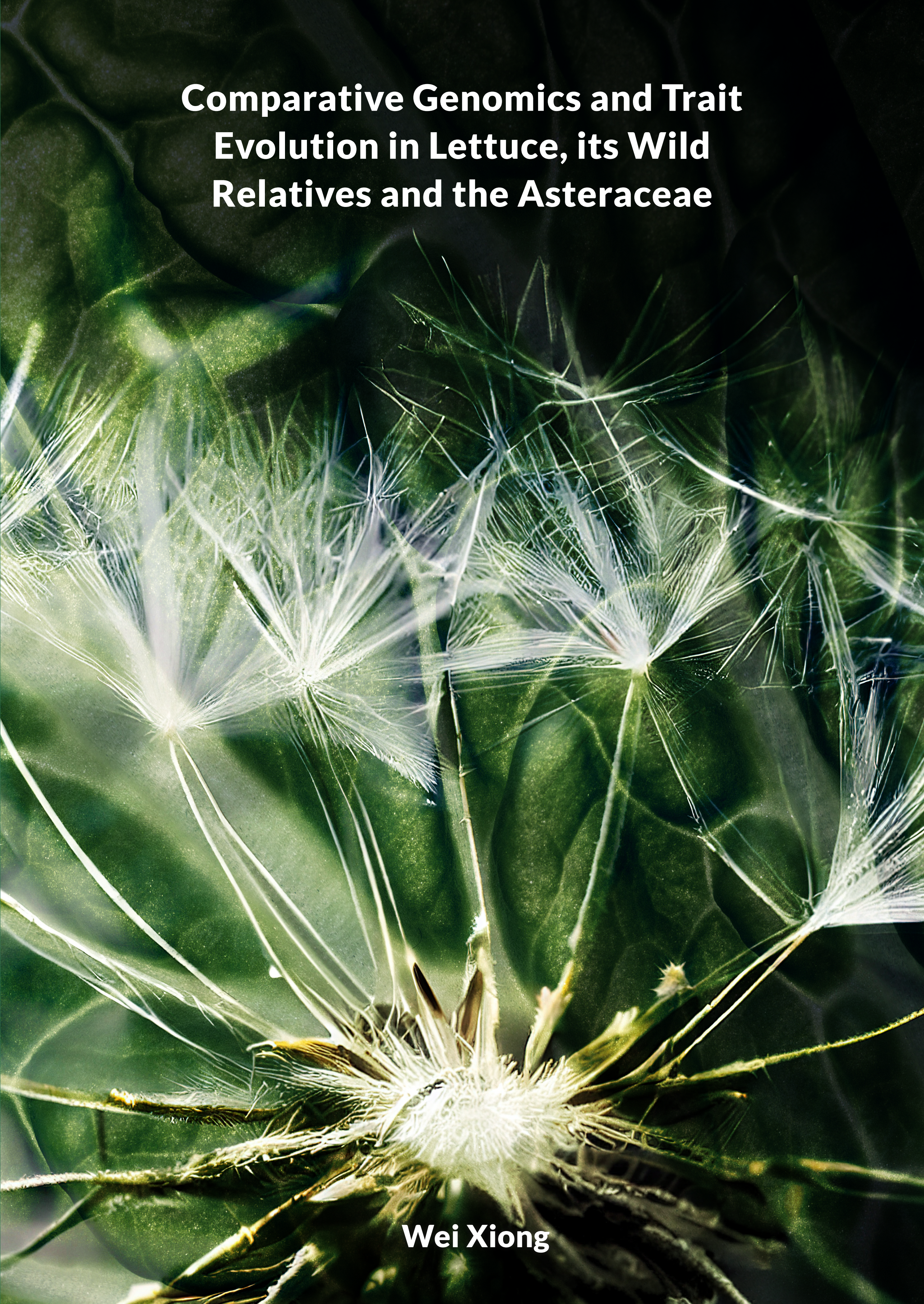


Comparative Genomics and Trait Evolution in Lettuce, its Wild Relatives and the Asteraceae

Wei Xiong

Propositions

1. Effective recognition of RLK/RLP conveys complete resistance to lettuce downy mildew in *Lactuca saligna*.

(this thesis)

2. Transposition of duplicated genes contributes to the diversity and success of the Asteraceae family.

(this thesis)

3. The Matthew effect is integral for understanding the trajectory of any scientific endeavor.

4. At present, high-throughput phenotyping is the bottleneck for Genome-wide association studies.

5. Science has no national borders, while scientists do.

6. The fastest pace is your own pace.

7. Training of expectation management is essential for a Ph.D. candidate.

Propositions belonging to the thesis, entitled

Comparative Genomics and Trait Evolution in Lettuce, its Wild Relatives and the Asteraceae

Wei Xiong
Wageningen, 28th of March 2023

Comparative Genomics and Trait Evolution in Lettuce, its Wild Relatives and the Asteraceae

Wei Xiong

Thesis committee

Promoter
Prof. Dr M.E. Schranz
Professor of Biosystematics
Wageningen University & Research

Other members
Prof. Dr Richard Immink, Wageningen University & Research
Dr Robert van Loo, Wageningen University & Research
Dr Dmitry Lapin, Utrecht University
Dr Saulo Alves Aflitos, Bejo Zaden, the Netherlands

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences.

