Q., D. Fan, and G. Wang. 2015. Heteromeric geranyl(geranyl) diphosphate synthase is

involved in monoterpene biosynthesis in arabidopsis flowers. Mol. Plant 8:1434–1437. doi:10.1016/j.molp.2015.05.001

Chen, Z.J. 2010. Molecular mechanisms of polyploidy and hybrid vigor. Trends Plant Sci. 15:57–71. doi:10.1016/j.tplants.2009.12.003

Chevreux, B., T. Wetter, and S. Suhai. 1999. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Comput. Sci. Biol. Proc. Ger. Conf. Bioinforma. '99, GCB, Hann. Ger. 45–56

Chikhi, R., and P. Medvedev. 2014. Informed and automated k-mer size selection for genome assembly. Bioinformatics 30:31–37. doi:10.1093/bioinformatics/btt310

Choulet, F., T. Wicker, C. Rustenholz, E. Paux, J. Salse, P. Leroy, S. Schlub, M.C. le Paslier, G. Magdelenat, C. Gonthier, A. Couloux, H. Budak, J. Breen, M. Pumphrey, S. Liu, X. Kong, J. Jia, M. Gut, D. Brunel, J.A. Anderson, B.S. Gill, R. Appels, B. Keller, and C. Feuillet. 2010. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22:1686–1701. doi:10.1105/tpc.110.074187

Chu, C., X. Li, and Y. Wu. 2019. GAPPadder: A sensitive approach for closing gaps on draft genomes with short sequence reads. BMC Genomics 20:426. doi:10.1186/s12864-019-5703-4

Claros, M.G., R. Bautista, D. Guerrero-Fernández, H. Benzerki, P. Seoane, and N. Fernández-Pozo. 2012. Why assembling plant genome sequences is so challenging. Biology (Basel). 1:439–459. doi:10.3390/biology1020439

Closa, M., E. Vranová, C. Bortolotti, L. Bigler, M. Arró, A. Ferrer, and W. Gruissem. 2010. The Arabidopsis thaliana FPP synthase isozymes have overlapping and specific functions in

isoprenoid biosynthesis, and complete loss of FPP synthase activity causes early developmental arrest. Plant J. 63:512–525. doi:10.1111/j.1365-313X.2010.04253.x

Combes, M.C., A. Cenci, H. Baraille, B. Bertrand, and P. Lashermes. 2012. Homeologous gene expression in response to growing temperature in a recent allopolyploid (Coffea arabica L.). J. Hered. 103:36–46. doi:10.1093/jhered/esr120

Compeau, P.E.C., P.A. Pevzner, and G. Tesler. 2011. How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. 29:987–991. doi:10.1038/nbt.2023

Cui, L., P.K. Wall, J.H. Leebens-Mack, B.G. Lindsay, D.E. Soltis, J.J. Doyle, P.S. Soltis, J.E. Carlson, K. Arumuganathan, A. Barakat, V.A. Albert, H. Ma, and C.W. DePamphilis. 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res. 16:738–749. doi:10.1101/gr.4825606

Dalilan, S., M. Rezaei-Tavirani, M. Nabiuni, S. Heidari-Keshel, M. Zamanian Azodi, and H. Zali. 2013. Aqueous extract of lavender Angustifolia inhibits lymphocytes proliferation of Hodgkin's lymphoma patients. Iran. J. Cancer Prev. 6:201–208

Dayarian, A., T.P. Michael, and A.M. Sengupta. 2010. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. BMC Bioinformatics 11:345. doi:10.1186/1471-2105-11-345

Demissie, Z.A., M.A. Cella, L.S. Sarker, T.J. Thompson, M.R. Rheault, and S.S. Mahmoud. 2012. Cloning, functional characterization and genomic organization of 1,8-cineole synthases from Lavandula. Plant Mol. Biol. 79:393–411. doi:10.1007/s11103-012-9920-3

Demissie, Z.A., L.A.E. Erland, M.R. Rheault, and S.S. Mahmoud. 2013. The biosynthetic origin of irregular monoterpenes in lavandula: Isolation and biochemical characterization of a novel cis-prenyl diphosphate synthase gene, lavandulyl diphosphate synthase. J. Biol.

Chem. 288:6333–6341. doi:10.1074/jbc.M112.431171

Demissie, Z.A., L.S. Sarker, and S.S. Mahmoud. 2011. Cloning and functional characterization of β-phellandrene synthase from Lavandula angustifolia. Planta 233:685–696. doi:10.1007/s00425-010-1332-5

Denner, S.S. 2009. Lavandula angustifolia miller: English lavender. Holist. Nurs. Pract. 23:57–64. doi:10.1097/01.HNP.0000343210.56710.fc

Dennis, A.B., G.I. Ballesteros, S. Robin, L. Schrader, J. Bast, J. Berghöfer, L.W. Beukeboom, M. Belghazi, A. Bretaudeau, J. Buellesbach, E. Cash, D. Colinet, Z. Dumas, M. Errbii, P. Falabella, J.L. Gatti, E. Geuverink, J.D. Gibson, C. Hertaeg, S. Hartmann, E. Jacquin-Joly, M. Lammers, B.I. Lavandero, I. Lindenbaum, L. Massardier-Galata, C. Meslin, N. Montagné, N. Pak, M. Poirié, R. Salvia, C.R. Smith, D. Tagu, S. Tares, H. Vogel, T. Schwander, J.C. Simon, C.C. Figueroa, C. Vorburger, F. Legeai, and J. Gadau. 2020. Functional insights from the GC-poor genomes of two aphid parasitoids, Aphidius ervi and Lysiphlebus fabarum. BMC Genomics 21:376. doi:10.1186/s12864-020-6764-0

Denoeud, F., L. Carretero-Paulet, A. Dereeper, G. Droc, R. Guyot, M. Pietrella, C. Zheng, A. Alberti, F. Anthony, G. Aprea, J.M. Aury, P. Bento, M. Bernard, S. Bocs, C. Campa, A. Cenci, M.C. Combes, D. Crouzillat, C. Da Silva, L. Daddiego, F. De Bellis, S. Dussert, O. Garsmeur, T. Gayraud, V. Guignon, K. Jahn, V. Jamilloux, T. Joët, K. Labadie, T. Lan, J. Leclercq, M. Lepelley, T. Leroy, L.T. Li, P. Librado, L. Lopez, A. Muñoz, B. Noel, A. Pallavicini, G. Perrotta, V. Poncet, D. Pot, Priyono, M. Rigoreau, M. Rouard, J. Rozas, C. Tranchant-Dubreuil, R. VanBuren, Q. Zhang, A.C. Andrade, X. Argout, B. Bertrand, A. De Kochko, G. Graziosi, R.J. Henry, Jayarama, R. Ming, C. Nagai, S. Rounsley, D. Sankoff, G. Giuliano, V.A. Albert, P. Wincker, and P. Lashermes. 2014. The coffee genome provides

insight into the convergent evolution of caffeine biosynthesis. Science (80-.). 345:1181–1184. doi:10.1126/science.1255274

Desvillechabrol, D., C. Bouchier, S. Kennedy, and T. Cokelaer. 2018. Sequana coverage: Detection and characterization of genomic variations using running median and mixture models. Gigascience 7. doi:10.1093/gigascience/giy110

Dohmen, E., L.P.M. Kremer, E. Bornberg-Bauer, and C. Kemena. 2016. DOGMA: Domain-based transcriptome and proteome quality assessment. Bioinformatics 32:2577–2581. doi:10.1093/bioinformatics/btw231

Doležel, J., and J. Bartoš. 2005. Plant DNA flow cytometry and estimation of nuclear genome size. Ann. Bot. 95:99–110. doi:10.1093/aob/mci005

Doolittle, W.F., and C. Sapienza. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601–603. doi:10.1038/284601a0

Dragan, M.A., I. Moghul, A. Priyam, C. Bustos, and Y. Wurm. 2016. GeneValidator: Identify problems with protein-coding gene predictions. Bioinformatics 32:1559–1561. doi:10.1093/bioinformatics/btw015

Duarte, J.M., L. Cui, P.K. Wall, Q. Zhang, X. Zhang, J. Leebens-Mack, H. Ma, N. Altman, and C.W. DePamphilis. 2006. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. Mol. Biol. Evol. 23:469–478. doi:10.1093/molbev/msj051

Earl, D., K. Bradnam, J. St. John, A. Darling, D. Lin, J. Fass, H.O.K. Yu, V. Buffalo, D.R. Zerbino, M. Diekhans, N. Nguyen, P.N. Ariyaratne, W.K. Sung, Z. Ning, M. Haimel, J.T. Simpson, N.A. Fonseca, I. Birol, T.R. Docking, I.Y. Ho, D.S. Rokhsar, R. Chikhi, D. Lavenier, G. Chapuis, D. Naquin, N. Maillet, M.C. Schatz, D.R. Kelley, A.M. Phillippy, S.

Koren, S.P. Yang, W. Wu, W.C. Chou, A. Srivastava, T.I. Shaw, J.G. Ruby, P. Skewes-Cox, M. Betegon, M.T. Dimon, V. Solovyev, I. Seledtsov, P. Kosarev, D. Vorobyev, R. Ramirez-Gonzalez, R. Leggett, D. MacLean, F. Xia, R. Luo, Z. Li, Y. Xie, B. Liu, S. Gnerre, I. MacCallum, D. Przybylski, F.J. Ribeiro, T. Sharpe, G. Hall, P.J. Kersey, R. Durbin, S.D. Jackman, J.A. Chapman, X. Huang, J.L. DeRisi, M. Caccamo, Y. Li, D.B. Jaffe, R.E. Green, D. Haussler, I. Korf, and B. Paten. 2011. Assemblathon 1: A competitive assessment of de novo short read assembly methods. Genome Res. 21:2224–2241. doi:10.1101/gr.126599.111

Eckardt, N.A. 2004. Two genomes are better than one: Widespread paleopolyploidy in plants and evolutionary effects. Plant Cell 16:1647–1649. doi:10.1105/tpc.160710

Eichler, E.E. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends Genet. 17:661–669. doi:10.1016/S0168-9525(01)02492-1

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. DeWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. 2009. Real-time DNA sequencing from single polymerase molecules. Science (80-.). 323:133–138. doi:10.1126/science.1162986

Eilbeck, K., B. Moore, C. Holt, and M. Yandell. 2009. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics.

doi:10.1186/1471-2105-10-67

Ekblom, R., and J. Galindo. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity (Edinb). 107:1–15. doi:10.1038/hdy.2010.152

El-Gebali, S., J. Mistry, A. Bateman, S.R. Eddy, A. Luciani, S.C. Potter, M. Qureshi, L.J. Richardson, G.A. Salazar, A. Smart, E.L.L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S.C.E. Tosatto, and R.D. Finn. 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47:D427–D432. doi:10.1093/nar/gky995

Elsik, C.G., A.J. Mackey, J.T. Reese, N. V. Milshina, D.S. Roos, and G.M. Weinstock. 2007. Creating a honey bee consensus gene set. Genome Biol. 8:R13. doi:10.1186/gb-2007-8-1-r13

Erland, L.A.E., and S.S. Mahmoud. 2014. An efficient method for regeneration of lavandin (Lavandula x intermedia cv. 'Grosso'). Vitr. Cell. Dev. Biol. - Plant 50:646–654. doi:10.1007/s11627-014-9614-4

Estep, M.C., J.D. Debarry, and J.L. Bennetzen. 2013. The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. Heredity (Edinb). 110:194–204. doi:10.1038/hdy.2012.99

Evandri, M.G., L. Battinelli, C. Daniele, S. Mastrangelo, P. Bolle, and G. Mazzanti. 2005. The antimutagenic activity of Lavandula angustifolia (lavender) essential oil in the bacterial reverse mutation assay. Food Chem. Toxicol. 43:1381–1387. doi:10.1016/j.fct.2005.03.013

Falara, V., T.A. Akhtar, T.T.H. Nguyen, E.A. Spyropoulou, P.M. Bleeker, I. Schauvinhold, Y. Matsuba, M.E. Bonini, A.L. Schilmiller, R.L. Last, R.C. Schuurink, and E. Pichersky. 2011. The tomato terpene synthase gene family. Plant Physiol. 157:770–789. doi:10.1104/pp.111.179648

Fanning, S., S. Proos, K. Jordan, and S. Srikumar. 2017. A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. Front. Microbiol. 8. doi:10.3389/fmicb.2017.01829

Farrant, G.K., M. Hoebeke, F. Partensky, G. Andres, E. Corre, and L. Garczarek. 2015. WiseScaffolder: An algorithm for the semi-automatic scaffolding of Next Generation Sequencing data. BMC Bioinformatics 16:281. doi:10.1186/s12859-015-0705-y

Fay, J.C., and C.I. Wu. 2003. Sequence Divergence, Functional Constraint, and Selection in Protein Evolution. Annu. Rev. Genomics Hum. Genet.. doi:10.1146/annurev.genom.4.020303.162528

Feldman, M., and A.A. Levy. 2009. Genome evolution in allopolyploid wheat-a revolutionary reprogramming followed by gradual changes. J. Genet. Genomics 36:511–518. doi:10.1016/S1673-8527(08)60142-3

Feng, X., J. Jiang, A. Padhi, C. Ning, J. Fu, A. Wang, R. Mrode, and J.F. Liu. 2017. Characterization of genome-wide segmental duplications reveals a common genomic feature of association with immunity among domestic animals. BMC Genomics 18:293. doi:10.1186/s12864-017-3690-x

Feng, Y. 2003. Plant MITEs: useful tools for plant genetics and genomics.. Genomics, proteomics Bioinforma. / Beijing Genomics Inst. 1:90–100. doi:10.1016/s1672-0229(03)01013-1

Ferragina, P., and G. Manzini. 2000. . Opportunistic data structures with applications. Pages 390–398 in Annual Symposium on Foundations of Computer Science - Proceedings

Feschotte, C., and E.J. Pritham. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. 41:331–368. doi:10.1146/annurev.genet.40.110405.090448

Finnegan, D.J. 1989. Eukaryotic transposable elements and genome evolution. Trends Genet. 5:103–107. doi:10.1016/0168-9525(89)90039-5

Flavell, R.B. 1986. Repetitive DNA and chromosome evolution in plants.. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 312:227–242. doi:10.1098/rstb.1986.0004

Flynn, J.M., R. Hubley, C. Goubert, J. Rosen, A.G. Clark, C. Feschotte, and A.F. Smit. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. U. S. A. 117:9451–9457. doi:10.1073/pnas.1921046117

Fomeju, B.F., C. Falentin, G. Lassalle, M.J. Manzanares-Dauleux, and R. Delourme. 2015. Comparative genomic analysis of duplicated homoeologous regions involved in the resistance of brassica napus to stem canker. Front. Plant Sci. 6:772. doi:10.3389/fpls.2015.00772

Galbraith, D.W., K.R. Harkins, J.M. Maddox, N.M. Ayres, D.P. Sharma, and E. Firoozabady. 1983. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. Science (80-.). 220:1049–1051. doi:10.1126/science.220.4601.1049

Gao, S., D. Bertrand, B.K.H. Chia, and N. Nagarajan. 2016. OPERA-LG: Efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees. Genome Biol. 17:102. doi:10.1186/s13059-016-0951-y

Gao, S., N. Nagarajan, and W.K. Sung. 2011. Opera: Reconstructing Optimal Genomic Scaffolds with High-Throughput Paired-End Sequences. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 6577 LNBI:437–451. doi:10.1007/978-3-642-20036-6_40

Gaubert, H., D.H. Sanchez, H.G. Drost, and J. Paszkowski. 2017. Developmental restriction of retrotransposition activated in Arabidopsis by environmental stress. Genetics 207:813–821.

doi:10.1534/genetics.117.300103

Geer, L.Y., A. Marchler-Bauer, R.C. Geer, L. Han, J. He, S. He, C. Liu, W. Shi, and S.H. Bryant. 2009. The NCBI BioSystems database. Nucleic Acids Res. 38:D492-6. doi:10.1093/nar/gkp858

Geib, S.M., B. Hall, T. Derego, F.T. Bremer, K. Cannoles, and S.B. Sim. 2018. Genome Annotation Generator: a simple tool for generating and correcting WGS annotation tables for NCBI submission. Gigascience 7:1–5. doi:10.1093/gigascience/giy018

Gill, N., P. SanMiguel, B.D.S. Dhillon, B. Abernathy, H.R. Kim, L. Stein, D. Ware, R. Wing, and S.A. Jackson. 2010. Dynamic Oryza genomes: Repetitive DNA sequences as genome modeling agents. Rice 3:251–269. doi:10.1007/s12284-010-9054-7

Giray, F.H. 2018. An Analysis of World Lavender Oil Markets and Lessons for Turkey. J. Essent. Oil-Bearing Plants 21:1612–1623. doi:10.1080/0972060X.2019.1574612

Girgis, H.Z. 2015. Red: An intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics 16:227. doi:10.1186/s12859-015-0654-5

Giussani, L.M., J.H. Cota-Sánchez, F.O. Zuloaga, and E.A. Kellogg. 2001. A molecular phylogeny of the grass subfamily Panicoideae (Poaceae) shows multiple origins of C4 photosynthesis. Am. J. Bot. 88:1993–2012. doi:10.2307/3558427

Gnerre, S., I. MacCallum, D. Przybylski, F.J. Ribeiro, J.N. Burton, B.J. Walker, T. Sharpe, G. Hall, T.P. Shea, S. Sykes, A.M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E.S. Lander, and D.B. Jaffe. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U. S. A. 108:1513–1518. doi:10.1073/pnas.1017351108

Goerner-Potvin, P., and G. Bourque. 2018. Computational tools to unmask transposable

elements. Nat. Rev. Genet. 19:688–704. doi:10.1038/s41576-018-0050-x

Goff, S.A., D. Ricke, T.H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchison, C. Martin, F. Katagiri, B.M. Lange, T. Moughamer, Y. Xia, P. Budworth, J. Zhong, T. Miguel, U. Paszkowski, S. Zhang, M. Colbert, W.L. Sun, L. Chen, B. Cooper, S. Park, T.C. Wood, L. Mao, P. Quail, R. Wing, R. Deans, Y. Yu, A. Zharkikh, R. Shen, S. Sahasrabudhe, A. Thomas, R. Cannings, A. Gutin, D. Pruss, J. Reid, S. Tavtigian, J. Mitchell, G. Eldredge, T. Scholl, R.M. Miller, S. Bhatnagar, N. Adey, T. Rubano, N. Tusneem, R. Robinson, J. Feldhaus, T. Macalma, A. Oliphant, and S. Briggs. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science (80-.). 296:92–100. doi:10.1126/science.1068275

Gonnella, G., and S. Kurtz. 2012. Readjoiner: a fast and memory efficient string graph-based sequence assembler. BMC Bioinformatics 13:82. doi:10.1186/1471-2105-13-82

González, L.G., and M.K. Deyholos. 2012. Identification, characterization and distribution of transposable elements in the flax (Linum usitatissimum L.) genome. BMC Genomics 13:644. doi:10.1186/1471-2164-13-644

Goodstein, D.M., S. Shu, R. Howson, R. Neupane, R.D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, and D.S. Rokhsar. 2012. Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res. 40:D1178–D1186. doi:10.1093/nar/gkr944

Grover, C.E., and J.F. Wendel. 2010. Recent Insights into Mechanisms of Genome Size Change in Plants. J. Bot. 2010:1–8. doi:10.1155/2010/382732

Guo, Y.L. 2013. Gene family evolution in green plants with emphasis on the origination and evolution of Arabidopsis thaliana genes. Plant J. 73:941–951. doi:10.1111/tpj.12089

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. 2013. QUAST: Quality assessment tool for

genome assemblies. Bioinformatics 29:1072–1075. doi:10.1093/bioinformatics/btt086

Haas, B.J., A.L. Delcher, S.M. Mount, J.R. Wortman, R.K. Smith, L.I. Hannick, R. Maiti, C.M. Ronning, D.B. Rusch, C.D. Town, S.L. Salzberg, and O. White. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31:5654–5666. doi:10.1093/nar/gkg770

Haas, B.J., A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M.D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C.N. Dewey, R. Henschel, R.D. Leduc, N. Friedman, and A. Regev. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494–1512. doi:10.1038/nprot.2013.084

Haas, B.J., S.L. Salzberg, W. Zhu, M. Pertea, J.E. Allen, J. Orvis, O. White, C.R. Robin, and J.R. Wortman. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9:R7. doi:10.1186/gb-2008-9-1-r7

Haberer, G., S. Young, A.K. Bharti, H. Gundlach, C. Raymond, G. Fuks, E. Butler, R.A. Wing, S. Rounsley, B. Birren, C. Nusbaum, K.F.X. Mayer, and J. Messing. 2005. Structure and architecture of the maize genome. Plant Physiol. 139:1612–1624. doi:10.1104/pp.105.068718

Hamilton, J.P., and C.R. Buell. 2014. Timber! Felling the loblolly pine genome. Genome Biol. 15:111. doi:10.1186/gb4170

Hardie, D.C., T.R. Gregory, and P.D.N. Hebert. 2002. From pixels to picograms: A beginners' guide to genome quantification by Feulgen image analysis densitometry. J. Histochem.