Comparative Genomics and Trait Evolution in Lettuce, its Wild Relatives and the Asteraceae

Wei Xiong

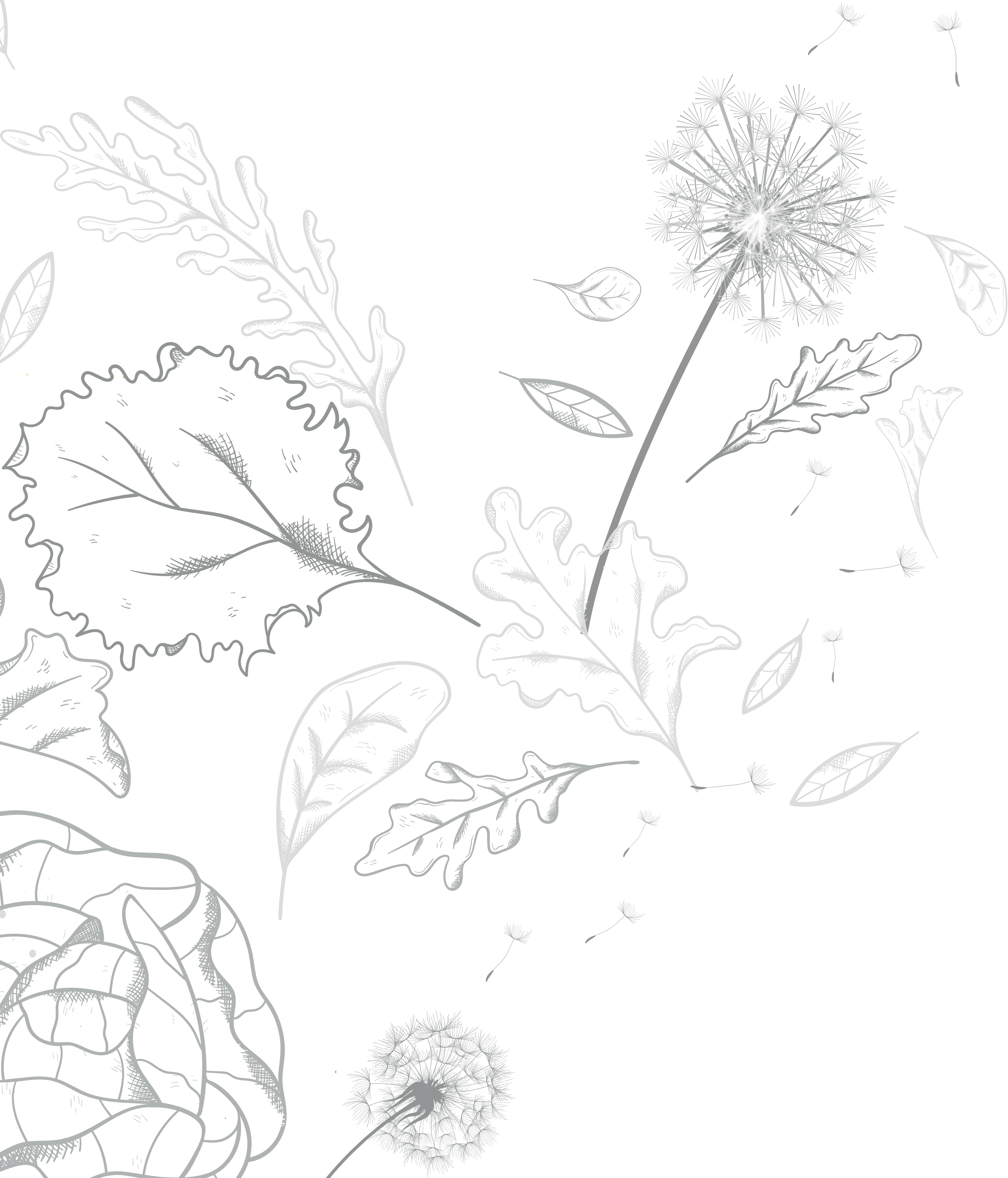
Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board to be defended in public
on Tuesday 28th March 2023
at 1.30 p.m. in the Omnia Auditorium

Wei Xiong
Comparative Genomics and Trait Evolution in Lettuce, its Wild Relatives and the Asteraceae
180 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2023)
With references, with summaries in English
ISBN: 978-94-6447-615-6
DOI: https://doi.org/10.18174/589142

Content

Chapter 1



General introduction

1. Sequencing plant genome

1.1 Crop improvement requires a better understanding of plant genomes

Recently, our planet has welcomed its eight billionth human citizen. However, this welcoming also comes with a warning: The Earth is reaching its limits for crop production. As the population grows, in addition to the need to end starvation, there is also an increasing need for healthier and more nutritious foods (Winson, 2010; Trewern et al., 2021). Practical solutions are required to enhance crop production yield and quality, like, improving disease management and reducing pesticide usage. In addition, new demands raised by changing environments and by innovations in farming systems also need to be addressed (Jorasch, 2019). All these demands require improved traits in crops. Such challenges cannot be tackled without a better understanding of the genetic diversity of crops and their wild relatives leading to novel genes or alleles for economically important traits. Moreover, interspecific compatibility also needs to be assessed on genome organization difference by indicators like structural variants, which can cause hybrid sterility (Shen et al., 2017). With such findings, the breeding of resilient and innovative crops will then be realized. Assembling the genomes of crop species and their relatives is essential to achieve this ambition by providing a high-resolution genome map.