Cytochem. 50:735–749. doi:10.1177/002215540205000601

Hayashi, K. ichiro, H. Kawaide, M. Notomi, Y. Sakigi, A. Matsuo, and H. Nozaki. 2006. Identification and functional analysis of bifunctional ent-kaurene synthase from the moss Physcomitrella patens. FEBS Lett. 580:6175–6181. doi:10.1016/j.febslet.2006.10.018

Hernandez, D., P. François, L. Farinelli, M. Østerås, and J. Schrenzel. 2008. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. Genome Res. 18:802–809. doi:10.1101/gr.072033.107

Herrmann, L., D. Lesueur, L. Bräu, J. Davison, T. Jairus, H. Robain, A. Robin, M. Vasar, W. Wiriyakitnateekul, and M. Öpik. 2016. Diversity of root-associated arbuscular mycorrhizal fungal communities in a rubber tree plantation chronosequence in Northeast Thailand. Mycorrhiza 26:863–877. doi:10.1007/s00572-016-0720-5

Heydari, M., G. Miclotte, P. Demeester, Y. Van de Peer, and J. Fostier. 2017. Evaluation of the impact of Illumina error correction tools on de novo genome assembly. BMC Bioinformatics 18. doi:10.1186/s12859-017-1784-8

Hoff, K.J., A. Lomsadze, M. Borodovsky, and M. Stanke. 2019. Whole-genome annotation with BRAKER. Methods Mol. Biol. 1962:65–95. doi:10.1007/978-1-4939-9173-0_5

Holt, C., and M. Yandell. 2011. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491. doi:10.1186/1471-2105-12-491

Hu, J., Y. Zheng, and X. Shang. 2018. MiteFinderII: A novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes. BMC Med. Genomics 11. doi:10.1186/s12920-018-0418-y

Hu, T.T., P. Pattyn, E.G. Bakker, J. Cao, J.F. Cheng, R.M. Clark, N. Fahlgren, J.A. Fawcett, J.

Grimwood, H. Gundlach, G. Haberer, J.D. Hollister, S. Ossowski, R.P. Ottilar, A.A. Salamov, K. Schneeberger, M. Spannagl, X. Wang, L. Yang, M.E. Nasrallah, J. Bergelson, J.C. Carrington, B.S. Gaut, J. Schmutz, K.F.X. Mayer, Y. Van De Peer, I. V. Grigoriev, M. Nordborg, D. Weigel, and Y.L. Guo. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat. Genet. 43:476–483. doi:10.1038/ng.807

Huang, X., and A. Madan. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9:868–877. doi:10.1101/gr.9.9.868

Huang, Y.T., and C.F. Liao. 2016. Integration of string and de Bruijn graphs for genome assembly. Bioinformatics 32:1301–1307. doi:10.1093/bioinformatics/btw011

Humann, J.L., T. Lee, S. Ficklin, and D. Main. 2019. Structural and functional annotation of eukaryotic genomes with GenSAS. Methods Mol. Biol. 1962:29–51. doi:10.1007/978-1-4939-9173-0_3

Ibarra-Laclette, E., E. Lyons, G. Hernández-Guzmán, C.A. Pérez-Torres, L. Carretero-Paulet, T.H. Chang, T. Lan, A.J. Welch, M.J.A. Juárez, J. Simpson, A. Fernández-Cortés, M. Arteaga-Vázquez, E. Góngora-Castillo, G. Acevedo-Hernández, S.C. Schuster, H. Himmelbauer, A.E. Minoche, S. Xu, M. Lynch, A. Oropeza-Aburto, S.A. Cervantes-Pérez, M. De Jesús Ortega-Estrada, J.I. Cervantes-Luevano, T.P. Michael, T. Mockler, D. Bryant, A. Herrera-Estrella, V.A. Albert, and L. Herrera-Estrella. 2013. Architecture and evolution of a minute plant genome. Nature 498:94–98. doi:10.1038/nature12132

Idury, R.M., and M.S. Waterman. 1995. A New Algorithm for DNA Sequence Assembly. J. Comput. Biol. 2:291–306. doi:10.1089/cmb.1995.2.291

Jaillon, O., J.M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D.

Jublot, J. Poulain, C. Bruyère, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M.E. Pè, G. Valle, M. Morgante, M. Caboche, A.F. Adam-Blondon, J. Weissenbach, F. Quétier, and P. Wincker. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467. doi:10.1038/nature06148

Jiao, Y., and A.H. Paterson. 2014. Polyploidy-associated genome modifications during land plant evolution. Philos. Trans. R. Soc. B Biol. Sci. 369. doi:10.1098/rstb.2013.0355

Jiao, Y., P. Peluso, J. Shi, T. Liang, M.C. Stitzer, B. Wang, M.S. Campbell, J.C. Stein, X. Wei, C.S. Chin, K. Guill, M. Regulski, S. Kumari, A. Olson, J. Gent, K.L. Schneider, T.K. Wolfgruber, M.R. May, N.M. Springer, E. Antoniou, W.R. McCombie, G.G. Presting, M. McMullen, J. Ross-Ibarra, R.K. Dawe, A. Hastie, D.R. Rank, and D. Ware. 2017. Improved maize reference genome with single-molecule technologies. Nature 546:524–527. doi:10.1038/nature22971

Jullien, F., S. Moja, A. Bony, S. Legrand, C. Petit, T. Benabdelkader, K. Poirot, S. Fiorucci, Y. Guitton, F. Nicolè, S. Baudino, and J.L. Magnard. 2014. Isolation and functional characterization of a τ-cadinol synthase, a new sesquiterpene synthase from Lavandula angustifolia. Plant Mol. Biol. 84:227–241. doi:10.1007/s11103-013-0131-3

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110:462–467. doi:10.1159/000084979

Kara, N., and H. Baydar. 2013. Determination of lavender and lavandin cultivars (Lavandula sp.) containing high quality essential oil in Isparta, Turkey. Turkish J. F. Crop. 18:58–65. doi:10.17557/tjfc.86804

Kaul, S., H.L. Koo, J. Jenkins, M. Rizzo, T. Rooney, L.J. Tallon, T. Feldblyum, W. Nierman, M.I. Benito, X. Lin, C.D. Town, J.C. Venter, C.M. Fraser, S. Tabata, Y. Nakamura, T. Kaneko, S. Sato, E. Asamizu, T. Kato, H. Kotani, S. Sasamoto, J.R. Ecker, A. Theologis, N.A. Federspiel, C.J. Palm, B.I. Osborne, P. Shinn, K. Dewar, C.J. Kim, E. Buehler, P. Dunn, Q. Chao, H. Chen, A. Theologis, B.I. Osborne, V.S. Vysotskaia, C.A. Lenz, C.J. Kim, N.F. Hansen, S.X. Liu, E. Buehler, H. Alta, H. Sakano, P. Dunn, B. Lam, P.K. Pham, Q. Chao, M. Nguyen, G. Yu, H. Chen, A. Southwick, J.M. Lee, M. Miranda, M.J. Toriumi, R.W. Davis, N.A. Federspiel, C.J. Palm, A.B. Conway, L. Conn, N.F. Hansen, A. Hootan, B. Lam, R. Wambutt, G. Murphy, A. Düsterhöft, W. Stiekema, T. Pohl, K.D. Entian, N. Terryn, G. Volckaert, M. Salanoubat, N. Choisne, F. Artiguenave, J. Weissenbach, F. Quetier, M. Rieger, W. Ansorge, M. Unseld, B. Fartmann, G. Valle, R.K. Wilson, M. Sekhon, K. Pepin, J. Murray, D. Johnson, L. Hillier, M. de la Bastide, E. Huang, L. Spiegel, L. Gnoj, K. Habermann, N. Dedhia, L. Parnell, R. Preston, M. Marra, W.R. McCombie, E. Chen, R. Martienssen, K. Mayer, K. Lemcke, B. Haas, D. Haase, S. Rudd, H. Schoof, D. Frishman, B. Morgenstern, P. Zaccaria, H.W. Mewes, O. White, T.H. Creasy, C. Bielke, R. Maiti, J. Peterson, M. Ermolaeva, M. Pertea, J. Quackenbush, N. Volfovsky, D. Wu, S.L. Salzberg, M. Bevan, T.M. Lowe, S. Rounsley, D. Bush, S. Subramaniam, I. Levin, S. Norris, R. Schmidt, A. Acarkan, I. Bancroft, A. Brennicke, J.A. Eisen, T. Bureau, B.A. Legault, Q.H. Le, N. Agrawal, Z. Yu, G.P. Copenhaver, S. Luo, D. Preuss, C.S. Pikaard, I.T. Paulsen, M. Sussman, A.B. Britt, D.A. Selinger, R. Pandey, V.L. Chandler, R.A.

Jorgensen, D.W. Mount, C. Pikaard, G. Juergens, E.M. Meyerowitz, J. Dangl, J.D.G. Jones, M. Chen, J. Chory, and C. Somerville. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815. doi:10.1038/35048692

Kawahara, Y., M. de la Bastide, J.P. Hamilton, H. Kanamori, W.R. Mccombie, S. Ouyang, D.C. Schwartz, T. Tanaka, J. Wu, S. Zhou, K.L. Childs, R.M. Davidson, H. Lin, L. Quesada-Ocampo, B. Vaillancourt, H. Sakai, S.S. Lee, J. Kim, H. Numa, T. Itoh, C.R. Buell, and T. Matsumoto. 2013. Improvement of the oryza sativa nipponbare reference genome using next generation sequence and optical map data. Rice 6:3–10. doi:10.1186/1939-8433-6-4

Kent, W.J. 2002. BLAT - The BLAST-like alignment tool. Genome Res. 12:656–664. doi:10.1101/gr.229202.

Kew, R.B.G. 2008. Seed Information Database (SID). Version 7.1. http://data.kew.org/sid/

Kim, Y.B., S.M. Kim, M.K. Kang, T. Kuzuyama, J.K. Lee, S.C. Park, S.C. Shin, and S.U. Kim. 2009. Regulation of resin acid synthesis in Pinus densiflora by differential transcription of genes encoding multiple 1-deoxy-d-xylulose 5-phosphate synthase and 1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase genes. Tree Physiol. 29:737–749. doi:10.1093/treephys/tpp002

Kopcsayová, D., and E. Vranová. 2019. Functional Gene Network of Prenyltransferases in Arabidopsis thaliana. Molecules 24:4556. doi:10.3390/molecules24244556

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

Koulivand, P.H., M. Khaleghi Ghadiri, and A. Gorji. 2013. Lavender and the nervous system. Evidence-based Complement. Altern. Med. 2013. doi:10.1155/2013/681304

Kriventseva, E. V., F. Tegenfeldt, T.J. Petty, R.M. Waterhouse, F.A. Simão, I.A. Pozdnyakov, P.

Ioannidis, and E.M. Zdobnov. 2015. OrthoDB v8: Update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res. 43:D250–D256. doi:10.1093/nar/gku1220

Kufleitner, M. 2009. On bijective variants of the Burrows-Wheeler transform. Proc. Prague Stringology Conf. 2009 65–79. doi:10.1.1.37.6774

Kusano, H., H. Li, H. Minami, Y. Kato, H. Tabata, and K. Yazaki. 2019. Evolutionary developments in plant specialized metabolism, exemplified by two transferase families. Front. Plant Sci. 10:794. doi:10.3389/fpls.2019.00794

Kyriakidou, M., H.H. Tai, N.L. Anglin, D. Ellis, and M. V. Strömvik. 2018. Current strategies of polyploid plant genome sequence assembly. Front. Plant Sci. 871:15. doi:10.3389/fpls.2018.01660

Lagesen, K., P. Hallin, E.A. Rødland, H.H. Stærfeldt, T. Rognes, and D.W. Ussery. 2007. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35:3100–3108. doi:10.1093/nar/gkm160

Landmann, C., B. Fink, M. Festner, M. Dregus, K.H. Engel, and W. Schwab. 2007. Cloning and functional characterization of three terpene synthases from lavender (Lavandula angustifolia). Arch. Biochem. Biophys. 465:417–429. doi:10.1016/j.abb.2007.06.011

Lane, A., A. Boecklemann, G.N. Woronuk, L. Sarker, and S.S. Mahmoud. 2010. A genomics resource for investigating regulation of essential oil production in Lavandula angustifolia. Planta 231:835–845. doi:10.1007/s00425-009-1090-4

Lee, E., G.A. Helt, J.T. Reese, M.C. Munoz-Torres, C.P. Childers, R.M. Buels, L. Stein, I.H. Holmes, C.G. Elsik, and S.E. Lewis. 2013. Web Apollo: A web-based genomic annotation editing platform. Genome Biol. 14:R93. doi:10.1186/gb-2013-14-8-r93

Lee, S.-I., and N.-S. Kim. 2014. Transposable Elements and Genome Size Variations in Plants. Genomics Inform. 12:87. doi:10.5808/gi.2014.12.3.87

Lerat, E. 2010. Identifying repeats and transposable elements in sequenced genomes: How to find your way through the dense forest of programs. Heredity (Edinb). 104:520–533. doi:10.1038/hdy.2009.165

Li, F., G. Fan, C. Lu, G. Xiao, C. Zou, R.J. Kohel, Z. Ma, H. Shang, X. Ma, J. Wu, X. Liang, G. Huang, R.G. Percy, K. Liu, W. Yang, W. Chen, X. Du, C. Shi, Y. Yuan, W. Ye, X. Liu, X. Zhang, W. Liu, H. Wei, S. Wei, G. Huang, X. Zhang, S. Zhu, H. Zhang, F. Sun, X. Wang, J. Liang, J. Wang, Q. He, L. Huang, J. Wang, J. Cui, G. Song, K. Wang, X. Xu, J.Z. Yu, Y. Zhu, and S. Yu. 2015. Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. Nat. Biotechnol. 33:524–530. doi:10.1038/nbt.3208

Li, F., G. Fan, K. Wang, F. Sun, Y. Yuan, G. Song, Q. Li, Z. Ma, C. Lu, C. Zou, W. Chen, X. Liang, H. Shang, W. Liu, C. Shi, G. Xiao, C. Gou, W. Ye, X. Xu, X. Zhang, H. Wei, Z. Li, G. Zhang, J. Wang, K. Liu, R.J. Kohel, R.G. Percy, J.Z. Yu, Y.X. Zhu, J. Wang, and S. Yu. 2014. Genome sequence of the cultivated cotton gossypium arboreum. Nat. Genet. 46:567–572. doi:10.1038/ng.2987

Li, F.W., and A. Harkess. 2018. A guide to sequence your favorite plant genomes. Appl. Plant Sci. 6:e1030. doi:10.1002/aps3.1030

Li, H. 2012. Exploring single-sample snp and indel calling with whole-genome de novo assembly. Bioinformatics 28:1838–1844. doi:10.1093/bioinformatics/bts280

Li, J., Y. Wang, Y. Dong, W. Zhang, D. Wang, H. Bai, K. Li, H. Li, and L. Shi. 2021. The chromosome-based lavender genome provides new insights into Lamiaceae evolution and

terpenoid biosynthesis. Hortic. Res. 8:53. doi:10.1038/s41438-021-00490-6

Li, P., Z.C. Qi, L.X. Liu, T. Ohi-Toma, J. Lee, T.H. Hsieh, C.X. Fu, K.M. Cameron, and Y.X. Qiu. 2017. Molecular phylogenetics and biogeography of the mint tribe Elsholtzieae (Nepetoideae, Lamiaceae), with an emphasis on its diversification in East Asia. Sci. Rep. 7:2057. doi:10.1038/s41598-017-02157-6