1.2 Techniques for *de novo* genome sequencing and assembly scaffolding

Since the emergence and advancement of sequencing technology, thousands of organisms have been sequenced. For example, over 1,700 genome assemblies are currently available on the National Center for Biotechnology Information (NCBI) platform (Accessed at Nov 2022: https://www.ncbi.nlm.nih.gov/genome/gdv/). Construction of genome assembly contains four major steps: i) *de novo* sequence DNA molecules to generate reads (i.e., sequence bases of DNA); ii) contiguously assemble reads into contigs as a draft assembly using bioinformatic tools; iii) generate mapping data; iv) apply bioinformatic tools on mapping data for scaffolding (i.e., stitched contigs with gaps) with gap filling to improve the draft assembly (Dominguez Del Angel et al., 2018). It is challenging to produce a high-quality assembly due to complex regions in a genome, like repetitive or heterozygous sequences. Among these, repeats are the most prominent obstacle. Unresolved repetitive regions will sabotage the sequence joining resulting in many small contigs or mis-assemblies; and consequently, in incomplete or misrepresented genomes (Sedlazeck et al., 2018). Plant genomes are especially challenging in this aspect due to their typically high repeat-content, which can be up to 90 % in some species (Mehrotra and Goyal, 2014). Here, I focus on the technologies and platforms for step 1 and 3 from the aspect of data generation. Many sequencing and mapping technologies are available but with technical specialties or limitations. The following summary sketches an overview of various techniques back to ~2015 at the set-up for this PhD project (Figure 1).

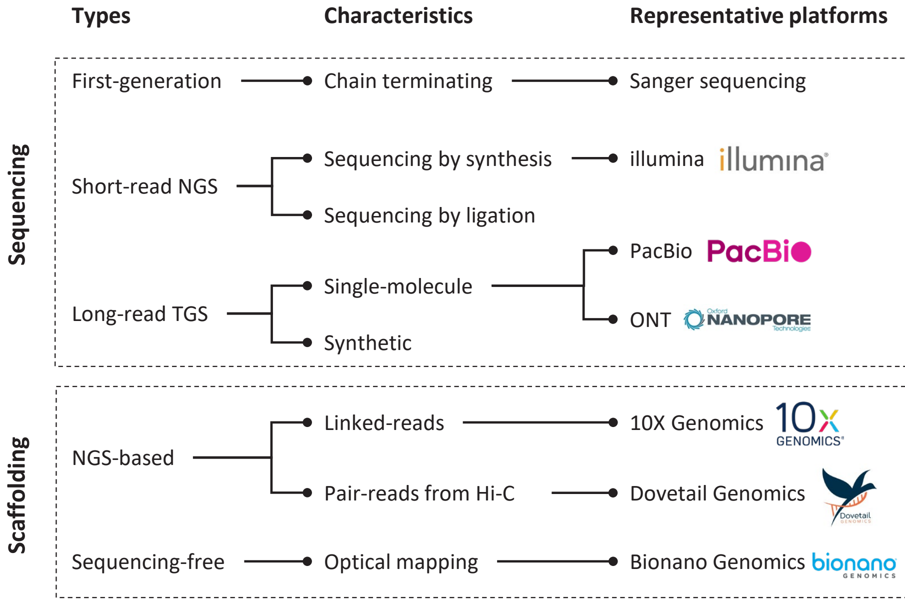


Figure 1. Summary of sequencing and scaffolding technologies for *de novo* genome assembly**.** The representative platforms are examples of platforms for different techniques, and used in sequencing projects described in this PhD thesis, except for Sanger sequencing. NGS, next-generation sequencing; TGS, third-generation sequencing. PacBio, Pacific Biosciences; ONT, Oxford Nanopore Technologies; Hi-C, high-throughput chromosome conformation capture.

In the late 1980s, the first-generation method was developed to sequence the DNA molecules based on the chain-terminating induced by labeled dideoxynucleotides (Sanger et al., 1977). This technique was automated by Sanger Sequencing and enabled first-time sequencing of a broad spectrum of creatures, including the hallmarks of the first genome assembly of human finished in 2003 (Lander et al., 2001; Craig Venter et al., 2001), and model plant *Arabidopsis* in 2000 (Kaul et al., 2000). Despite such achievements, the preparation of first-generation sequencing was laborious and costly, moreover, the runtime was long and throughput is low. Thus, second-generation sequencing was then developed after the human genome project (2004 – 2006) to expand sequencing capacity, also known as next-generation sequencing (NGS). NGS is a high-throughput sequencing approach to generate short-reads from fragmented DNA molecules, where it can efficiently produce a vast amount of data with reduced labor and cost (Liu et al., 2012). NGS was achieved by the advancement of nanotechnology, which facilitated massively parallel sequencing reactions (up to million) from amplified DNA clones (Hu et al., 2021). NGS sequencing can be classified into two types: sequencing-by-ligation (SBL) and sequencing-by-synthesis (SBS), where imaging signals of probe hybridization or nucleotide addition are detected by fluorophore and polymerase on a solid surface respectively (Goodwin et al., 2016). The SBS has quickly replaced Sanger Sequencing and SBS platforms of Illumina technology are the most dominant. NGS was widely applied in sequencing projects of *de novo* genome assemblies and genetic variation. Two famous examples are 'The 1000 Genomes Project' for the global human population (Auton et al., 2015), and '1000 Plant Genomes Project' (Leebens-Mack et al., 2019). While population-scale research has been made possible by SBS, there are, however, still significant restrictions due to innate properties. For example, the short read-length limits the spanning of structural variants (SVs) and causes ambiguity on repetitive regions for genome assembly, while DNA amplification can produce artifacts (Sedlazeck et al., 2018; Hu et al., 2021).

To go beyond the limitations of NGS, long-read sequencing, also known as third-generation sequencing (TGS), was developed. Especially, TGS can produce extremely long-reads (>10 kbp) compared to NGS (Sedlazeck et al., 2018) with less throughput (i.e., tradeoff: length vs. throughput). There are two categories of long-reads: real long-reads of single-molecules, and synthetic long-reads constructed from short-reads *in vitro* (Goodwin et al., 2016). The long-read technologies are represented by single-molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio; Eid *et al.*, 2009) and nanopore-based sequencing from Oxford Nanopore Technologies (ONT; Clarke *et al.*, 2009). Among them, PacBio has been the more widely used platform (Goodwin et al., 2016). Although single-molecule long-read sequencing can eliminate ambiguity and provide high resolution for repeat regions. Long-read technologies suffer from a high-error rate (> 10%) compared to short-read sequencing (< 2%) when they were initially released in the 2010s (Eid et al., 2009; Laver et al., 2015), especially for nanopore sequencing (65% - 88%; Lu, Giordano, and Ning 2016). To correct errors, additional short-reads can be exploited to improve the TGS contigs (i.e., contiguously assembled sequences of long-reads), known as hybrid error correction, which delivers high-quality draft assemblies (Koren et al., 2012; Goodwin et al., 2015).

After sequencing and assembling, different mapping technologies can be applied to scaffold the draft assemblies further and elevate them to near chromosome-level reaching a better resolution for genomes. There are currently three leading genome scaffolding platforms: i) 10X Genomics from Chromium (Zheng et al., 2016) applies barcoding to link short-reads from the second-generation sequencer, spanning hundreds of thousands of bases (~100 kbp); ii) Another platform based on SBS is Hi-C (Burton et al., 2013), where long-range pair-reads (< 100 bp) are created via crosslinking DNA from chromatin interaction; and iii) The optical mapping from Bionano Genomics (Cao et al., 2014) can label a vast range of genome regions with fluorescent probes permitting excellent scaffolding, while moderately costing less than the other two platforms.

To summarize, producing a high-quality plant genome assembly requires the complementation between short- and long-read sequencing and the incorporation of different scaffolding techniques: long-range reads can better resolve the repetitive and complex regions, while short-reads with higher accuracy can correct errors and polish long-read contigs. To increase assembly continuity (e.g., chromosome-level), different mapping tools can elevate the contigs from draft assembly to scaffold or super-scaffold (i.e., longer length). A strategy of such a combination is essential but also challenging, and needs careful pre-design and post-evaluation.

2. The importance of wild relatives in lettuce breeding

2.1 Lettuce as an important leafy crop

Lettuce (*Lactuca sativa* L.) is an important leafy-crop worldwide and contributes to the healthy and nutritious eating habits and goals of many people. Lettuce production has steadily climbed in the last decade, reaching 29 million tons in 2020 with a total value of 20 billion US dollars (FAOSTAT, 2020). In general, lettuce is mainly consumed as a leafy vegetable, but also for its stalk and oil-enriched seeds. Lettuce cultivars can be classified into seven horticultural types by morphological characteristics and edible parts: butterhead, crisphead (iceberg), loose-leaf (cutting), romaine (cos), Latin, stalk (stem), and oilseed (de Vries, 1997; Křístková et al., 2008). Regardless of morphological diversity, similar features like the entire leaf, loss of seed shattering and absence of spines are shared by various lettuce cultivars (de Vries 1997). Lettuce domestication is believed to originate in South West Asia, between Egypt and Iran (Boukema et al., 1990). Many lettuce wild relatives are found in this region from the Euphrates to the Tigris rivers (de Vries, 1997). Wall-paintings of lettuce stored in Egyptian tombs help researchers to date its cultivation history back to 2500BC (Lindqvist, 1960a), when oilseed mostly grown in Egypt is considered as the most ancient domesticated type. After thousands of years of cultivation, breeders are still facing many challenges. One primary goal for lettuce breeding is to develop resilient cultivars against abiotic and biotic stress. In addition, innovative techniques, like vertical farming and LED illumination, have also imposed novel demands on lettuce breeding. To cope with these challenges, knowledge of genetic and genomic makeup is urgently required for wild lettuce to provide variation and overcome introgression hybridization in lettuce breeding.