Li, R., W. Fan, G. Tian, H. Zhu, L. He, J. Cai, Q. Huang, Q. Cai, B. Li, Y. Bai, Z. Zhang, Y. Zhang, W. Wang, J. Li, F. Wei, H. Li, M. Jian, J. Li, Z. Zhang, R. Nielsen, D. Li, W. Gu, Z. Yang, Z. Xuan, O.A. Ryder, F.C.C. Leung, Y. Zhou, J. Cao, X. Sun, Y. Fu, X. Fang, X. Guo, B. Wang, R. Hou, F. Shen, B. Mu, P. Ni, R. Lin, W. Qian, G. Wang, C. Yu, W. Nie, J. Wang, Z. Wu, H. Liang, J. Min, Q. Wu, S. Cheng, J. Ruan, M. Wang, Z. Shi, M. Wen, B. Liu, X. Ren, H. Zheng, D. Dong, K. Cook, G. Shan, H. Zhang, C. Kosiol, X. Xie, Z. Lu, H. Zheng, Y. Li, C.C. Steiner, T.T.Y. Lam, S. Lin, Q. Zhang, G. Li, J. Tian, T. Gong, H. Liu, D. Zhang, L. Fang, C. Ye, J. Zhang, W. Hu, A. Xu, Y. Ren, G. Zhang, M.W. Bruford, Q. Li, L. Ma, Y. Guo, N. An, Y. Hu, Y. Zheng, Y. Shi, Z. Li, Q. Liu, Y. Chen, J. Zhao, N. Qu, S. Zhao, F. Tian, X. Wang, H. Wang, L. Xu, X. Liu, T. Vinar, Y. Wang, T.W. Lam, S.M. Yiu, S. Liu, H. Zhang, D. Li, Y. Huang, X. Wang, G. Yang, Z. Jiang, J. Wang, N. Qin, L. Li, J. Li, L. Bolund, K. Kristiansen, G.K.S. Wong, M. Olson, X. Zhang, S. Li, H. Yang, J. Wang, and J. Wang. 2010. The sequence and de novo assembly of the giant panda genome. Nature 463:311–317. doi:10.1038/nature08696

Li, X., and M.S. Waterman. 2003. Estimating the repeat structure and length of DNA sequences using ℓ-tuples. Genome Res. 13:1916–1922. doi:10.1101/gr.1251803

Liang, P., Y. Zhang, K. Lin, and J. Hu. 2014. . A fast sequence assembly method based on compressed data structures. Pages 326–329 in 2014 36th Annual International Conference

of the IEEE Engineering in Medicine and Biology Society, EMBC 2014

Liao, X., M. Li, Y. Zou, F.X. Wu, Yi-Pan, and J. Wang. 2019. Current challenges and solutions of de novo assembly. Quant. Biol. 7:90–109. doi:10.1007/s40484-019-0166-9

Liu, B., Y. Shi, J. Yuan, X. Hu, H. Zhang, N. Li, Z. Li, Y. Chen, D. Mu, and W. Fan. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv: Genomics

Lohse, M., A. Nagel, T. Herter, P. May, M. Schroda, R. Zrenner, T. Tohge, A.R. Fernie, M. Stitt, and B. Usadel. 2014. Mercator: A fast and simple web server for genome scale functional annotation of plant sequence data. Plant, Cell Environ. 37:1250–1258. doi:10.1111/pce.12231

Lowe, T.M., and S.R. Eddy. 1996. TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25:955–964. doi:10.1093/nar/25.5.0955

Lu, C., J. Chen, Y. Zhang, Q. Hu, W. Su, and H. Kuang. 2012. Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in oryza sativa. Mol. Biol. Evol. 29:1005–1017. doi:10.1093/molbev/msr282

Lu, H., F. Giordano, and Z. Ning. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics, Proteomics Bioinforma. 14:265–279. doi:10.1016/j.gpb.2016.05.004

Lukaszewski, A.J., A. Alberti, A. Sharpe, A. Kilian, A.M. Stanca, B. Keller, B.J. Clavijo, B. Friebe, B. Gill, B. Wulff, B. Chapman, B. Steuernagel, C. Feuillet, C. Viseux, C. Pozniak, D.S. Rokhsar, D. Klassen, D. Edwards, E. Akhunov, E. Paux, F. Alfama, F. Choulet, F. Kobayashi, G.J. Muehlbauer, H. Quesneville, H. Šimková, H. Rimbert, H. Gundlach, H.

Budak, H. Sakai, H. Handa, H. Kanamori, J. Batley, J. Vrána, J. Rogers, J. Číhalíková, J. Doležel, J. Chapman, J.A. Poland, J. Wu, J. Khurana, J. Wright, K.C. Bader, K. Eversole, K. Barry, K. McLay, K.F.X. Mayer, K. Singh, L. Clissold, L. Pingault, L. Couderc, L. Cattivelli, M. Spannagl, M. Kubaláková, M. Caccamo, M. Mascher, M. Bellgard, M. Pfeifer, M. Zytnicki, M. Febrer, M. Alaux, M.M. Martis, M. Loaec, M. Colaiacovo, N.K. Singh, N. Glover, N. Guilhot, N. Stein, O.A. Olsen, P.R. Maclachlan, P. Chhuneja, P. Wincker, P. Sourdille, P. Faccioli, R.H. Ramirez-Gonzalez, R. Waugh, R. Šperková, R. Knox, R. Appels, S. Sharma, S. Ayling, S. Praud, S. Wang, S. Lien, S.R. Sandve, T. Matsumoto, T.R. Endo, T. Itoh, T. Nussbaumer, T. Wicker, T. Tanaka, U. Scholz, V. Barbe, V. Jamilloux, Y. Ogihara, and Z. Dubská. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science (80-.). 345. doi:10.1126/science.1251788

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D.W. Cheung, S.M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.W. Lam, and J. Wang. 2012. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. Gigascience 1:2047. doi:10.1186/2047-217X-1-18

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D.W. Cheung, S.M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.W. Lam, and J. Wang. 2015. Erratum to “SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler” [GigaScience, (2012), 1, 18]. Gigascience 4:1. doi:10.1186/s13742-015-0069-2

MacCallum, I., D. Przybylski, S. Gnerre, J. Burton, I. Shlyakhter, A. Gnirke, J. Malek, K. McKernan, S. Ranade, T.P. Shea, L. Williams, S. Young, C. Nusbaum, and D.B. Jaffe. 2009. ALLPATHS 2: Small genomes assembled accurately and with high continuity from short paired reads. Genome Biol. 10:r103. doi:10.1186/gb-2009-10-10-r103

Madden, T. 2013. The BLAST Sequence Analysis Tool

Magrane, M., and U. Consortium. 2011. UniProt Knowledgebase: A hub of integrated protein data. Database 2011. doi:10.1093/database/bar009

Malli, R.P.N., A.M. Adal, L.S. Sarker, P. Liang, and S.S. Mahmoud. 2019. De novo sequencing of the Lavandula angustifolia genome reveals highly duplicated and optimized features for essential oil production. Planta 249:251–256. doi:10.1007/s00425-018-3012-9

Marçais, G., and C. Kingsford. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–770. doi:10.1093/bioinformatics/btr011

Marcussen, T., S.R. Sandve, L. Heier, M. Spannagl, M. Pfeifer, K.S. Jakobsen, B.B.H. Wulff, B. Steuernagel, K.F.X. Mayer, O.A. Olsen, J. Rogers, J. Doležel, C. Pozniak, K. Eversole, C. Feuillet, B. Gill, B. Friebe, A.J. Lukaszewski, P. Sourdille, T.R. Endo, M. Kubaláková, J. Šíhalíková, Z. Dubská, J. Vrána, R. Šperková, H. Šimková, M. Febrer, L. Clissold, K. McLay, K. Singh, P. Chhuneja, N.K. Singh, J. Khurana, E. Akhunov, F. Choulet, A. Alberti, V. Barbe, P. Wincker, H. Kanamori, F. Kobayashi, T. Itoh, T. Matsumoto, H. Sakai, T. Tanaka, J. Wu, Y. Ogihara, H. Handa, P.R. Maclachlan, A. Sharpe, D. Klassen, D. Edwards, J. Batley, S.R. Sandve, S. Lien, B. Wulff, M. Caccamo, S. Ayling, R.H. Ramirez-Gonzalez, B.J. Clavijo, J. Wright, M.M. Martis, M. Mascher, J. Chapman, J.A. Poland, U. Scholz, K. Barry, R. Waugh, D.S. Rokhsar, G.J. Muehlbauer, N. Stein, H. Gundlach, M. Zytnicki, V. Jamilloux, H. Quesneville, T. Wicker, P. Faccioli, M. Colaiacovo, A.M.

Stanca, H. Budak, L. Cattivelli, N. Glover, L. Pingault, E. Paux, S. Sharma, R. Appels, M. Bellgard, B. Chapman, T. Nussbaumer, K.C. Bader, H. Rimbert, S. Wang, R. Knox, A. Kilian, M. Alaux, F. Alfama, L. Couderc, N. Guilhot, C. Viseux, M. Loaec, B. Keller, and S. Praud. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. Science (80-.). 345. doi:10.1126/science.1250092

Mardis, E.R. 2008. Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. 9:387–402. doi:10.1146/annurev.genom.9.081307.164359

Margarido, G.R.A., and D. Heckerman. 2015. ConPADE: genome assembly ploidy estimation from next-generation sequencing data. PLoS Comput. Biol. 11:e1004229. doi:10.1371/journal.pcbi.1004229

Marschinke, F., and I. Strömberg. 2008. Dual effects of TNFα on nerve fiber formation from ventral mesencephalic organotypic tissue cultures.

Martin, D.M., S. Aubourg, M.B. Schouwey, L. Daviet, M. Schalk, O. Toub, S.T. Lund, and J. Bohlmann. 2010. Functional Annotation, Genome Organization and Phylogeny of the Grapevine (Vitis vinifera) Terpene Synthase Gene Family Based on Genome Assembly, FLcDNA Cloning, and Enzyme Assays. BMC Plant Biol. 10:226. doi:10.1186/1471-2229-10-226

McCLINTOCK, B. 1950. The origin and behavior of mutable loci in maize.. Proc. Natl. Acad. Sci. U. S. A. 36:344–355. doi:10.1073/pnas.36.6.344

Melters, D.P., L. V. Paliulis, I.F. Korf, and S.W.L. Chan. 2012. Holocentric chromosomes: Convergent evolution, meiotic adaptations, and genomic analysis. Chromosom. Res. 20:579–593. doi:10.1007/s10577-012-9292-1

Mendoza-Poudereux, I., E. Kutzner, C. Huber, J. Segura, I. Arrillaga, and W. Eisenreich. 2017.

Dynamics of monoterpene formation in spike lavender plants. Metabolites 7:65. doi:10.3390/metabo7040065

Meyers, L.A., and D.A. Levin. 2006. on the Abundance of Polyploids in Flowering Plants. Evolution (N. Y). 60:1198. doi:10.1554/05-629.1

Michael, T.P. 2014. Plant genome size variation: Bloating and purging DNA. Briefings Funct. Genomics Proteomics 13:308–317. doi:10.1093/bfgp/elu005

Miller, J.R., S. Koren, and G. Sutton. 2010. Assembly algorithms for next-generation sequencing data. Genomics 95:315–327. doi:10.1016/j.ygeno.2010.03.001

Mitchell, A.L., T.K. Attwood, P.C. Babbitt, M. Blum, P. Bork, A. Bridge, S.D. Brown, H.Y. Chang, S. El-Gebali, M.I. Fraser, J. Gough, D.R. Haft, H. Huang, I. Letunic, R. Lopez, A. Luciani, F. Madeira, A. Marchler-Bauer, H. Mi, D.A. Natale, M. Necci, G. Nuka, C. Orengo, A.P. Pandurangan, T. Paysan-Lafosse, S. Pesseat, S.C. Potter, M.A. Qureshi, N.D. Rawlings, N. Redaschi, L.J. Richardson, C. Rivoire, G.A. Salazar, A. Sangrador-Vegas, C.J.A. Sigrist, I. Sillitoe, G.G. Sutton, N. Thanki, P.D. Thomas, S.C.E. Tosatto, S.Y. Yong, and R.D. Finn. 2019. InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 47:D351–D360. doi:10.1093/nar/gky1100

Muñoz-Bertomeu, J., I. Arrillaga, R. Ros, and J. Segura. 2006. Up-regulation of 1-Deoxy-D-xylulose-5-phosphate synthase enhances production of essential oils in transgenic spike lavender. Plant Physiol. 142:890–900. doi:10.1104/pp.106.086355

Munoz-Lopez, M., and J. Garcia-Perez. 2010. DNA Transposons: Nature and Applications in Genomics. Curr. Genomics 11:115–128. doi:10.2174/138920210790886871

Myers, E.W. 1995. Toward Simplifying and Accurately Formulating Fragment Assembly. J. Comput. Biol. 2:275–290. doi:10.1089/cmb.1995.2.275

Myers, E.W. 2005. The fragment assembly string graph. Bioinformatics 21:ii79–ii85. doi:10.1093/bioinformatics/bti1114

Naito, K., E. Cho, G. Yang, M.A. Campbell, K. Yano, Y. Okumoto, T. Tanisaka, and S.R. Wessler. 2006. Dramatic amplification of a rice transposable element during recent domestication. Proc. Natl. Acad. Sci. U. S. A. 103:17620–17625. doi:10.1073/pnas.0605421103

Narzisi, G., and B. Mishra. 2011. Comparing De Novo genome assembly: The long and short of it. PLoS One 6. doi:10.1371/journal.pone.0019175

Negi, P., A.N. Rai, and P. Suprasanna. 2016. Moving through the stressed genome: Emerging regulatory roles for transposons in plant stress response. Front. Plant Sci. 7. doi:10.3389/fpls.2016.01448

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418–426. doi:10.1093/oxfordjournals.molbev.a040410

Nelson, M.G., R.S. Linheiro, and C.M. Bergman. 2017. McClintock: An integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. G3 Genes, Genomes, Genet. 7:2763–2778. doi:10.1534/g3.117.043893

Ono, S. 1972. So much “junk” DNA in our genome.. Brookhaven Symp. Biol. 23:366–370

Ou, S., W. Su, Y. Liao, K. Chougule, J.R.A. Agda, A.J. Hellinga, C.S.B. Lugo, T.A. Elliott, D. Ware, T. Peterson, N. Jiang, C.N. Hirsch, and M.B. Hufford. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 20:275. doi:10.1186/s13059-019-1905-y

Panchy, N., M. Lehti-Shiu, and S.H. Shiu. 2016. Evolution of gene duplication in plants. Plant

Physiol. 171:2294–2316. doi:10.1104/pp.16.00523

Parra, G., K. Bradnam, and I. Korf. 2007. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067. doi:10.1093/bioinformatics/btm071

Pavesp, A., R. Percudani, and F. Conterio. 1997. A novel algorithm for the search of 5S rRNA genes in DNA databases: Comparison with other methods and identification of new potential 5S rRNA genes. Mitochondrial DNA 7:165–177. doi:10.3109/10425179709034032

Van De Peer, Y., S. Maere, and A. Meyer. 2009. The evolutionary significance of ancient genome duplications. Nat. Rev. Genet.. doi:10.1038/nrg2600

Pellicer, J., and I.J. Leitch. 2020. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. New Phytol. 226:301–305. doi:10.1111/nph.16261

Peterson, D.G., S.R. Schulze, E.B. Sciara, S.A. Lee, J.E. Bowers, A. Nagel, N. Jiang, D.C. Tibbitts, S.R. Wessler, and A.H. Paterson. 2002. Integration of cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res. 12:795–807. doi:10.1101/gr.226102

Pflug, J.M., V.R. Holmes, C. Burrus, J. Spencer Johnston, and D.R. Maddison. 2020. Measuring genome sizes using read-depth, k-mers, and flow cytometry: Methodological comparisons in beetles (Coleoptera). G3 Genes, Genomes, Genet. 10:3047–3060. doi:10.1534/g3.120.401028

Prashar, A., I.C. Locke, and C.S. Evans. 2004. Cytotoxicity of lavender oil and its major components to human skin cells. Cell Prolif. 37:221–229. doi:10.1111/j.1365-2184.2004.00307.x

Price, A.L., N.C. Jones, and P.A. Pevzner. 2005. De novo identification of repeat families in large genomes. Bioinformatics 21:i351-8. doi:10.1093/bioinformatics/bti1018

Priya, P., A. Yadav, J. Chand, and G. Yadav. 2018. Terzyme: A tool for identification and analysis of the plant terpenome. Plant Methods 14:4. doi:10.1186/s13007-017-0269-0

Proost, S., P. Pattyn, T. Gerats, and Y. Van De Peer. 2011. Journey through the past: 150 million years of plant genome evolution. Plant J. 66:58–65. doi:10.1111/j.1365-313X.2011.04521.x

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. 2005. InterProScan: Protein domains identifier. Nucleic Acids Res. 33:W116-20. doi:10.1093/nar/gki442

Quinlan, A.R. 2014. BEDTools: The Swiss-Army tool for genome feature analysis. Curr. Protoc. Bioinforma. 2014:11.12.1-11.12.34. doi:10.1002/0471250953.bi1112s47

Quinlan, A.R., and I.M. Hall. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. doi:10.1093/bioinformatics/btq033

Ranallo-Benavidez, T.R., K.S. Jaron, and M.C. Schatz. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. 11:1432. doi:10.1038/s41467-020-14998-3

Rang, F.J., W.P. Kloosterman, and J. de Ridder. 2018. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 19:90. doi:10.1186/s13059-018-1462-9

Rastogi, S., A. Kalra, V. Gupta, F. Khan, R.K. Lal, A.K. Tripathi, S. Parameswaran, C. Gopalakrishnan, G. Ramaswamy, and A.K. Shasany. 2015. Unravelling the genome of Holy basil: An “incomparable” “elixir of life” of traditional Indian medicine. BMC Genomics 16:413. doi:10.1186/s12864-015-1640-z

Rätsch, G., S. Sonnenburg, J. Srinivasan, H. Witte, K.R. Müller, R.J. Sommer, and B. Schölkopf. 2007. Improving the Caenorhabditis elegans genome annotation using machine learning. PLoS Comput. Biol. 3:0313–0322. doi:10.1371/journal.pcbi.0030020

Renny-Byfield, S., and J.F. Wendel. 2014. Doubling down on genomes: Polyploidy and crop plants. Am. J. Bot. 101:1711–1725. doi:10.3732/ajb.1400119

Rhoads, A., and K.F. Au. 2015. PacBio Sequencing and Its Applications. Genomics, Proteomics Bioinforma. 13:278–289. doi:10.1016/j.gpb.2015.08.002

Richter, A., I. Seidl-Adams, T.G. Köllner, C. Schaff, J.H. Tumlinson, and J. Degenhardt. 2015. A small, differentially regulated family of farnesyl diphosphate synthases in maize (Zea mays) provides farnesyl diphosphate for the biosynthesis of herbivore-induced sesquiterpenes. Planta 241:1351–1361. doi:10.1007/s00425-015-2254-z

Roth, C., and D.A. Liberles. 2006. A systematic search for positive selection in higher plants (Embryophytes). BMC Plant Biol. 6. doi:10.1186/1471-2229-6-12

Roulin, A., P.L. Auer, M. Libault, J. Schlueter, A. Farmer, G. May, G. Stacey, R.W. Doerge, and S.A. Jackson. 2013. The fate of duplicated genes in a polyploid plant genome. Plant J. 73:143–153. doi:10.1111/tpj.12026

Salzberg, S.L., A.M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T.J. Treangen, M.C. Schatz, A.L. Delcher, M. Roberts, G. Marcxais, M. Pop, and J.A. Yorke. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 22:557–567. doi:10.1101/gr.131383.111

Sanger, F., J.E. Donelson, A.R. Coulson, H. Kössel, and D. Fischer. 1973. Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA.. Proc. Natl. Acad. Sci. U. S. A. 70:1209–1213. doi:10.1073/pnas.70.4.1209

SanMiguel, P., B.S. Gaut, A. Tikhonov, Y. Nakajima, and J.L. Bennetzen. 1998. The paleontology of intergene retrotransposons of maize. Nat. Genet. 20:43–45. doi:10.1038/1695

Sanmiya, K., T. Iwasaki, M. Matsuoka, M. Miyao, and N. Yamamoto. 1997. Cloning of a cDNA that encodes farnesyl diphosphate synthase and the blue-light-induced expression of the corresponding gene in the leaves of rice plants. Biochim. Biophys. Acta - Gene Struct. Expr. 1350:240–246. doi:10.1016/S0167-4781(96)00231-X

Sarker, L.S., Z.A. Demissie, and S.S. Mahmoud. 2013. Cloning of a sesquiterpene synthase from Lavandula x intermedia glandular trichomes. Planta 238:983–989. doi:10.1007/s00425-013-1937-6

Sarker, L.S., and S.S. Mahmoud. 2015. Cloning and functional characterization of two monoterpene acetyltransferases from glandular trichomes of L. x intermedia.. Planta 242:709–719. doi:10.1007/s00425-015-2325-1

Sato, S., H. Hirakawa, S. Isobe, E. Fukai, A. Watanabe, M. Kato, K. Kawashima, C. Minami, A. Muraki, N. Nakazaki, C. Takahashi, S. Nakayama, Y. Kishida, M. Kohara, M. Yamada, H. Tsuruoka, S. Sasamoto, S. Tabata, T. Aizu, A. Toyoda, T. Shin-I, Y. Minakuchi, Y. Kohara, A. Fujiyama, S. Tsuchimoto, S. Kajiyama, E. Makigano, N. Ohmido, N. Shibagaki, J.A. Cartagena, N. Wada, T. Kohinata, A. Atefeh, S. Yuasa, S. Matsunaga, and K. Fukui. 2011. Sequence analysis of the genome of an oil-bearing tree, jatropha curcas L.. DNA Res. 18:65–76. doi:10.1093/dnares/dsq030

Sato, S., S. Tabata, H. Hirakawa, E. Asamizu, K. Shirasawa, S. Isobe, T. Kaneko, Y. Nakamura, D. Shibata, K. Aoki, M. Egholm, J. Knight, R. Bogden, C. Li, Y. Shuang, X. Xu, S. Pan, S. Cheng, X. Liu, Y. Ren, J. Wang, A. Albiero, F. Dal Pero, S. Todesco, J. Van Eck, R.M.

Buels, A. Bombarely, J.R. Gosselin, M. Huang, J.A. Leto, N. Menda, S. Strickler, L. Mao, S. Gao, I.Y. Tecle, T. York, Y. Zheng, J.T. Vrebalov, J. Lee, S. Zhong, L.A. Mueller, W.J. Stiekema, P. Ribeca, T. Alioto, W. Yang, S. Huang, Y. Du, Z. Zhang, J. Gao, Y. Guo, X. Wang, Y. Li, J. He, C. Li, Z. Cheng, J. Zuo, J. Ren, J. Zhao, L. Yan, H. Jiang, B. Wang, H. Li, Z. Li, F. Fu, B. Chen, B. Han, Q. Feng, D. Fan, Y. Wang, H. Ling, Y. Xue, D. Ware, W. Richard McCombie, Z.B. Lippman, J.M. Chia, K. Jiang, S. Pasternak, L. Gelley, M. Kramer, L.K. Anderson, S. Bin Chang, S.M. Royer, L.A. Shearer, S.M. Stack, J.K.C. Rose, Y. Xu, N. Eannetta, A.J. Matas, R. McQuinn, S.D. Tanksley, F. Camara, R. Guigó, S. Rombauts, J. Fawcett, Y. Van De Peer, D. Zamir, C. Liang, M. Spannagl, H. Gundlach, R. Bruggmann, K. Mayer, Z. Jia, J. Zhang, Z. Ye, G.J. Bishop, S. Butcher, R. Lopez-Cobollo, D. Buchan, I. Filippis, J. Abbott, I.R. Dixit, M. Singh, A. Singh, J.K. Pal, A. Pandit, P.K. Singh, A.K. Mahato, V. Dogra, K. Gaikwad, T.R. Sharma, T. Mohapatra, N.K. Singh, M. Causse, C. Rothan, C. Noirot, A. Bellec, C. Klopp, C. Delalande, H. Berges, J. Mariette, P. Frasse, S. Vautrin, T.M. Zouine, A. Latché, C. Rousseau, F. Regad, J.C. Pech, M. Philippot, M. Bouzayen, P. Pericard, S. Osorio, A.F. Del Carmen, A. Monforte, A. Granell, R. Fernandez-Muñoz, M. Conte, G. Lichtenstein, F. Carrari, G. De Bellis, F. Fuligni, C. Peano, S. Grandillo, P. Termolino, M. Pietrella, E. Fantini, G. Falcone, A. Fiore, G. Giuliano, L. Lopez, P. Facella, G. Perrotta, L. Daddiego, G. Bryan, B.M. Orozco, X. Pastor, D. Torrents, M.G.M. Van Schriek, R.M.C. Feron, J. Van Oeveren, P. De Heer, L. Da Ponte, S. Jacobs-Oomen, M. Cariaso, M. Prins, M.J.T. Van Eijk, A. Janssen, J.J. Van Haaren, S. -HwanJo, J. Kim, S.Y. Kwon, S. Kim, D.H. Koo, S. Lee, C. Clouser, A. Rico, A. Hallab, C. Gebhardt, K. Klee, A. Jöcker, J. Warfsmann, U. Göbel, S. Kawamura, K. Yano, J.D. Sherman, H. Fukuoka, S. Negoro, S. Bhutty, P. Chowdhury, D. Chattopadhyay, E. Datema, S. Smit,

E.G.W.M. Schijlen, J. Van De Belt, J.C. Van Haarst, S.A. Peters, M.J. Van Staveren, M.H.C. Henkens, P.J.W. Mooyman, T. Hesselink, R.C.H.J. Van Ham, G. Jiang, M. Droege, D. Choi, B.C. Kang, B.D. Kim, M. Park, S. Kim, S.I. Yeom, Y.H. Lee, Y. Do Choi, G. Li, J. Gao, Y. Liu, S. Huang, V. Fernandez-Pedrosa, C. Collado, S. Zuñ Iga, G. Wang, R. Cade, R.A. Dietrich, J. Rogers, S. Knapp, Z. Fei, R.A. White, T.W. Thannhauser, J.J. Giovannoni, M.A. Botella, L. Gilbert, F.R. Gonzalez, J.L. Goicoechea, Y. Yu, D. Kudrna, K. Collura, M. Wissotski, R. Wing, B.C. Meyers, A.B. Gurazada, P.J. Green, S. Mathur, S. Vyas, A.U. Solanke, R. Kumar, V. Gupta, A.K. Sharma, P. Khurana, J.P. Khurana, A.K. Tyagi, T. Dalmay, I. Mohorianu, B. Walts, S. Chamala, W.B. Barbazuk, J. Li, H. Guo, T.H. Lee, Y. Wang, D. Zhang, A.H. Paterson, X. Wang, H. Tang, A. Barone, M.L. Chiusano, M.R. Ercolano, N. D'Agostino, M. Di Filippo, A. Traini, W. Sanseverino, L. Frusciante, G.B. Seymour, M. Elharam, Y. Fu, A. Hua, S. Kenton, J. Lewis, S. Lin, F. Najar, H. Lai, B. Qin, R. Shi, C. Qu, D. White, J. White, Y. Xing, K. Yang, J. Yi, Z. Yao, L. Zhou, B.A. Roe, A. Vezzi, M. D'Angelo, R. Zimbello, R. Schiavon, E. Caniato, C. Rigobello, D. Campagna, N. Vitulo, G. Valle, D.R. Nelson, E. De Paoli, D. Szinay, H.H. De Jong, Y. Bai, R.G.F. Visser, R.K. Lankhorst, H. Beasley, K. McLaren, C. Nicholson, C. Riddle, and G. Gianese. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641. doi:10.1038/nature11119

Schirmer, M., R. D'Amore, U.Z. Ijaz, N. Hall, and C. Quince. 2016. Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics 17:125. doi:10.1186/s12859-016-0976-y

Schmidt, M.H.W., A. Vogel, A.K. Denton, B. Istace, A. Wormit, H. van de Geest, M.E. Bolger, S. Alseekh, J. Maß, C. Pfaff, U. Schurr, R. Chetelat, F. Maumus, J.M. Aury, S. Koren, A.R.

Fernie, D. Zamir, A.M. Bolger, and B. Usadel. 2017. De novo assembly of a new Solanum pennellii accession using nanopore sequencing. Plant Cell 29:2336–2348. doi:10.1105/tpc.17.00521

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D.L. Hyten, Q. Song, J.J. Thelen, J. Cheng, D. Xu, U. Hellsten, G.D. May, Y. Yu, T. Sakurai, T. Umezawa, M.K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X.C. Zhang, K. Shinozaki, H.T. Nguyen, R.A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R.C. Shoemaker, and S.A. Jackson. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463:178–183. doi:10.1038/nature08670

Schnable, P.S., D. Ware, R.S. Fulton, J.C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T.A. Graves, P. Minx, A.D. Reily, L. Courtney, S.S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S.M. Rock, E. Belter, F. Du, K. Kim, R.M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S.M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S.

Kumari, B. Faga, M.J. Levy, L. McMahan, P. Van Buren, M.W. Vaughn, K. Ying, C.T. Yeh, S.J. Emrich, Y. Jia, A. Kalyanaraman, A.P. Hsia, W.B. Barbazuk, R.S. Baucom, T.P. Brutnell, N.C. Carpita, C. Chaparro, J.M. Chia, J.M. Deragon, J.C. Estill, Y. Fu, J.A. Jeddeloh, Y. Han, H. Lee, P. Li, D.R. Lisch, S. Liu, Z. Liu, D.H. Nagel, M.C. McCann, P. Sanmiguel, A.M. Myers, D. Nettleton, J. Nguyen, B.W. Penning, L. Ponnala, K.L. Schneider, D.C. Schwartz, A. Sharma, C. Soderlund, N.M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T.K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J.L. Bennetzen, R.K. Dawe, J. Jiang, N. Jiang, G.G. Presting, S.R. Wessler, S. Aluru, R.A. Martienssen, S.W. Clifton, W.R. McCombie, R.A. Wing, and R.K. Wilson. 2009. The B73 maize genome: Complexity, diversity, and dynamics. Science (80-.). 326:1112–1115. doi:10.1126/science.1178534

Schweikert, G., J. Behr, A. Zien, G. Zeller, C.S. Ong, S. Sonnenburg, and G. Rätsch. 2009. mGene.web: A web service for accurate computational gene finding. Nucleic Acids Res. 37:W312–W316. doi:10.1093/nar/gkp479

Shulaev, V., D.J. Sargent, R.N. Crowhurst, T.C. Mockler, O. Folkerts, A.L. Delcher, P. Jaiswal, K. Mockaitis, A. Liston, S.P. Mane, P. Burns, T.M. Davis, J.P. Slovin, N. Bassil, R.P. Hellens, C. Evans, T. Harkins, C. Kodira, B. Desany, O.R. Crasta, R. V. Jensen, A.C. Allan, T.P. Michael, J.C. Setubal, J.M. Celton, D.J.G. Rees, K.P. Williams, S.H. Holt, J.J.R. Rojas, M. Chatterjee, B. Liu, H. Silva, L. Meisel, A. Adato, S.A. Filichkin, M. Troggio, R. Viola, T.L. Ashman, H. Wang, P. Dharmawardhana, J. Elser, R. Raja, H.D. Priest, D.W. Bryant, S.E. Fox, S.A. Givan, L.J. Wilhelm, S. Naithani, A. Christoffels, D.Y. Salama, J. Carter, E.L. Girona, A. Zdepski, W. Wang, R.A. Kerstetter, W. Schwab, S.S. Korban, J. Davik, A. Monfort, B. Denoyes-Rothan, P. Arus, R. Mittler, B. Flinn, A. Aharoni, J.L. Bennetzen,

S.L. Salzberg, A.W. Dickerman, R. Velasco, M. Borodovsky, R.E. Veilleux, and K.M. Folta. 2011. The genome of woodland strawberry (Fragaria vesca). Nat. Genet. 43:109–116. doi:10.1038/ng.740

Simão, F.A., R.M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E.M. Zdobnov. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. doi:10.1093/bioinformatics/btv351

Simillion, C., K. Vandepoele, M.C.E. Van Montagu, M. Zabeau, and Y. Van de Peer. 2002. The hidden duplication past of Arabidopsis thaliana. Proc. Natl. Acad. Sci. U. S. A. 99:13627–13632. doi:10.1073/pnas.212522399

Simpson, J.T., and R. Durbin. 2010. Efficient construction of an assembly string graph using the FM-index. Bioinformatics 26:i367--73. doi:10.1093/bioinformatics/btq217

Simpson, J.T., and R. Durbin. 2012. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 22:549–556. doi:10.1101/gr.126953.111

Simpson, J.T., K. Wong, S.D. Jackman, J.E. Schein, S.J.M. Jones, and I. Birol. 2009. ABySS: A parallel assembler for short read sequence data. Genome Res. 19:1117–1123. doi:10.1101/gr.089532.108

Simpson, K., L.F. Quiroz, M. Rodriguez-Concepción, and C.R. Stange. 2016. Differential contribution of the first two enzymes of the MEP pathway to the supply of metabolic precursors for carotenoid and chlorophyll biosynthesis in carrot (Daucus carota). Front. Plant Sci. 7:1344. doi:10.3389/fpls.2016.01344

Singh, R., R. Ming, and Q. Yu. 2016. Comparative Analysis of GC Content Variations in Plant Genomes. Trop. Plant Biol. 9:136–149. doi:10.1007/s12042-016-9165-4

Šmarda, P., P. Bureš, L. Horová, I.J. Leitch, L. Mucina, E. Pacini, L. Tichý, V. Grulich, and O.

Rotreklová. 2014. Ecological and evolutionary significance of genomic GC content diversity in monocots. Proc. Natl. Acad. Sci. U. S. A. 111:E4096–E4102. doi:10.1073/pnas.1321152111

Smit, A., and R. Hubley. 2018. RepeatModeler Open-1.0.. GItHub

Smit, A.F.A., R. Hubley, and P. Green. 2013. RepeatMasker Open-4.0. 2013-2015 .. http://www.repeatmasker.org

Soltis, D.E., V.A. Albert, J. Leebens-Mack, C.D. Bell, A.H. Paterson, C. Zheng, D. Sankoff, C.W. DePamphilis, P.K. Wall, and P.S. Soltis. 2009. Polyploidy and angiosperm diversification. Am. J. Bot. 96:336–348. doi:10.3732/ajb.0800079

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern. 2006. AUGUSTUS: A b initio prediction of alternative transcripts. Nucleic Acids Res. 34:W435--W439. doi:10.1093/nar/gkl200

Stanke, M., and S. Waack. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19:ii215–ii225. doi:10.1093/bioinformatics/btg1080

Sun, H., J. DIng, M. Piednoël, and K. Schneeberger. 2018. FindGSE: Estimating genome size variation within human and Arabidopsis using k -mer frequencies. Bioinformatics 34:550–557. doi:10.1093/bioinformatics/btx637

Swift, H. 1950. The constancy of desoxyribose nucleic acid in plant nuclei.. Proc. Natl. Acad. Sci. U. S. A. 36:643–654. doi:10.1073/pnas.36.11.643

Tholl, D., and S. Lee. 2011. Terpene Specialized Metabolism in Arabidopsis thaliana . Arab. B. 9:e0143. doi:10.1199/tab.0143

Tiley, G.P., M.S. Barker, and J. Gordon Burleigh. 2018. Assessing the performance of KS plots for detecting ancient whole genome duplications. Genome Biol. Evol. 10:2882–2898.

doi:10.1093/gbe/evy200

Turner, F.S. 2014. Assessment of insert sizes and adapter content in fastq data from NexteraXT libraries. Front. Genet. 5. doi:10.3389/fgene.2014.00005

Upson, T., and S. Andrews. 2005. The Genus Lavandula. Book News, Inc.

Urwin, N.A.R. 2014. Generation and characterisation of colchicine-induced polyploid Lavandula × intermedia. Euphytica 197:331–339. doi:10.1007/s10681-014-1069-5

Urwin, N.A.R., J. Horsnell, and T. Moon. 2007. Generation and characterisation of colchicine-induced autotetraploid Lavandula angustifolia. Euphytica 156:257–266. doi:10.1007/s10681-007-9373-y

Veleba, A., P. Šmarda, F. Zedek, L. Horová, J. Šmerda, and P. Bureš. 2017. Evolution of genome size and genomic GC content in carnivorous holokinetics (Droseraceae). Ann. Bot. 119:409–416. doi:10.1093/aob/mcw229

Vicient, C.M., and J.M. Casacuberta. 2017. Impact of transposable elements on polyploid plant genomes. Ann. Bot. 120:195–207. doi:10.1093/aob/mcx078

Vining, K.J., S.R. Johnson, A. Ahkami, I. Lange, A.N. Parrish, S.C. Trapp, R.B. Croteau, S.C.K. Straub, I. Pandelova, and B.M. Lange. 2017. Draft Genome Sequence of Mentha longifolia and Development of Resources for Mint Cultivar Improvement. Mol. Plant 10:323–339. doi:10.1016/j.molp.2016.10.018

Wang, C., Q. Chen, D. Fan, J. Li, G. Wang, and P. Zhang. 2016. Structural Analyses of Short-Chain Prenyltransferases Identify an Evolutionarily Conserved GFPPS Clade in Brassicaceae Plants. Mol. Plant 9:195–204. doi:10.1016/j.molp.2015.10.010

Wang, J., G. Zhang, X. Fang, X. Guo, L. Li, R. Luo, F. Xu, P. Yang, L. Zhang, X. Wang, H. Qi, Z. Xiong, H. Que, Y. Xie, P.W.H. Holland, J. Paps, Y. Zhu, F. Wu, Y. Chen, J. Wang, C.