2.2 Lettuce gene pool and phylogeny

Within the genus *Lactuca* (> 100 species), about 20 species fall into the lettuce gene pool (Lebeda et al., 2004; van Treuren et al., 2012). Among them, *Lactuca serriola*, *L. saligna* and *L. virosa* are representative wild lettuce species, which are extensively studied in lettuce breeding (Lebeda et al., 2014). Taxonomically, *L. sativa* (lettuce) and its putative progenitor *L. serriola* constitute the primary gene pool (i.e., germplasm) for lettuce (Figure 2A), together with six other *serriola*-resembling species from Asia and South Africa (Lindqvist, 1960b; Zohary, 1991; de Vries, 1997). While *L. serriola* is fully fertile with *L. sativa*, the other two representatives suffer from crossing difficulties with lettuce to different extents. Based on cross-compatibility with crop lettuce (Figure 2A), the partially interfertile *L. saligna* and naturally sterile *L. virosa* are therefore classified as the secondary and tertiary gene pool respectively (de Vries, 1997; Zohary, 1991; Křístková et al., 2008; Lebeda et al., 2009). Moreover, the development of molecular and DNA sequencing techniques enables the research of the origin and genetic diversity within the lettuce gene pool. Koopman *et al.* (2001) applied AFLP data for 20 *Lactuca* species (95 accessions) to infer phylogenetic relationships. However, the position of *L. saligna* and *L. virosa* remained uncertain. A more recent phylogenetic study of chloroplast genes confirmed that *L. saligna* and *L. virosa* are sisters to lettuce and its close relatives (Figure 2B; Wei *et al.*, 2017). Further, Zhang *et al*. (2017) indicated that all lettuce cultivars originated from a single domestication event (i.e., common ancestor) using transcriptomic data of 240 *Lactuca* accessions (Figure 2C). Recently, a re-sequencing study of 445 accessions pinpointed that the lettuce's original domestication happened in the Caucasus, accompanied by an iconic loss of seed shattering (Wei et al., 2021). This study's phylogenetic tree of single-copy nuclear genes also showed that *L. saligna* is closer to the primary gene pool species than *L. virosa* (Figure 2D), which is consistent with the classification based on intercrossing.

3. Lettuce wild relatives: *Lactuca saligna* and *L. virosa*

Lactuca saligna is annual wild lettuce (2n=2x=18) with an estimated genome size of 2.3Gb, while *L. virosa* is biennial (2n=2x=18) with an estimated genome size of 3.7Gb (Doležalová et al., 2002). Both wild species flower in summer and are predominant self-fertilizers (Zohary, 1991). They are broadly distributed across Eurasia from the Mediterranean region to temperate Europe and North America (Zohary, 1991; Lebeda et al., 2019). As weedy plants, they commonly grow in waste places or ruderal habitats, like roads, ditches, and river banks (Lebeda et al., 2019).

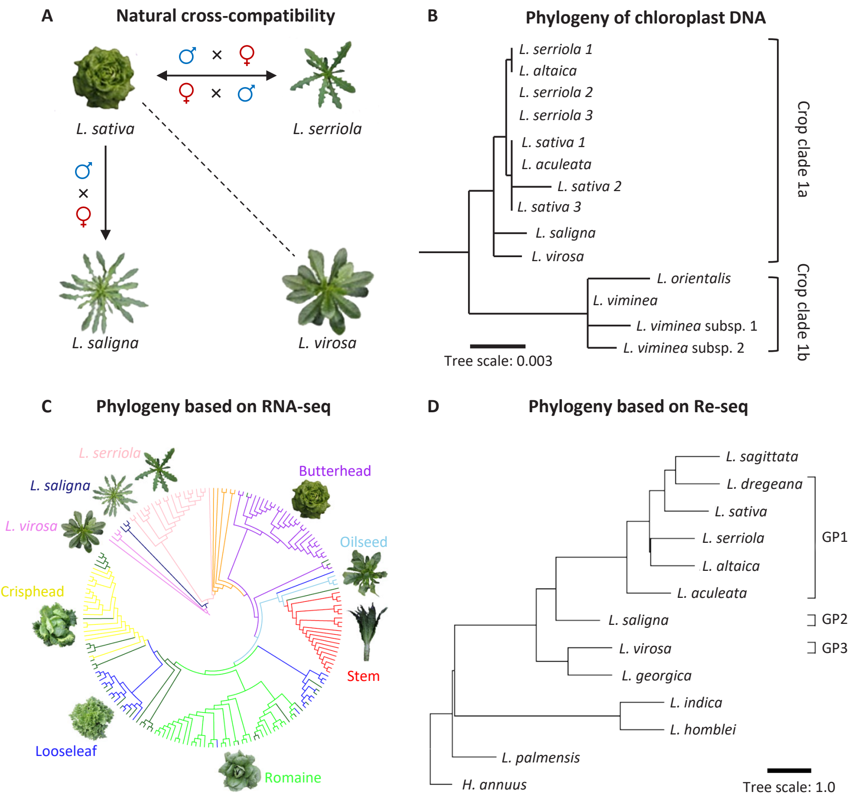


Figure 2. Summary of interspecific and evolutionary distances between lettuce and its close-relatives based on previous studies**.** **A**, cross-compatibility among *Lactuca* species. Arrows represent the pollination direction, and dashed line means natural barrier(s). **B**, Crop clade of *Lactuca* accessions from tree constructed by two chloroplast genes, adjusted from Wei *et al.* (2017). Numbers represent different accessions from the same species. **C**, phylogenetic tree built on SNPs derived from RNA-sequencing of 240 *Lactuca* accessions, adjusted from Zhang *et al.* (2017). Colors indicate different groups. **D**, phylogenetic tree built on SNPs derived from re-sequencing of 445 *Lactuca* a accessions, adjusted from Wei *et al.* (2021). GP1 to 3 correspond to the primary, second and tertiary gene pool mentioned in this introduction.