Peng, J. Meng, L. Yang, J. Liu, B. Wen, N. Zhang, Z. Huang, Q. Zhu, Y. Feng, A. Mount, D. Hedgecock, Z. Xu, Y. Liu, T. Domazet-Lošo, Y. Du, X. Sun, S. Zhang, B. Liu, P. Cheng, X. Jiang, J. Li, D. Fan, W. Wang, W. Fu, T. Wang, B. Wang, J. Zhang, Z. Peng, Y. Li, N. Li, J. Wang, M. Chen, Y. He, F. Tan, X. Song, Q. Zheng, R. Huang, H. Yang, X. Du, L. Chen, M. Yang, P.M. Gaffney, S. Wang, L. Luo, Z. She, Y. Ming, W. Huang, S. Zhang, B. Huang, Y. Zhang, T. Qu, P. Ni, G. Miao, J. Wang, Q. Wang, C.E.W. Steinberg, H. Wang, N. Li, L. Qian, G. Zhang, Y. Li, H. Yang, X. Liu, Y. Yin, and J. Wang. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490:49–54. doi:10.1038/nature11413

Wang, L., S. Yu, C. Tong, Y. Zhao, Y. Liu, C. Song, Y. Zhang, X. Zhang, Y. Wang, W. Hua, D. Li, D. Li, F. Li, J. Yu, C. Xu, X. Han, S. Huang, S. Tai, J. Wang, X. Xu, Y. Li, S. Liu, R.K. Varshney, J. Wang, and X. Zhang. 2014. Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. Genome Biol. 15:R39. doi:10.1186/gb-2014-15-2-r39

Wang, Y., D. Coleman-Derr, G. Chen, and Y.Q. Gu. 2015. OrthoVenn: A web server for genome wide comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res. 43:W78–W84. doi:10.1093/nar/gkv487

Wang, Z., Y. Chen, and Y. Li. 2004. A brief review of computational gene prediction methods.. Genomics. Proteomics Bioinformatics 2:216–221. doi:10.1016/S1672-0229(04)02028-5

Warren, R.L., C. Yang, B.P. Vandervalk, B. Behsaz, A. Lagman, S.J.M. Jones, and I. Birol. 2015. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. Gigascience 4. doi:10.1186/s13742-015-0076-3

Wells, R., F. Truong, A.M. Adal, L.S. Sarker, and S.S. Mahmoud. 2018. Lavandula essential oils: A current review of applications in medicinal, food, and cosmetic industries of

lavender. Nat. Prod. Commun. 13:1403–1417. doi:10.1177/1934578x1801301038

Wells, R.S., A.M. Adal, L. Bauer, E. Najafianashrafi, and S.S. Mahmoud. 2020. Cloning and functional characterization of a floral repressor gene from Lavandula angustifolia. Planta 251. doi:10.1007/s00425-019-03333-w

Wicker, T., F. Sabot, A. Hua-Van, J.L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A.H. Schulman. 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8:973–982. doi:10.1038/nrg2165

Wilhelm, J., A. Pingoud, and M. Hahn. 2003. Real-time PCR-based method for the estimation of genome sizes.. Nucleic Acids Res. 31:56e – 56. doi:10.1093/nar/gng056

Woronuk, G., Z. Demissie, M. Rheault, and S. Mahmoud. 2011. Biosynthesis and therapeutic properties of lavandula essential oil constituents. Planta Med. 77:7–15. doi:10.1055/s-0030-1250136

Xie, C., X. Mao, J. Huang, Y. Ding, J. Wu, S. Dong, L. Kong, G. Gao, C.Y. Li, and L. Wei. 2011. KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 39:W316–W322. doi:10.1093/nar/gkr483

Xu, L., Z. Dong, L. Fang, Y. Luo, Z. Wei, H. Guo, G. Zhang, Y.Q. Gu, D. Coleman-Derr, Q. Xia, and Y. Wang. 2019. OrthoVenn2: A web server for whole-genome comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res. 47:W52–W58. doi:10.1093/nar/gkz333

Xu, X., S. Pan, S. Cheng, B. Zhang, D. Mu, P. Ni, G. Zhang, S. Yang, R. Li, J. Wang, G. Orjeda, F. Guzman, M. Torres, R. Lozano, O. Ponce, D. Martinez, G. De La Cruz, S.K. Chakrabarti, V.U. Patil, G. Skryabin, B.B. Kuznetsov, N. V. Ravin, T. V. Kolganova, A. V. Beletsky, A.

V. Mardanov, A. Di Genova, D.M. Bolser, D.M.A. Martin, G. Li, Y. Yang, H. Kuang, Q. Hu, X. Xiong, G.J. Bishop, B. Sagredo, N. Mejía, W. Zagorski, R. Gromadka, J. Gawor, P. Szczesny, S. Huang, Z. Zhang, C. Liang, J. He, Y. Li, Y. He, J. Xu, Y. Zhang, B. Xie, Y. Du, D. Qu, M. Bonierbale, M. Ghislain, M.D.R. Herrera, G. Giuliano, M. Pietrella, G. Perrotta, P. Facella, K. O'Brien, S.E. Feingold, L.E. Barreiro, G.A. Massa, L. Diambra, B.R. Whitty, B. Vaillancourt, H. Lin, A.N. Massa, M. Geoffroy, S. Lundback, D. DellaPenna, C.R. Buell, S.K. Sharma, D.F. Marshall, R. Waugh, G.J. Bryan, M. Destefanis, I. Nagy, D. Milbourne, S.J. Thomson, M. Fiers, J.M.E. Jacobs, K.L. Nielsen, M. Sønderkær, M. Iovene, G.A. Torres, J. Jiang, R.E. Veilleux, C.W.B. Bachem, J. De Boer, T. Borm, B. Kloosterman, H. Van Eck, E. Datema, B.L. Hekkert, A. Goverse, R.C.H.J. Van Ham, and R.G.F. Visser. 2011. Genome sequence and analysis of the tuber crop potato. Nature 475:189–195. doi:10.1038/nature10158

Xue, W., J.T. Li, Y.P. Zhu, G.Y. Hou, X.F. Kong, Y.Y. Kuang, and X.W. Sun. 2013. L_RNA_scaffolder: Scaffolding genomes with transcripts. BMC Genomics 14:604. doi:10.1186/1471-2164-14-604

Yandell, M., and D. Ence. 2012. A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. 13:329–342. doi:10.1038/nrg3174

Yang, J., D. Liu, X. Wang, C. Ji, F. Cheng, B. Liu, Z. Hu, S. Chen, D. Pental, Y. Ju, P. Yao, X. Li, K. Xie, J. Zhang, J. Wang, F. Liu, W. Ma, J. Shopan, H. Zheng, S.A. Mackenzie, and M. Zhang. 2016. The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection. Nat. Genet. 48:1225–1232. doi:10.1038/ng.3657

Ye, C., Z.S. Ma, C.H. Cannon, M. Pop, and D.W. Yu. 2012. Exploiting sparseness in de novo

genome assembly.. BMC Bioinformatics 13 Suppl 6:S1. doi:10.1186/1471-2105-13-S6-S1

Ye, J., Y. Zhang, H. Cui, J. Liu, Y. Wu, Y. Cheng, H. Xu, X. Huang, S. Li, A. Zhou, X. Zhang, L. Bolund, Q. Chen, J. Wang, H. Yang, L. Fang, and C. Shi. 2018. WEGO 2.0: A web tool for analyzing and plotting GO annotations, 2018 update. Nucleic Acids Res. 46:W71–W75. doi:10.1093/nar/gky400

Yi, C.G., T.T. Hieu, S.H. Lee, B.R. Choi, M. Kwon, and Y.J. Ahn. 2016. Toxicity of Lavandula angustifolia oil constituents and spray formulations to insecticide-susceptible and pyrethroid-resistant Plutella xylostella and its endoparasitoid Cotesia glomerata. Pest Manag. Sci. 72:1202–1210. doi:10.1002/ps.4098

Yu, J., S. Hu, J. Wang, G.K.S. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, J. Li, Z. Liu, Q. Qi, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, W. Zhao, P. Li, W. Chen, Y. Zhang, J. Hu, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, M. Tao, L. Zhu, L. Yuan, and H. Yang. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science (80-.). 296:79–92. doi:10.1126/science.1068037

Yu, Z., C. Zhao, G. Zhang, J.A. Teixeira da Silva, and J. Duan. 2020. Genome-wide identification and expression profile of tps gene family in dendrobium officinale and the role of dotps10 in linalool biosynthesis. Int. J. Mol. Sci. 21:1–22. doi:10.3390/ijms21155419

Zapata, L., J. Ding, E.M. Willing, B. Hartwig, D. Bezdan, W.B. Jiao, V. Patel, G.V. James, M. Koornneef, S. Ossowski, and K. Schneeberger. 2016. Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. Proc. Natl. Acad. Sci. U. S. A. 113:E4052–E4060. doi:10.1073/pnas.1607532113

Zerbe, P., and J. Bohlmann. 2015. Plant diterpene synthases: Exploring modularity and metabolic diversity for bioengineering. Trends Biotechnol. 33:419–428. doi:10.1016/j.tibtech.2015.04.006

Zerbino, D.R., and E. Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829. doi:10.1101/gr.074492.107

Zhang, F., W. Liu, J. Xia, J. Zeng, L. Xiang, S. Zhu, Q. Zheng, H. Xie, C. Yang, M. Chen, and Z. Liao. 2018. Molecular characterization of the 1-deoxy-D-xylulose 5-phosphate synthase gene family in Artemisia annua. Front. Plant Sci. 9:952. doi:10.3389/fpls.2018.00952

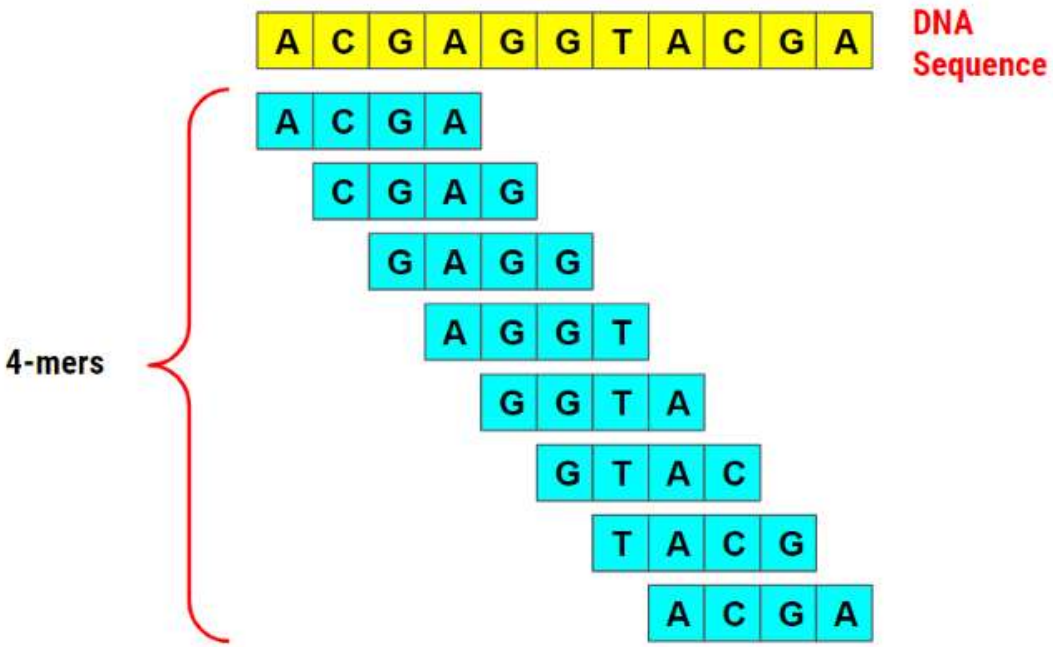
Zhang, Z., J. Li, X.Q. Zhao, J. Wang, G.K.S. Wong, and J. Yu. 2006. KaKs_Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging. Genomics, Proteomics Bioinforma. 4:259–263. doi:10.1016/S1672-0229(07)60007-2

Zhou, F., and E. Pichersky. 2020. The complete functional characterisation of the terpene synthase family in tomato. New Phytol. 226:1341–1360. doi:10.1111/nph.16431

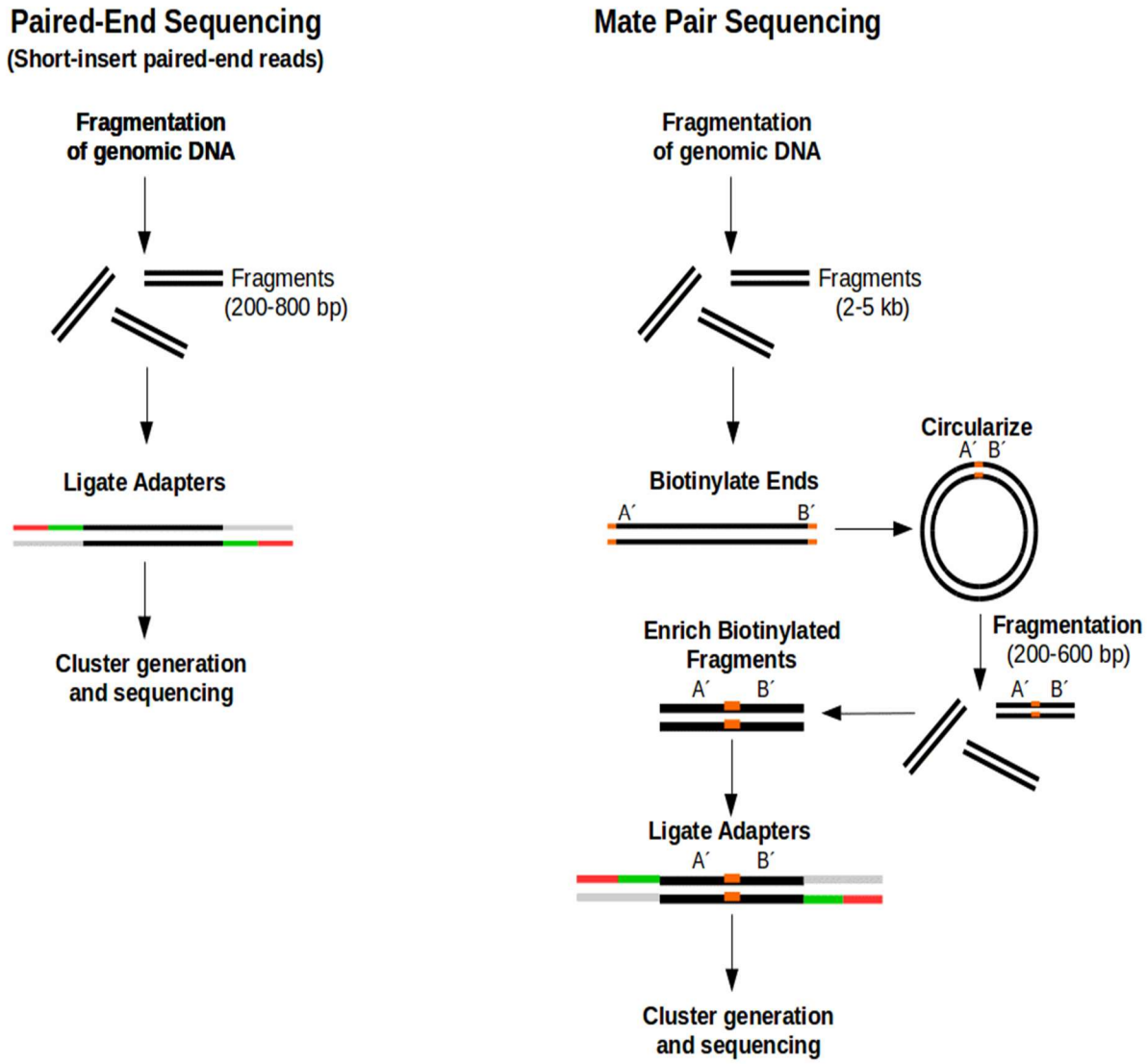
Zonneveld, B.J.M., I.J. Leitch, and M.D. Bennett. 2005. First nuclear DNA amounts in more than 300 angiosperms. Ann. Bot. 96:229–244. doi:10.1093/aob/mci170

VI. Appendixes

Appendix A – Additional tables and figures



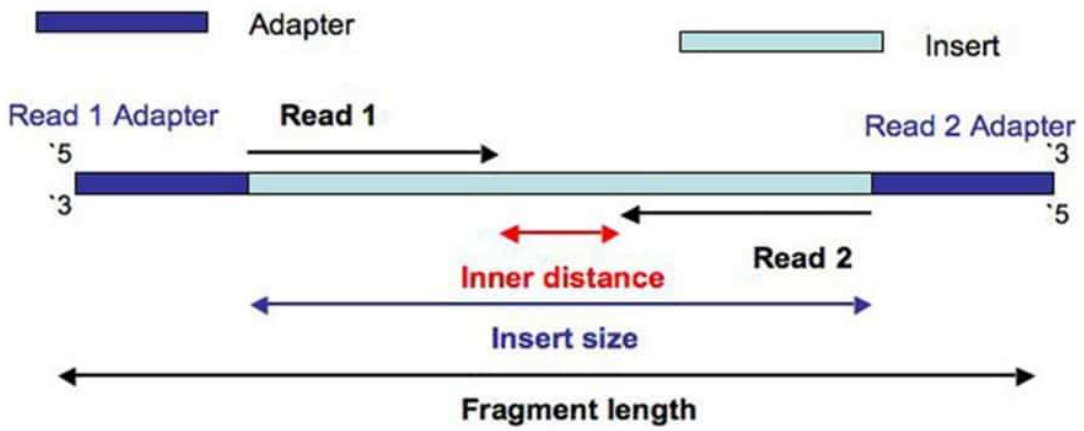
A.F.1. Figure showing k-mers (4-mers) generated from a DNA sequence.