3.1 Interspecific hybridization among *Lactuca* species

Due to evolutionary distance, the introgression of genes from *L. saligna* and *L. virosa* into crop lettuce via sexual hybridization is challenging; but has had varying successes. For *L. saligna*, it is relatively easier to perform an interspecific cross with *L. sativa*. A successful hybridization requires *L. saligna* as the female parent, but still causes reduced fertility in the F_1 hybrid, which will be restored later in F_2 and F_3 offspring (Zohary, 1991). Based on such work, hybridizations were performed to foster the defense of lettuce in several studies. For example, *L. saligna* was crossed with and backcrossed into *L. sativa* for resistance to virus and oomycete diseases (Zdravkovic et al., 2001; Jeuken and Lindhout, 2004). For *L.*

virosa, interspecific hybridization with lettuce can be indirectly achieved using *L. serriola* as a bridge (Thompson and Ryder, 1961; Whitaker, 1974). This approach led to the development of parental lines for “Salinas” and “Vanguard” (Figure 3), which are the most important ancestors of modern crisphead cultivars and estimated to constitute 36% of genes in modern crisphead varieties (Mikel, 2013). Additionally, Maisonneuve also reported successful hybridization between *L. sativa* and *L. virosa* using embryo rescue and protoplast fusion technique (Maisonneuve et al., 1995; Maisonneuve, 2003). These hybridization studies were performed to exploit the genetic variation and enhance the agronomic traits and resistance of lettuce crops: for example, *L. virosa* contributes to a more robust root system and mitigated defoliation in crisphead cultivars (Mikel, 2007), and the *L. virosa* is a resource for resistance to lettuce downy mildew (*Bremia lactucae*), lettuce mosaic virus (LMV), beet western yellows virus (BWYV) and aphids (Maisonneuve, 2003).

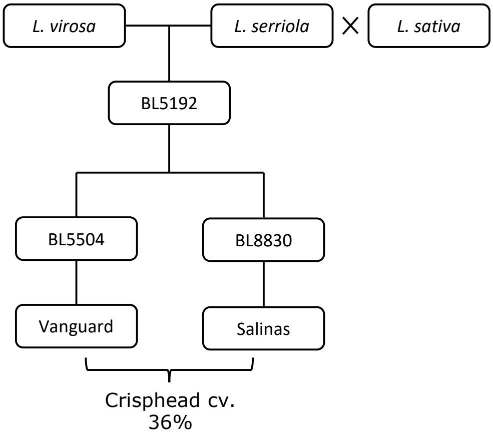


Figure 3. Simplified pedigree of line BL5192**.** *Lactuca serriola* was used as a bridge to enable the interspecific hybridization between lettuce (*L. sativa*) and its distant relative *L. virosa* (Thompson and Ryder, 1961). The resulting line BL5192 led the development of lines BL5504 and BL8830, which are parental lines of modern crisphead cultivar (cv.) Vanguard (24%) and Salinas (12%) contributing ~36% genetic content (Mikel, 2013).

3.2 Resistance is the most crucial improvement trait of modern lettuce breeding

In lettuce breeding projects, the major goal is to boost the immunity of cultivars by introducing novel resistance from lettuce’s germplasm. As outgroup species of the primary gene pool, *L. saligna* and *L. virosa* were both intensively studied to combat the economically significant disease or pests for lettuce production (Table 1). Among different biotic stresses, *L. saligna* and *L. virosa* are known for resistance to lettuce downy mildew and currant-lettuce aphid, respectively.

Table 1. Resistance to vital diseases or pests in *L. saligna* and *L. virosa**.

Disease (pathogen or pest)	Donors of resistance
Big-vein (*Mirafiori lettuce big-vein virus*)	*L. virosa*
Lettuce dieback (*Tomato bushy stunt virus* & *Lettuce necrotic stunt virus*)	*L. saligna, L. virosa*
Lettuce mosaic (*Lettuce mosaic virus*)	*L. saligna*, *L. virosa*
(*Tomato spotted wilt virus* and *Impatiens necrotic spot virus*)	*L. saligna*
Corky root (*Rhizomonas suberifaciens*)	*L. saligna*, *L. virosa*
Lettuce downy mildew (*Bremia lactucae*)	*L. saligna*, *L. virosa*
Lettuce powdery mildew (*Golovinomyces cichoracearum*)	*L. saligna*, *L. virosa*
Verticillium wilt (*Verticillium dahliae*)	*L. virosa*
Lettuce drop (*Sclerotinia minor* & *Sclerotinia sclerotiorum*)	*L. virosa*
Lettuce aphid (*Nasonovia ribisnigri*)	*L. virosa*
Leaf miners (*Liriomyza langei*, *L. trifolii* & *L. sativae*)	*L. saligna*, *L. virosa*

* Adjusted from Table 5.1 of the book section “Wild *Lactuca* Species in North America” (Lebeda et al., 2019).

3.2.1 Resistance of *Lactuca saligna* to lettuce downy mildew

Lettuce downy mildew is caused by the oomycete pathogen *Bremia lactucae* and is one of the most severe diseases for lettuce, causing massive global yield loss (Bonnier et al., 1991; Michelmore and Wong, 2008; Crute, 1992). The interaction between *B. lactucae* and lettuce is race-specific, ruled by classic gene-for-gene interactions (Crute and Johnson, 1976; Lebeda, 1984; Farrara and Michelmore, 1987; Lebeda et al., 2005). Hence, the common resistance breeding against *B. lactucae* relies heavily on introducing race-specific resistance (*R*) genes, also known as *Dm* genes, from diverse lettuce germplasm. However, resistant cultivars with Dm genes are rapidly overcome by the quick adaption of *Bremia*, which can be illustrated by the broken effectiveness of *Dm16* and *R18* genes in the 1990s (Lebeda and Zinkernagel, 2003). Thus, durable resistance to *B. lactucae* is an urgent demand of lettuce growers (Michelmore and Wong, 2008; Lebeda et al., 2014).

L. saligna has been considered a promising donor due to its crossing-ability with lettuce and complete resistance to *B. lactucae*. (Bonnier et al., 1991; Jeuken and Lindhout, 2002; Zhang et al., 2009a). For *L. saligna*, such broad resistance to *B. lactucae* resembles the features of a non-host species: “All genotypes of a species are resistant against all genotypes of a specific pathogen” (Heath, 1981), which is also known as non-host resistance (NHR). Though most *L. saligna* accessions presented NHR to *Bremia* races at different developmental stages, some accessions still showed sporulation under ideal conditions (Bonnier et al., 1991; Lebeda and Reinink, 1994). *Lactuca saligna* should therefore be classified as an intermediate host for *Bremia* (Niks, 1987, 1988), and its NHR is strictly phenotypic, potentially comprising different molecular mechanisms or factors (Panstruga and Moscou, 2020). Histologically, *L. saligna* demonstrated a distinct defense pattern after *Bremia* penetration compared to resistant *L. sativa* or natural non-host species (Figure 4). Unlike other incompatible cultivars, in *L. saligna*, the growth of *Bremia* invasion is stopped with a malformed-hypha before haustorium-formation, (Lebeda and Reinink, 1994; Jeuken and Lindhout, 2002; Zhang et al., 2009b). Genetic studies of *L. saligna* have shown a complicated system underlying its resistance to *B. lactucae*, where an F_2 population derived from a cross between a resistant *L. saligna* and susceptible *L. sativa* shows both quantitative and race-specific resistance (Lebeda and Reinink, 1994; Jeuken and Lindhout, 2002). Furthermore, a later study by Giesbers et al. (2017) confirmed that the *R* gene is dispensable for NHR in *L. saligna*. Thus, the mechanism of NHR in *L. saligna* remains unknown.

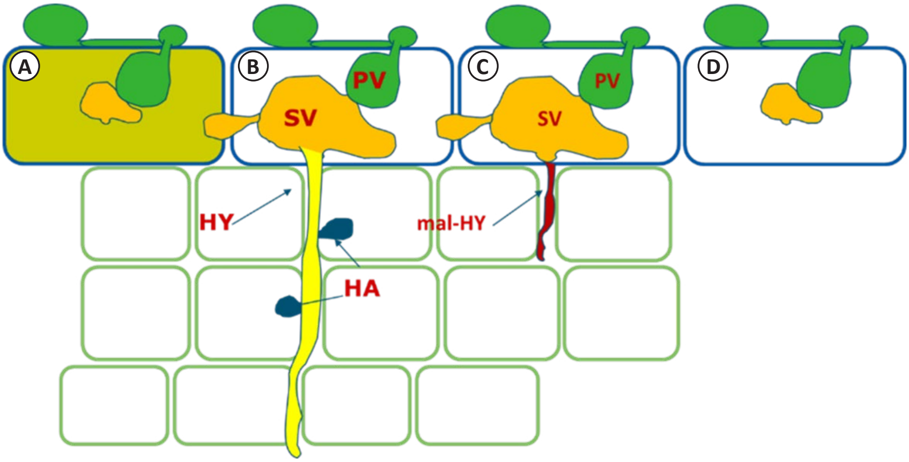


Figure 4. Histology of lettuce downy mildew infection in different hosts. There are four different plant-pathogen interactions of lettuce downy mildew infection: **A**, The incompatible reaction of host *L. sativa* with R genes. **B**, the compatible reaction of host *L. sativa*. **C**, Incompatible reaction host *L. saligna* with NHR. **D**, The incompatible reaction of non-host species. Adapted from Figure 1 page 4, PhD thesis Zhang, N. (2008). Genetic dissection of non-host resistance of wild lettuce, *Lactuca saligna*, to downy mildew. (https://edepot.wur.nl/10917). PV, primary vesicle; SV, secondary vesicle; HY, hyphae; HA, haustorium; mal-HY, malformed hyphae.