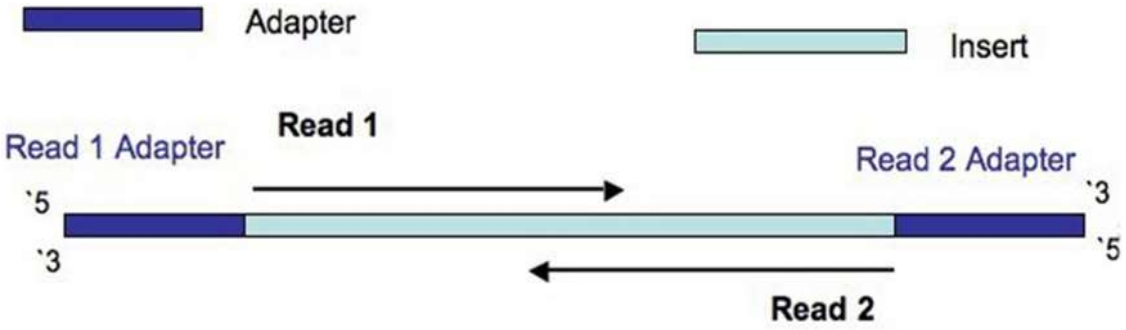
A.F.2. Figure showing two major types of sequencing reads, the paired-end read and the mate-pair reads.

This figure was adapted from https://www.ecseq.com. The left panel shows the generation of paired-end reads and the right panel showing the generation of mate-pair reads.

A.

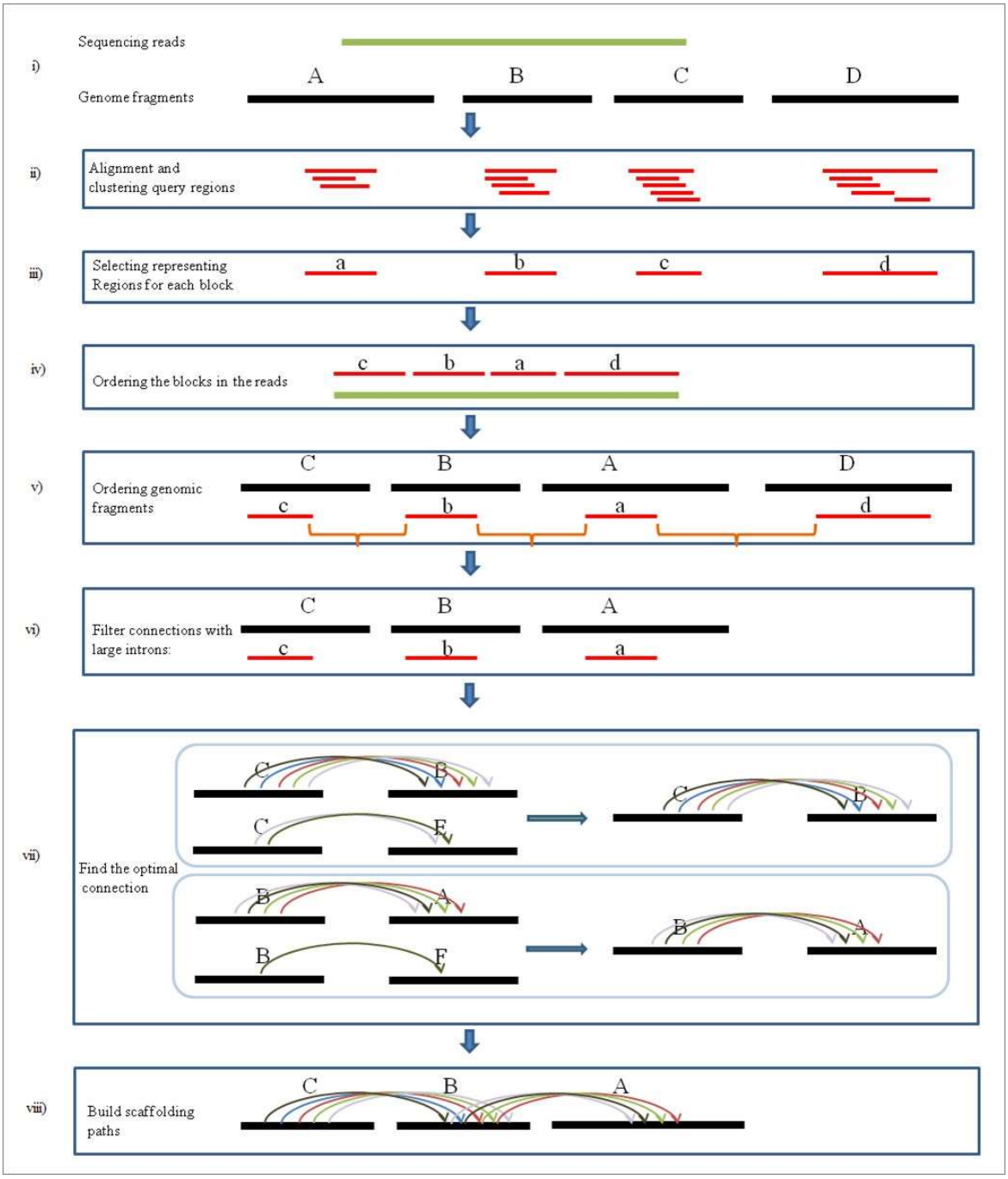


B.



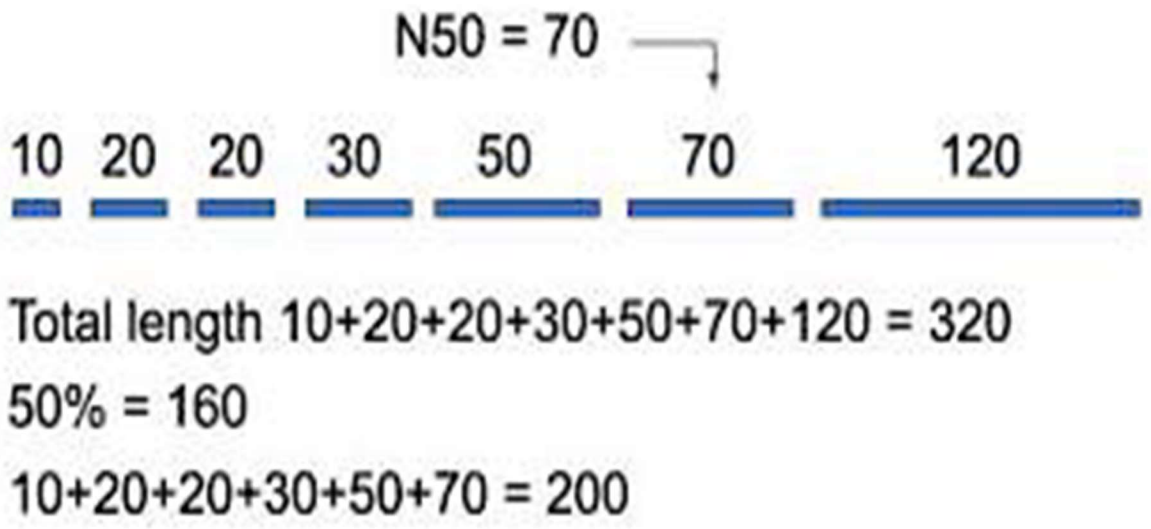
A.F.3. Schematic view of an Illumina paired-end read.

This figure is adapted from "Assessment of insert sizes and adapter content in fastq data from NexteraXT libraries" (Turner, 2014) A. Paired-end read showing the adapter regions, orientation of the forward and reverse reads. The insert size is longer than the length of both reads. B. Overlap read where the insert size is smaller than that of the length of both the reads.



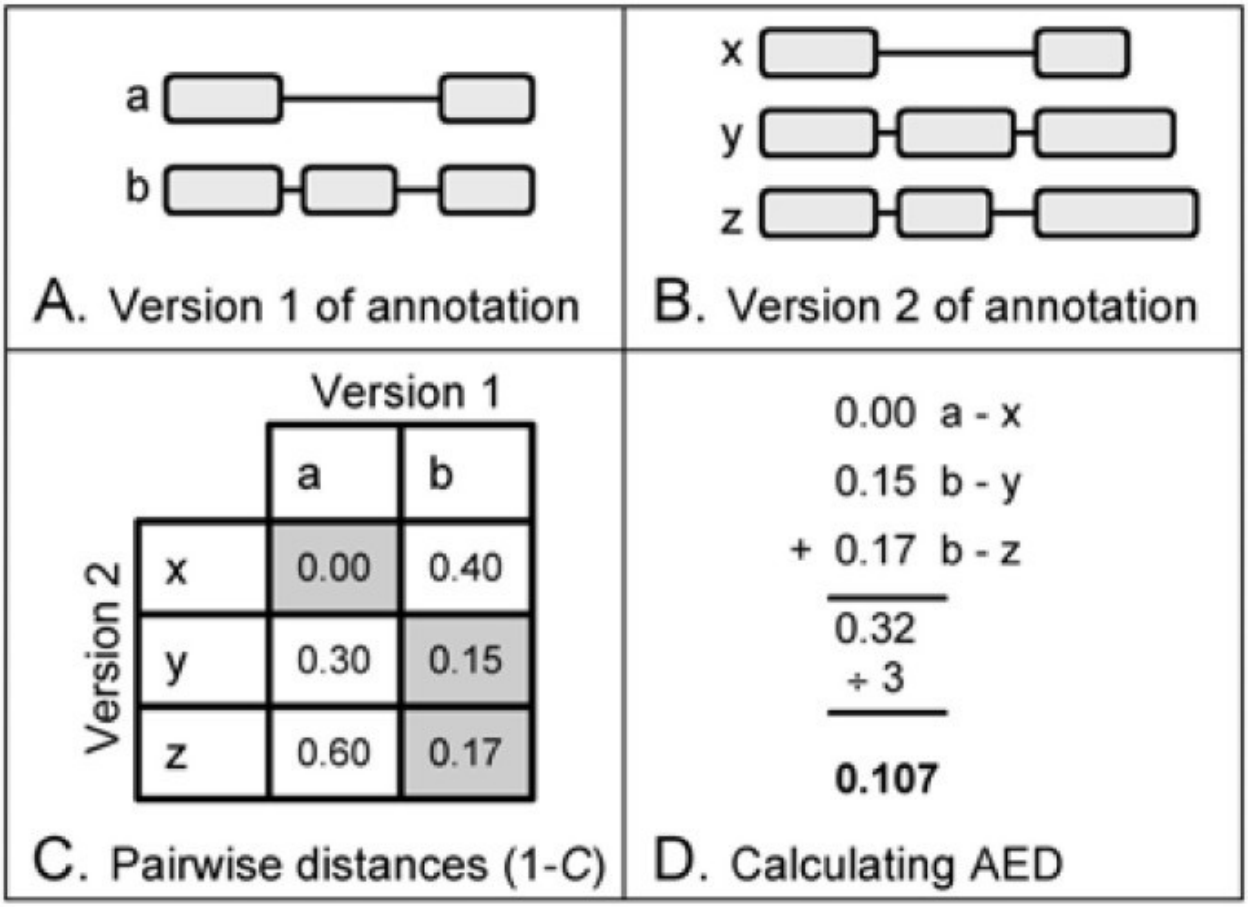
A.F.4. Schematic representation of the L_RNA_scaffolder steps in improving genome assembly.

This figure is adapted from "L_RNA_scaffolder: scaffolding genomes with transcripts" by Xue et al. (Xue et al., 2013). Steps show the selection of the 'guide' transcripts, ordering the fragments, estimating potential introns, building scaffolding paths, finding optimal connection between the scaffolds and joining the scaffolds. The full description is available in the corresponding article.



A.F.5. Schematic representation of N50 metric calculation.

The figure is adapted from https://bioinformaticshome.com. The assembled contigs are listed in ascending order starting from the shortest, the total length and the 50% of the total are then calculated. The N50 value is the length of the last contig when adding up the lengths of the contigs starting from smallest until the total length reaches 50% or more of the total length.



A.F.6. Calculating AED score between two versions of the same annotation.

This figure is adapted from "Quantitative measures for the management and comparison of annotated genomes" by Eilbeck K. et al (Eilbeck et al., 2009). The panels A and B represent two versions of the same annotation. Panel C indicates the pair wise distances between these 2 annotations and highlighting the minimum distances. The AED is calculated by summing the minimum distance (0.32) or represented by normalizing the number of transcript-pairs to give an average per-transcript pair (0.107). A detailed description for calculating the pair-wise distances 1-C is described in the publication by the author (Eilbeck et al., 2009).

A.1 Genome size estimation using qPCR

Genome	Gene	C-values in pg		Genome size (Mbp)
Lavender	*DXR*	E1	1.056	876
		E2	0.928	
		E3	0.926	
		Average	**0.970**	
	HMGS	E1	1.119	870
		E2	0.862	
		E3	0.879	
		Average	**0.953**	
Arabidopsis	*DXR*	E1	0.171	157
		E2	0.174	
		E3	0.172	
		Average	**0.172**	156
	HMGS	E1	0.167	
		E2	0.167	
		E3	0.167	
		Average	**0.167**	

A.2 Number of IPR codes assigned to lavender protein sequences.

IPR assigned	count_freq	Freq
1	19949	50.7763185
2	10567	26.8962533
3	7878	20.0519243
4	5020	12.7774384
5	2940	7.48320098
6	1387	3.53034005
7	749	1.90643453
8	392	0.99776013
9	229	0.58287518
10	103	0.26216656
11	41	0.10435756
12	16	0.0407249
13	11	0.02799837
14	5	0.01272653
15	6	0.01527184
16	3	0.00763592
17	1	0.00254531

A.3 Lavender protein sequence matches against model plant proteins using KOBAS*

Lavender protein sequences	Model plants	Protein sequence matches	% Match
60,819	*Arabidopsis*	50,516	83%
	Rice	58,241	95%
	Maize	57,689	94%

*- KOBAS (Xie et al., 2011) - KEGG Orthology Based Annotation System, is a webserver for functional annotation of gene or protein sequences.

Appendix B – A list of commands/programs used and their parameter settings

B.1. Computational genome size estimation using K-mergenie

```
./kmergenie list_files.txt &>run_kmergenie_4all.log &
list_files.txt
/work/lianglab/wgs/lavender/FASTQ/C48NNACXX/DNA_SG_ACAGTG_L005.R1.fastq
/work/lianglab/wgs/lavender/FASTQ/C48NNACXX/DNA_SG_ACAGTG_L005.R2.fastq
/work/lianglab/wgs/lavender/FASTQ/C48NNACXX/DNA_SG_GTGAAA_L005.R1.fastq
/work/lianglab/wgs/lavender/FASTQ/C48NNACXX/DNA_SG_GTGAAA_L005.R2.fastq
/work/lianglab/wgs/lavender/FASTQ/C48LWACXX/DNA_SG_ACAGTG_L007.R1.fastq
/work/lianglab/wgs/lavender/FASTQ/C48LWACXX/DNA_SG_ACAGTG_L007.R2.fastq
/work/lianglab/wgs/lavender/FASTQ/C48LWACXX/DNA_SG_GTGAAA_L007.R1.fastq
/work/lianglab/wgs/lavender/FASTQ/C48LWACXX/DNA_SG_GTGAAA_L007.R2.fastq
/work/lianglab/wgs/lavender/FASTQ/OverlapLib/PE_l50T230.1.fastq
/work/lianglab/wgs/lavender/FASTQ/OverlapLib/PE_l50T230.2.fastq
```

The “list_files.txt” contains all the sequencing reads which are used as input by Kmergenie tool to estimate the genome size of the lavender plant.

B.2. Sequencing data pre-processing

```
/work/lianglab/bin/fastqc  -q -t 4 -o /work/rn13ow/fastqc/ ../PE_L5A.1.fastq &>run_fastqc_L5A.log &
```

B.3. Contig generation

```
/home/rn13ow/work/fermi-1.1/run_fermi.pl -Pt 6 -e /home/rn13ow/work/fermi-1.1/fermi -p
lavender_raw_PE /work/lianglab/wgs/lavender/FASTQ/C48NNACXX/DNA_SG_ACAGTG_L005.R1.fastq
/work/lianglab/wgs/lavender/FASTQ/C48NNACXX/DNA_SG_ACAGTG_L005.R2.fastq
/work/lianglab/wgs/lavender/FASTQ/C48NNACXX/DNA_SG_GTGAAA_L005.R1.fastq
/work/lianglab/wgs/lavender/FASTQ/C48NNACXX/DNA_SG_GTGAAA_L005.R2.fastq
/work/lianglab/wgs/lavender/FASTQ/C48LWACXX/DNA_SG_ACAGTG_L007.R1.fastq
/work/lianglab/wgs/lavender/FASTQ/C48LWACXX/DNA_SG_ACAGTG_L007.R2.fastq
/work/lianglab/wgs/lavender/FASTQ/C48LWACXX/DNA_SG_GTGAAA_L007.R1.fastq
/work/lianglab/wgs/lavender/FASTQ/C48LWACXX/DNA_SG_GTGAAA_L007.R2.fastq
/work/lianglab/wgs/lavender/FASTQ/OverlapLib/PE_l50T230.1.fastq
/work/lianglab/wgs/lavender/FASTQ/OverlapLib/PE_l50T230.2.fastq
/work/lianglab/wgs/lavender/FASTQ/OverlapLib/PE_l50t2229.1.fastq
/work/lianglab/wgs/lavender/FASTQ/OverlapLib/PE_l50t2229.2.fastq  >PE_All_raw_overlap_fermi.mak
make -f PE_All_raw_overlap_fermi.mak -j 6 &> run_fermi_4PE_All_raw_overlap.log &
```

B.4. Scaffold generation

B.4.a. Preprocess sequencing reads to be used as input for running OPERA-LG

```
sqsub --mpp 30G -r 167h -q threaded -n 10 -o run_opera_preproc_PEL5A.log perl
/home/rn13ow/work/opera_v2.0.1/bin/preprocess_reads.pl ../lavender_raw_fermi.p5.fa
/work/lianglab/wgs/lavender/FASTQ/C48NNACXX/DNA_SG_ACAGTG_L005.R1.fastq
/work/lianglab/wgs/lavender/FASTQ/C48NNACXX/DNA_SG_ACAGTG_L005.R2.fastq PEL5A.map.bam bwa &>
run_opera_preproc_PEL5A.log
  /home/rn13ow/work/OPERA-LG_v2.0.2/bin/OPERA-LG multiLib.config &>
run_operaLG_4fermi_alllib_overlap-iq.log &
```

B.4.b. MultiLib.config sample file for processing OPERA-LG

```
# Essential Parameters
#
# Output folder for final results
output_folder=test_dataset/results

# Contig file
contig_file=test_dataset/contigs.fa

#Samtools directory; to leave out this option if samtools is on PATH
#samtools_dir=
kmer=##

# Specify if repeats will be scaffolded or not
# no: scaffold repeats
# yes: do not scaffold repeats (default)
#filter_repeat=no

# coverage for haploid sequence can also be specified
(recommend to calculate this value by OPERA-LG)
#haploid_coverage=10

# Mapped read locations
[LIB]
map_file=test_dataset/lib_1.bam
cluster_threshold=5
#lib_mean=10000
#lib_std=1000

[LIB]
map_file=test_dataset/lib_2.bam
cluster_threshold=5
#lib_mean=300
#lib_std=30
```

B.5.GapCloser

```
GapCloser -b config_file -a /home/rn13ow/work/scaffolds_complete/fermi_raw_reads_opera_k79.fasta  -o
fermi_opera_k79_scaffold_gapfill.fa &
```

Post-assembly genome improvement process

B.6. L_RNA_Scafolder

```
blat -minIdentity=98 -minScore=100 -noHead
/home/rn13ow/work/scaffolds_complete/fermi_raw_reads_opera_k59.fasta
/work/lianglab/transcriptome/Lavandula/lavender_RNA-seq/L.ang/L.ang_best_transcripts.fa
fermi_contigs_opera_k79_L.ang_best_trans.psl &> run_blat_4fermi_opera_k79_L.ang_best_trans.log

module load bioperl
/work/lianglab/bin/L_RNA_scaffolder/L_RNA_scaffolder.sh -d /work/lianglab/bin/L_RNA_scaffolder/ -i
fermi_contigs_opera_k79_L.ang_best_trans.psl -j
/home/rn13ow/work/scaffolds_complete/fermi_raw_reads_opera_k79.fasta &>
run_LRNA_scaf_4fermi_opera_k79_L.ang_best_trans.log &
```

B.7. rDNA_Linker –in-house tool

```
rdna.pl –-/home/rn13ow/work/scaffolds_complete/fermi_raw_reads_opera_k79.fasta –-fermi_blat.psl
```

B.8. Quality check of contigs/scaffolds using QUAST tool.

```
module load intel/12.1.3
module load python/intel/2.7.8
/work/lianglab/bin/quast-2.3/quast.py -o /home/rn13ow/scratch/quast_4fermi_alllib_R1R2/ -e -T 5 -s --est-ref-size 879915704 /home/rn13ow/work/complete_contigs/lavender_raw_fermi.p5.fa &>run_quast_4fermi_alllib_R1R2.log &
```

B.9. Genome completeness check using BUSCO

```
python3 /home/rn13ow/work/BUSCO_v2/BUSCO_v2/BUSCO.py -i /home/rn13ow/work/1Dec_2017_Lav_Ref_Assembly/fermi_PE_filt_overlap_opera_k79_scaf_new_Feb10.fa -o LA_fermi_opera_genome_initial -l /home/rn13ow/work/BUSCO_v2/BUSCO_v2/embryophyta_odb9/ -m geno -c 5 -sp arabidopsis &>run_BUSCOv2_LA_genome_initial.log &
```

B.10. *De novo* genome annotation

A.10.a. Maker genome annotation pipeline - .ctl files and run command

A.10.a.1. MAKER pipeline – control files - .ctl files
maker_exe.ctl

```
#-----Location of Executables Used by MAKER/EVALUATOR
makeblastdb=/home/rn13ow/work/blast-2.6/ncbi-blast-2.6.0+/bin/makeblastdb #location of NCBI+ makeblastdb executable
blastn=/home/rn13ow/work/blast-2.6/ncbi-blast-2.6.0+/bin/blastn #location of NCBI+ blastn executable
blastx=/home/rn13ow/work/blast-2.6/ncbi-blast-2.6.0+/bin/blastx #location of NCBI+ blastx executable
tblastx=/home/rn13ow/work/blast-2.6/ncbi-blast-2.6.0+/bin/tblastx #location of NCBI+ tblastx executable
formatdb=/home/pliang/bin/formatdb #location of NCBI formatdb executable
blastall=/work/lianglab/bin/blastall #location of NCBI blastall executable
xdformat= #location of WUBLAST xdformat executable
blast= #location of WUBLAST blast executable
RepeatMasker=/work/lianglab/bin/RepeatMasker/RepeatMasker #location of RepeatMasker executable
exonerate=/work/lianglab/bin/exonerate #location of exonerate executable

#-----Ab-initio Gene Prediction Algorithms
snap=/work/rn13ow/maker_p/maker/bin/../exe/snap/snap #location of snap executable
gmhmme3= #location of eukaryotic genemark executable
gmhmmp= #location of prokaryotic genemark executable
augustus=/home/rn13ow/work/augustus.2.5.5/bin/augustus #location of augustus executable
fgenesh= #location of fgenesh executable
tRNAscan-SE=/home/rn13ow/bin/tRNAscan-SE #location of trnascan executable
snoscan= #location of snoscan executable

#-----Other Algorithms
probuild= #location of probuild executable (required for genemark)
```

B.10.a.2. Maker_bopts.ctl – Blast options control file

```
#-----BLAST and Exonerate Statistics Thresholds
blast_type=ncbi+ #set to 'ncbi+', 'ncbi' or 'wublast'

pcov_blastn=0.8 #Blastn Percent Coverage Threhold EST-Genome Alignments
pid_blastn=0.85 #Blastn Percent Identity Threshold EST-Genome Alignments
eval_blastn=1e-10 #Blastn eval cutoff
bit_blastn=40 #Blastn bit cutoff
depth_blastn=0 #Blastn depth cutoff (0 to disable cutoff)

pcov_blastx=0.5 #Blastx Percent Coverage Threhold Protein-Genome Alignments
pid_blastx=0.4 #Blastx Percent Identity Threshold Protein-Genome Alignments
eval_blastx=1e-06 #Blastx eval cutoff
bit_blastx=30 #Blastx bit cutoff
depth_blastx=0 #Blastx depth cutoff (0 to disable cutoff)

pcov_tblastx=0.8 #tBlastx Percent Coverage Threhold alt-EST-Genome Alignments
pid_tblastx=0.85 #tBlastx Percent Identity Threshold alt-EST-Genome Alignments
eval_tblastx=1e-10 #tBlastx eval cutoff
bit_tblastx=40 #tBlastx bit cutoff
depth_tblastx=0 #tBlastx depth cutoff (0 to disable cutoff)

pcov_rm_blastx=0.5 #Blastx Percent Coverage Threhold For Transposable Element Masking
pid_rm_blastx=0.4 #Blastx Percent Identity Threshold For Transposbale Element Masking
eval_rm_blastx=1e-06 #Blastx eval cutoff for transposable element masking
bit_rm_blastx=30 #Blastx bit cutoff for transposable element masking

ep_score_limit=20 #Exonerate protein percent of maximal score threshold
en_score_limit=20 #Exonerate nucleotide percent of maximal score threshold
```

B.10.a.3. maker_opts.ctl

```
#-----Genome (these are always required)
genome= #genome sequence (fasta file or fasta embeded in GFF3 file)
organism_type=eukaryotic #eukaryotic or prokaryotic. Default is eukaryotic
#-----Re-annotation Using MAKER Derived GFF3
maker_gff= #MAKER derived GFF3 file
est_pass=0 #use ESTs in maker_gff: 1 = yes, 0 = no
altest_pass=0 #use alternate organism ESTs in maker_gff: 1 = yes, 0 = no
protein_pass=0 #use protein alignments in maker_gff: 1 = yes, 0 = no
rm_pass=0 #use repeats in maker_gff: 1 = yes, 0 = no
model_pass=0 #use gene models in maker_gff: 1 = yes, 0 = no
pred_pass=0 #use ab-initio predictions in maker_gff: 1 = yes, 0 = no
other_pass=0 #passthrough anyything else in maker_gff: 1 = yes, 0 = no
#-----EST Evidence (for best results provide a file for at least one)
est=/work/lianglab/transcriptome/Lavandula/assembly/L.ang_allSeq.cap3min100.fa.annot #set of ESTs or
assembled mRNA-seq in fasta format
altest= #EST/cDNA sequence file in fasta format from an alternate organism
est_gff= #aligned ESTs or mRNA-seq from an external GFF3 file
altest_gff= #aligned ESTs from a closely relate species in GFF3 format
#-----Protein Homology Evidence (for best results provide a file for at least one)
protein=/work/lianglab/DB/blastDB/plant/planRefPep.fa  #protein sequence file in fasta format (i.e.
from multiple organisms)
protein_gff=  #aligned protein homology evidence from an external GFF3 file
#-----Repeat Masking (leave values blank to skip repeat masking)
model_org=all #select a model organism for RepBase masking in RepeatMasker
rmlib= #provide an organism specific repeat library in fasta format for RepeatMasker
repeat_protein=/work/rn13ow/maker/data/te_proteins.fasta #provide a fasta file of transposable
element proteins for RepeatRunner
rm_gff= #pre-identified repeat elements from an external GFF3 file
```

```
prok_rm=0 #forces MAKER to repeatmask prokaryotes (no reason to change this), 1 = yes, 0 = no
softmask=1 #use soft-masking rather than hard-masking in BLAST (i.e. seg and dust filtering)
#-----Gene Prediction
snaphmm= #SNAP HMM file
gmhmm= #GeneMark HMM file
augustus_species=arabidopsis #Augustus gene prediction species model
fgenesh_par_file= #FGENESH parameter file
pred_gff= #ab-initio predictions from an external GFF3 file
model_gff= #annotated gene models from an external GFF3 file (annotation pass-through)
est2genome=1 #infer gene predictions directly from ESTs, 1 = yes, 0 = no
protein2genome=1 #infer predictions from protein homology, 1 = yes, 0 = no
trna=0 #find tRNAs with tRNAscan, 1 = yes, 0 = no
snoscan_rrna= #rRNA file to have Snoscan find snoRNAs
unmask=0 #also run ab-initio prediction programs on unmasked sequence, 1 = yes, 0 = no
#-----Other Annotation Feature Types (features MAKER doesn't recognize)
other_gff= #extra features to pass-through to final MAKER generated GFF3 file
#-----External Application Behavior Options
alt_peptide=C #amino acid used to replace non-standard amino acids in BLAST databases
cpus=4 #max number of cpus to use in BLAST and RepeatMasker (not for MPI, leave 1 when using MPI)
#-----MAKER Behavior Options
max_dna_len=100000 #length for dividing up contigs into chunks (increases/decreases memory usage)
min_contig=1 #skip genome contigs below this length (under 10kb are often useless)
pred_flank=200 #flank for extending evidence clusters sent to gene predictors
pred_stats=0 #report AED and QI statistics for all predictions as well as models
AED_threshold=1 #Maximum Annotation Edit Distance allowed (bound by 0 and 1)
min_protein=0 #require at least this many amino acids in predicted proteins
alt_splice=0 #Take extra steps to try and find alternative splicing, 1 = yes, 0 = no
always_complete=0 #extra steps to force start and stop codons, 1 = yes, 0 = no
map_forward=0 #map names and attributes forward from old GFF3 genes, 1 = yes, 0 = no
keep_preds=0 #Concordance threshold to add unsupported gene prediction (bound by 0 and 1)
split_hit=10000 #length for the splitting of hits (expected max intron size for evidence alignments)
single_exon=0 #consider single exon EST evidence when generating annotations, 1 = yes, 0 = no
single_length=250 #min length required for single exon ESTs if 'single_exon is enabled'
correct_est_fusion=0 #limits use of ESTs in annotation to avoid fusion genes
tries=5 #number of times to try a contig if there is a failure for some reason
clean_try=0 #remove all data from previous run before retrying, 1 = yes, 0 = no
clean_up=1 #removes theVoid directory with individual analysis files, 1 = yes, 0 = no
TMP= #specify a directory other than the system default temporary directory for temporary files
```

B.10.b. MAKER pipeline – Run command

```
/home/rn13ow/work/maker/bin/maker -genome sub_1 -RM_off --ignore_nfs_tmp -TMP
/home/rn13ow/scratch/Oct_lav_draft_100_chnks/sub_1_maker  maker_opts.ctl maker_bopts.ctl
maker_exe.ctl &>run_maker_sub_1_iq_test.log1 &

### - .ctl files review each of the file - to highlight key parameters and options provided.

### - Running MAKER pipeline on split draft genome sequences
/home/rn13ow/work/maker/bin/gff3_merge -d
/work/lianglab/wgs/lavender/Annotation/Oct_lav_775_chnks/sub_6_maker/sub_6.maker.output/sub_6_master
_datastore_index.log -n -o sub_6.gff &
 ## Merging individually generated gff3 files into a combined .gff3 file
for i in
{757,758,759,760,761,764,773,775,80,81,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,9,8,68,74,
75,79,6}; do /home/rn13ow/work/maker/bin/gff3_merge -d
/work/lianglab/wgs/lavender/Annotation/Oct_lav_775_chnks/sub_"$i"_maker/sub_"$i".maker.output/sub_"$
i"_master_datastore_index.log  -n -o sub_"$i".gff; done &
```

B.11. Post-annotation improvement pipeline

B.11.a. Process_orf. Pl

```
#!/usr/bin/perl -w

use strict;
(@ARGV &&  -s $ARGV[0]) || die "Need a file or the file is not exist.\n";

open (IN, "<$ARGV[0]") or die "can not open $ARGV[0]\n";
my ($gorf,$nstart,$nstop,$bad)=(0,0,0,0);
while(my $record = get_next_record(*IN)){
    if (!$record){next}
    my ($defline,$seq) = $record =~/^(.+?)\n(.+)$/sm;
    $seq =~ s/[\s\d\n\-]//g; #remove space, number, and newline
    my $len = length($seq);
    if (!$len || $len<30){next}
    my $orf;
    if ($seq=~/^M[A-W]{29,}\*$/){$gorf++; print "$record"}
    elsif($seq=~/^[A-W]{30,}\*$/){$nstart++; print "$record"}
    elsif($seq=~/[A-W]{30,}/){$nstop++; print STDERR "$record"}
#    else{$bad++; print STDERR "$record"}
}
close IN;
print STDERR "$gorf perfect ORFs, $nstart with no start condon; $nstop with no stop codon, $bad
being anything else.\n";

exit;
####################end of main#########################
sub get_next_record {

    my($fh) = @_;
    my($record) = '';
    my($save_input_separator) = $/;

    $/ = "\n>";
    $record = <$fh>;
    if (!$record){return 0}
    $record =~ tr/>//d; $record =">".$record;
    $/ = $save_input_separator;
    return $record;
}
```

B.11.b Blastp running

```
for f in {1..5}; do sqsub --mpp 6G -r 8h -q threaded -n 10 -o run_blastp_sub_$f.log blastp -db
PlantRefPep -query sub_$f -evalue 1e-20 -max_target_seqs 3 -out sub_”$f”.bl6 -outfmt 7 -num_threads
10 &>run_final_blastp_sub_$f.log ; done
B.1
```

B.11.c. GAG to generate initial protein sequence

```
module load intel/12.1.3
module load python/intel/2.7.8

/home/rn13ow/work/GAG/genomeannotation-GAG-98da78e/gag.py --fasta
/home/rn13ow/scratch/Lav_final_assembly_l200.fa --gff ../genome.all.gff --out Lav_final_gff_new_out
&>run_GAG_final_genome_gff.log &
```

B.11.d. Genevalidator – run command and filtering based on 3 important criteria (perl run command)

```
for f in {51..100}; do sqsub --mpp 10G -r 5h -q threaded -n 3 -o run_gv_sub_"$f"_200.log
/home/rn13ow/bin/bin/genevalidator -d ../planRefPep.fa -v all -n 3 sub_$f
&>run_gv_sub_"$f"_tmp_nonstop.log; done

### filtering the protein sequences and generating the id for using in the generat_final_files.pl
perl -ne 'my @f=split(/\t/,$_); my $s=$f[1]; my ($len)=$f[5]=~/^(\d+)/; my
($c,$e,$m)=$f[8]=~/^(\d+).+?(\d+).+?(\d+)/; if ($s>=35 && $c>40 && $e<=50 && $m<=40){print "$_"}'
Final_GV_results.list |cut -f3 >Lav_prot_Sep_Final_GV_filtered.id
```

B.11.e. Generate_final_files.pl – to generate final annotation files.

```
 #!/bin/perl -w

#this script is used to generate the final gff, protein sequences, and cDNA sequences for the
proteins we want to keep.

#use strict;

if (!$ARGV[0] || ! -s $ARGV[0]){print "$0 input_LV_prot_GV_filteredList.\n"; exit}

#read in the final protein list from LV_prot_GV_filtered.id
open(ID, "<$ARGV[0]") or die "can't open $ARGV[0]";  #maker-opera_scaffold_1032-augustus-gene-0.241-
mRNA-1
my @ID=<ID>; close (ID);
chomp @ID;


#collect the scalffold IDs, and organize the gene IDs into scaffolds
my (%sf2Genes, %GoodS);
my($SN,$GN,$skipped,$modified,$unmod)=(0,0,0,0,0);
foreach my $id (@ID){
    $GN++;
    my ($tmp,$SID)=split(/\-/,$id);
    if (!$sf2Genes{$SID}){$SN++}
    $sf2Genes{$SID}=$SID;
    push @{$sf2Genes{$SID}}, $id;
}
print STDERR "$GN gene(s) in $SN contigs/scaffolds included in the input.\n";
($SN,$GN)=(0,0); #reset the scaffold and gene numbers, such they are to report the actual number
processed in the 2nd round
open (GFF,">$ARGV[0].gff") or die "can't open $ARGV[0].gff:$!\n";
open (PROT,">$ARGV[0].faa") or die "can't open $ARGV[0].faa:$!\n";
open (CDS,">$ARGV[0].fna") or die "can't open $ARGV[0].fna:$!\n";
open (MOD,">$ARGV[0].modified") or die "can't open $ARGV[0].modified:$!\n";
open (SKIP,">$ARGV[0].skipped") or die "can't open $ARGV[0].skipped:$!\n";

#Collect top match refprot info: sequence length and annotation
#NP_001031628    440     S-adenosyl-L-homocysteine hydrolase [Arabidopsis thaliana]

open(SLEN,"< RefProt_tophit.acc_len_annot") or die "can't open  RefProt_tophit.acc_len_annot:$!.\n";
my %REFP;
while(<SLEN>){
    chomp;
    my($id,$len,$annot)=split("\t",$_);
    if ($annot){$REFP{$id}={L=>$len,A=>$annot}}
}
close SLEN;

#process all genes in groups of scaffolds
print GFF "##gff-version 3\n";# print GFF "\n";
my ($k,$seq);
foreach $k (keys %sf2Genes){
    #obtain the sequence from the assembly
    $SN++;
    print STDERR "Processing contig/scaffold: $k ....\n";
    $seq=`fatools -W $k -p nd
/work/lianglab/wgs/lavender/Annotation/Oct_lav_775_chnks/Lav_final_assembly_l200.fa`;
    if (!$seq){print SKIP "$k has no sequence extracted.\n\n"; next}
    my $slen=length($seq);
    #print the scaffold sequence as a file for running tblastn, do it once for a scaffold
    open(OUT,">tmp_$k.fna");
    print OUT ">$k\n$seq\n";
```

```
    close OUT;
    #process all genes within a scaffold
    my @genes=@{$sf2Genes{$k}};
    my $gn=scalar @genes;
    print STDERR "$gn gene(s) included in scaffold/contig: $k.\n";
    if (@genes){print GFF "$k\t.\t.\tscaffold\t1\t$slen\t.\t.\t.\tID=$k;Name=$k\n"}#print
scaffold line

    foreach my $g (@genes){ #process genes in a scaffold one by one
    #collect the protein sequence
          $GN++;
          print STDERR "processing: $g in $k ....\n";
          my $pseq=`fatools -D $g -p nd t3.fa`;
          my $plen=length($pseq); #length the final protein sequenc
          my $note;
          #collect the regions matched to plantRefseq
          #augustus-opera_scaffold_1766-processed-gene-0.17-mRNA-1
gi|747065062|ref|XP_011079143.1|  54.92 122 4 84 271 389 327 444 5e-27  117
          #GRKYCAVDRET*ELLGLQDRRNEGF*YRDSVQAFAFGLDYHSSCRIGFHARESFNFSSLGTGS**IWPETSYNIWHIFSGYIQYNFWC*
RKLLDGLFYKVSSWKFMWNARTFKD*HFLGHRLGHWSCCWRLSCPACRKISNHLLSKLYIWKVSIFFALPCDLNICSGVLTPLLLAS*NCSHP**RKQSP
RFSVKESCNGGPYS*VH*VKESGISKKPLKKLALDVCYNGVLHFPTS*HGLY*DILFVGYQPKETWGIKF*LGRCW*RSFNHRGWNAILSTHSLYLDRED
TWTGHGFSYRSCHYNTIAIKLPIFSQIIGSRSCTAA*FCFNLEECTLYIYNNWFVSSSKQSCAERAKRSCKWCIHDCYVSL*SNWPCSWRFPALIRSNPS
*CRLLTRCSPGFFHVKRG*VYWSRYDLQTDSGSPGRRQATEFGSGPRKPLKL
          #genes with stop codon in the aligned protein sequences are considered pseudogenes
          my @f =split(/\t/,`grep $g blastp_top1hit.bl6.txt`);  #the file contains the single top
hit for each protein to the plantRef protein sequences
 #maker-opera_scaffold_3121-augustus-gene-0.132-mRNA-1    gi|747076267|ref|XP_011085198.1|
59.28   830     306     7       6       820     83      895     0.0       920
          my ($qs,$qe)=@f[6,7];  #alignment start and end for query
          my ($ss,$se)=@f[8,9];  #alignment start and end for the subject
          my @blastp=split(/\|/,$f[1]); my ($acc)=split(/\./,$blastp[3]);  #remove the verson number
for the acc#
          my ($sslen,$sannot)=($REFP{$acc}->{L},$REFP{$acc}->{A}); #length and annotation for the
matched refProtei
          if (!$sslen){($sslen,$sannot)=($REFP{$blastp[3]}->{L},$REFP{$blastp[3]}->{A})} #some acc
have the version number
          $note="|$sannot"; #annotation of the matched ref Protein
          if ($pseq=~/^M[A-Z]{30,}\*$/){$note.="|N-term:complete|C-term:complete";
print_gff($g,$pseq,$note); next} #print gff lines without modification if the protein sequence
starts from M and ends with a stop codon and is minimally 31 AA,

          my $aligned_s=substr($pseq,$qs-1,$qe-$qs+1); #extract the aligned region
          my $aligned_s1=$aligned_s;
#         if ($aligned_s=~/[A-Z]{5,}\*[A-Z]{5,}/){print SKIP
">$g|original:NoGoodProtein\n$pseq\n>$g|Aligned\n$aligned_s\n\n"; $skipped++; next} #skip squences
with stop codon in the middle of the aligned sequence
          ($aligned_s)=$aligned_s=~/\*{0,1}([A-Z]{30,}\*{0,1})/;  #extract the longest non-stop
sequence in the aligned region
          if (!$aligned_s){print SKIP ">$g|original:ProteinTooShort\n$pseq\n>$g|Aligned_s\n\n";
$skipped++; next}
#         my $as=$qs-50; if ($as<0){$as=1} #expan the alignment region on query by 20 AA
#         my $ae=$qe+50; if ($ae>$plen){$ae=$plen}
#         my $aligned_b=substr($pseq,$as-1,$ae-$as+1); #extract the expanded protein sequences
around the aligned region
#         my ($aligned)=$aligned_b=~/\*{0,1}([A-Z]*${aligned_s}[A-Z]*\*{0,1})/;  #extract the
maximal ORF from the expanded aligned sequences
          my ($aligned)=$pseq=~/\*{0,1}([A-Z]*${aligned_s}[A-Z]*\*{0,1})/;  #expand the aligned
region to get the maximal ORF from the original protein sequence, including the optional last stop
codon
        #if there is start codon at the N-termal at position equivalent to the subject protein, then
use this M as the start of the proteins sequences
          if ($aligned=~/[A-LN-Z]{0,$ss+20}M[A-Z]{29,}\*{0,1}$/){#Take the 1st M located with
equivalenent position in subject sequence as the start codon, need to add a start codon line for the
gff file late
              ($aligned)=$aligned=~/[A-LN-Z]{0,$ss+20}(M[A-Z]{29,}\*{0,1})$/
```

```
        }
        if (!$aligned){print SKIP
">$g|original:NoGoodProtein\n$pseq\n>$g|Aligned\n$aligned_s\n\n"; $skipped++; next}

        #need to decide whether a protein is complete on the N- and C-terminal based on alignment
with the refSeq

        my $qs2=index $pseq,$aligned; $qs2++; #convert to 1-based position
        my $qe2=$qs2+length($aligned)-1; #the end position of the final aligned sequence
        if ($ss<=30){$note.="|N-term:complete"} #if the alignment is at the beginning for the
subject, then the N-terminal are cosnidered complete
        elsif($ss>30 && $qs2<=10){my $missN=$ss-1-($qs-$qs2); $note.="|N-term:incomplete,$missN"}
        else{$note.="|N-term:unknown"}
        if ($sslen-$se<=30 && $aligned=~/\*$/){$note.="|C-term:complete"} #if the aligned region
reaches to the end of subject and the sequence ends with a stop codon
        elsif ($sslen-$se>30 && $aligned!~/\*$/){my $missC=$sslen-$se-abs($qe2-$qe); $note.="|C-
term:incomplete,$missC"} #subject has more than 30AA left at the C-terminus, and the sequence
doesn't ends with a stop codon
        else{$note.="|C-term:unknown"}
        if ($aligned eq $pseq){print_gff($g,$pseq,$note); next};  #entries with no adjustment to
the sequneces, copy the existing gff lines

        $plen=length($aligned); #update protein length
        #save the modified protein sequence and the scaffold sequence for running tblastn
        open(OUT,">tmp_$g.faa");
        print OUT ">$g\n$aligned\n";
        close OUT;

        #obtain the genomic positions correspond to protein sequences represented by $aligne using
bl2seq
#       my @hits=`bl2seq -p tblastn -i $g.fa -j $k.fna -D 1 -F F -e 1e-3`;
        open(BL2SEQ, "bl2seq -p tblastn -i tmp_$g.faa -j tmp_$k.fna -D 1 -F F -e 1e-3|");
      #Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q.
start, q. end, s. start, s. end, e-value, bit score
        #maker-opera_scaffold_79802-exonerate_protein2genome-gene-0.0-mRNA-1
opera_scaffold_79802    100.00  41      0       0       1       41      374     252     1e-29   84.3
        my @hits=<BL2SEQ>; close (BL2SEQ);
        if (!@hits){print SKIP ">$g|original:NoMatch2DNA\n$pseq\n>$g|Aligned\n$aligned\n\n";
$skipped++; next}
        my ($gs,$ge,$st)=(0,0,"+");
        foreach my $l (@hits){
            if (!$l ||$l=~/^#/){next} #skip the header and empty lines
            #print "$l"; #debug
            my @f=split(/\t/,$l);
            if ($f[8]>$f[9]){$st="-"} #update strand info based on tbastn result
            my ($s,$e)=@f[8,9];  #alignment position on the subject sequence
            if ($st eq "+"){
                  if(!$gs ||$s<$gs){$gs=$s}  #obtain minimal start
                  elsif(!$ge ||$e>$ge){$ge=$e} #obtain maximal end position of the gene in the
scaffold
            }else{
                  if(!$gs || $e<$gs){$gs=$e}  #obtain minimal start
                  elsif(!$ge || $s>$ge){$ge=$s} #obtain maximal end position of the gene in the
scaffold
            }
        }#$gs is always the 5' position and $ge is the 3' position in the scaffold
        if ($aligned=~/[A-Z]+\*$/){  #if the modified protein sequence ends with a stop codon,
extend gene end position by 3
            if ($st eq "+"){$ge+=3}
            else{$gs-=3} #if the protein is on the minus strand
        }
        print_gff($g,$pseq,$note,$st,$aligned_s,$aligned,$gs,$ge);
        unlink("tmp_$g.faa");
    }
    unlink("tmp_$k.fna");
```

```
}
close (GFF); close (CDS); close (PROT); close (MOD), close (SKIP);
print STDERR "process completed for $SN scaffolds and $GN genes ($unmod unchanged, $modified
modified, $skipped skipped).\n";
exit 0;


sub print_gff{ #extrat the gff lines from the large gff file for a gene, making necessary
modification/filtering before printing
    my ($g,$pseq,$note,$st,$aligned_s,$aligned,$gs,$ge)=@_;
    print STDERR "process gff lines for $g....\n";
    #extract and modified the gff lines
    my $G=$g; #save the original ID in $G
    $g=~s/\-mRNA\-1//;  #obtain gene ID
    open(GFFL,"grep -e $g -e $G LV_final_raw.gff|"); #extract all gff lines for the gene from the
large gff files
    my @gff=<GFFL>; close (GFFL);
    if (!@gff){print SKIP ">$g|original:NoMatch2DNA\n$pseq\n\n"; $skipped++; next}
    $g=$G; #recover the gene ID to the mRNA ID
    my ($CDS,$GFF,$seenStart,@mRNA);
          #retrieve and modify gff lines
          #RDNAU_58_10161  maker   gene    176     457     .       +       .       ID=maker-
RDNAU_58_10161-augustus-gene-0.10;Name=maker-RDNAU_58_10161-augustus-gene-0.10
          #RDNAU_58_10161  maker   mRNA    176     457     .       +       .       ID=maker-
RDNAU_58_10161-augustus-gene-0.10-mRNA-1;Parent=maker-RDNAU_58_10161-augustus-gene-0.10
          #RDNAU_58_10161  maker   exon    176     198     .       +       .       ID=maker-
RDNAU_58_10161-augustus-gene-0.10-mRNA-1:exon:860;Parent=maker-RDNAU_58_10161-augustus-gene-0.10-
mRNA-1
          #RDNAU_58_10161  maker   exon    292     457     .       +       .       ID=maker-
RDNAU_58_10161-augustus-gene-0.10-mRNA-1:exon:861;Parent=maker-RDNAU_58_10161-augustus-gene-0.10-
mRNA-1
          #RDNAU_58_10161  maker   CDS     176     198     .       +       0       ID=maker-
RDNAU_58_10161-augustus-gene-0.10-mRNA-1:cds;Parent=maker-RDNAU_58_10161-augustus-gene-0.10-mRNA-1
          #RDNAU_58_10161  maker   CDS     292     457     .       +       1       ID=maker-
RDNAU_58_10161-augustus-gene-0.10-mRNA-1:cds;Parent=maker-RDNAU_58_10161-augustus-gene-0.10-mRNA-1
          #RDNAU_58_10161  maker   stop_codon      455     457     .       +       .       ID=maker-
RDNAU_58_10161-augustus-gene-0.10-mRNA-1:stop;Parent=maker-RDNAU_58_10161-augustus-gene-0.10-mRNA-1
        #there are also line for start codon
    #print "Gene: $g; start-end: $gs - $ge\n"; #>$g|original\n$pseq\n$>$g|aligned\n$aligned\n";
    foreach my $l (@gff){ #process all gff lines for the gene and collect CDS sequence
          my @f=split(/\t/,$l);
          $st=$f[6];
          if ($gs && $ge){#needs modifications of feature positions
              #print "adjusted genepositon: $gs-$ge:$l"; #for debugging
              if ($f[4]<$gs ||$f[3]>$ge){next} #igore any feature outside of the gne regions based
on the adjusted coordinates
              if ($f[2] eq "gene"){$f[3]=$gs;$f[4]=$ge} #gene line, update the coordinates based on
$gs and $ge
              if ($f[2] eq "mRNA"){#mRNA line, update the coordinates based on $gs and $ge
                    $f[3]=$gs;$f[4]=$ge;
                    @mRNA=@f;  #save the mRNA line for printing it later
              }
              if ($f[2]=~/exon|CDS/){ #CDS lines
                    if ($gs>$f[3] && $gs<$f[4]){$f[3]=$gs} #adjust the left position for overlapping
CDS
                    if ($ge>$f[3] && $ge<$f[4]){$f[4]=$ge} #adjust the right position for
overlapping CDS
              }
              if ($f[2] eq "stop_codon" && ($f[3]<$gs ||$f[4]>$ge)){next}  #ignoe stop codon lines
outside of $gs-$ge
              if ($f[2] eq "start_codon"){ #process start codon lines
                    if ($f[3]>$gs+2 && $f[4]<$ge-2){next} #ingore start codon in the middle of the
sequence
                    $seenStart=1; #used to determine whether a new start codon line is needed for
modified sequence
```

```
            }
            $l=join("\t",@f); #re-generate the modified gff line
        }

        $GFF .= $l;  #collect the gff lines
        #collect CDS sequences
        if ($f[2] =~/CDS|stop_codon/){#extract the DNA sequences for the CDS including the stop
codon
            #print "$l"; #for debugging
            $CDS .=substr($seq,$f[3]-1,$f[4]-$f[3]+1); #append to existing CDS
            #print "$f[3]-$f[4]: $CDS\n"; #for debugging
        }
    }
    if ($aligned && $aligned=~/^M/ && !$seenStart && $gs && $ge){#no start codon line for the
adjusted protein sequences starting from M, add
        my ($start,$sts,$ste); #start codon, and it's position in the scaffold
        if ($st eq "-"){$sts=$ge-2; $ste=$ge} #minus strand
        else{$sts=$gs; $ste=$gs+2} #plus strand
        $mRNA[2]="start_codon"; #modified the start codon line
        $mRNA[3]=$sts;
        $mRNA[4]=$ste;
        $start=join("\t",@mRNA);
        $GFF .="$start\n";  #add start codon line
    }
    if ($GFF && $CDS){
        print GFF "$GFF";
        #print the modified protein sequence to file
        my $PSEQ; #final protein sequence to print
        if ($aligned){#use the modified
            $PSEQ=$aligned
        }else{$PSEQ=$pseq}#used the original protein sequence
        print PROT ">$g$note\n$PSEQ\n";
        #print the updated cDNA sequence to file
        if ($st eq "-"){ #reverse complement the CDS for genes on minus strand
            $CDS=reverse($CDS);
            $CDS=~tr/[ACGTacgt]/[TGCAtgca]/;
        }
        print CDS ">$g$note\n$CDS\n";
        #record entries with adjustment made to the protein sequences
        if ($aligned && $pseq ne $aligned){print MOD
">$g|original\n$pseq\n>$g|Aligned\n$aligned_s\n>$g|Final\n$aligned\n\/\/\n";$modified++}
        else{$unmod++}
    }elsif (!$CDS){
        print SKIP ">$g|original:NoCDS|$st|$gs-$ge\n$pseq\n>$g|Aligned\n$aligned_s\n\n";
$skipped++;
    }else{print SKIP ">$g|original:unknown\n$pseq\n>$g|Aligned\n$aligned_s\n\n"; $skipped++;}
#    unlink ("tmp_$g.faa");
    return 0;
}
```

B.12. Non-coding sequences identification

B.12.a. RNAmmer – rRNA sequences identification

```
for i in {6..100}; do cd sub_"$i"_RNAmmer && sqsub --mpp 1G -r 3h -o run_RNAmmer_sub_$i.log
~/work/RNAmmer/rnammer -S euk -m lsu,ssu,tsu -f sub_$i_res.fasta -gff sub_$i.gff -h sub_$i.hmm
../sub_$i.fa &>run_"$i".log && cd ../; done &
```

B.12.b. tRNAScan – tRNA sequence identification

```
./tRNAscan-SE -G -o /home/rn13ow/scratch/V2A_data/tRNAscan/LA_V2_genome_tRNAscan -f
/home/rn13ow/scratch/V2A_data/tRNAscan/LA_v2_sec_st -m
/home/rn13ow/scratch/V2A_data/tRNAscan/LA_v2genome_tRNAscan_run_stats -H ../LA_genome_final.fa
&>run_tRNAScan-SE_LA_V2_genome.log &
```

B.13. RepeatMasker and RepeatModeler

B.13.a. RepeatModeler command

```
perl ~/work/RepeatModeler/RepeatModeler-open-1.0.11/BuildDatabase -name LA_REF -engine ncbi
../LA_genome_final.fa &>run_BUILD_DB_REPMOD_test.log &
```

B.13.b. RepeatMasker command

```
perl ~/work/RMasker_new/RepeatMasker/RepeatMasker -e ncbi -s -pa 12 -lib consensi.fa.classified -u -
lcambig -xsmall -poly -source -gff -a -frag 40000 LA_genome_final.fa &>run_RM_4RModel_out.log2 &
```

B.14. InterProScan – to classify and add functional information to protein sequences and validate existing annotation.

```
module load intel/12.1.3
module load python/intel/2.7.8
/home/rn13ow/work/InterProScan/InterProScan-5.20-59.0/InterProScan.sh -i
../Lav_GAG_gff_out/genome.proteins.fasta -pa - goterms -iprlookup -dp T
/home/rn13ow/scratch/maker_gff3_merge_final/InterProScan_Lav_GAG_gff_out
&>run_InterProScan_4GAG_out_protein.log &
```

B.15 ConPADE – Estimating ploidy from sequencing reads

```
./conpade -bamName LA_all.bam
./conpade -bamName LA_EO.bam
```

B.16. MITEFINDER II – Identification of MITE elements in plant genomes

```
./bin/miteFinder -input genome_file -output output.txt -pattern_scoring
./profile/pattern_scoring.txt -threshold 0.5
```