3.2.2 Lettuce aphid resistance of *Lactuca virosa*

Apart from the *B. lactucae* pathogen, the currant-lettuce aphid (*Nasonovia ribisnigri*) pest can also cause considerable economic loss of lettuce production. The colonization of aphids usually begins with the young leaves of lettuce head (Liu, 2004; McCreight, 2008), and later spreads to the frame leaves as the population grows (Liu, 2004). Ultimately, the large number of aphids will damage the leaves and seedlings. Indirectly, even the presence of a few living aphids can sabotage the outward lettuce appearance and make it unmarketable. Moreover, *N. ribisnigri* can transmit viruses to lettuce (Subbarao et al., 2017). Different approaches can be applied to protect the lettuce plants from the lettuce aphid, including chemical control, biological control and cultural practice (ten Broeke, 2013). Nevertheless, the most economical and sustainable is believed to be the host plant resistance.

Since 1982, *L. sativa* has employed host plant resistance to control the *N. ribisnigri* aphid species, which is mediated by dominant *Nr*-gene originating from the wild lettuce species *L. virosa* (Eenink et al., 1982; Eenink and Dieleman, 1983). This gene provided complete resistance and was used in many modern cultivars by breeding companies (van der Arend, 2003). However, the reported rise in virulent *N. ribisnigri* aphids since 2007 has rendered this resistance ineffective (Thabuis et al., 2011). This virulence may be caused by an effector protein located in the salivary secretion of the aphids, which suppresses the resistance of the lettuce towards the avirulent aphids. Behavioral studies have been done to unravel the resistance mechanism mediated by the *Nr*-gene against *N. ribisnigri* (ten Broeke et al., 2017). The resistance primarily expresses in the phloem, and aphids may encounter unknown deter-compound(s) in phloem sap as they traverse the pathway. The resistance factor(s) are only produced in the shoot, and the intact vascular system is required for complete resistance. The virulence of aphids on resistant or susceptible varieties from different population, associated with a fitness cost (ten Broeke, 2013). In a more recent study from Walley et al. (2017), a lettuce diversity panel was combined with genetic markers to identify the novel resistance factors against the currant-lettuce aphid. Several single-nucleotide polymorphisms (SNPs) were found to be significantly associated with the resistance with diversifying function, for example, the LS1_51 and LS1_729 markers whose homolog in *Arabidopsis* encode peptidase and receptor-like kinase proteins, respectively (Walley et al., 2017). Although, this study revealed the loci of resistance to *N. ribisnigri* and its polygenic nature. Likely, the identified SNPs are indirectly associated with the nearby causal genes, and the exact mechanism of *Nr*-gene needs to be explored soon. Consequently, unraveling of the resistance mechanism(s) or allele(s) will quench the demand of resistant cultivars.

With the development of sequencing, researchers can thoroughly study and compare the genetic diversity within or between species. Producing assemblies of *L. saligna* and *L. virosa* can reveal their genetic architecture and difference compared to lettuce. Subsequently, the gained knowledge will boost the biological research of the mechanism underlying crossing barrier and resistance. Moreover, it can also benefit the study of pedigree history for lettuce breeding, for example, by tracing the genetic content that introgressed from *L. virosa* to *L. sativa*.

4. Dandelion: an important outgroup of the *Lactuca* genus

4.1 Genome editing extends the gene pool scope for modern lettuce breeding

Plant breeding is traditionally very time-consuming, especially if it involves introgression from wild species. Nowadays, marker-assisted selection makes the introgression hybridization of a trait more precise and efficient. However, breeders are still wrestling with additional challenges (e.g., linkage-drag and off-target sites). Moreover, only the variation in crossable relatives could be accessed. Over the last ten years, developments of new breeding techniques represented by genome editing (GEd) facilitate precise and site-directed modifications in the genome of many plant species (Modrzejewski et al., 2019). This advancement enables more predictable breeding and hopefully rapid crop adaption. GEd uses two types of variants generated by site-directed nucleases (SDNs) or oligo-nucleotide-directed mutagenesis (Sprink et al., 2022). To date, the SDNs type technology based on CRISPR-system is the most trending approach and was applied in many reported studies, especially for the CRISPR-Cas9 system (Huang and Puchta, 2021).

4.2 Applications of dandelion in lettuce breeding and trait evolution study

For lettuce breeding, the study of apomixis in dandelion is an outstanding example of how GEd can facilitate the study of traits of interest and extend the common gene pool to more distant species. Apomixis, or reproduction via clonal seed, holds great potential to revolutionize the plant breeding industry as a tool to produce and fix breeding lines (Nogler, 1984; Ozias-Akins and Van Dijk, 2007). This trait is rare but convergently present across flowering plants (Mogie, 1992). One of the most well-known and widespread apomicts is the common dandelion, *Taraxacum officinale*, which relies heavily on apomixis for its success in ecology and evolution (Van Dijk, 2003). *Taraxacum officinale* has sexual type (2n=2x=16) and asexual (i.e., apomictic) type (2n=3x=24; Tas and Van Dijk, 1999). The apomixis in dandelion is gametophytic-type because of its sexual-like female gametophytes. In contrast to sexual type, the apomixis trait of dandelion is regulated by two separate loci responsible for diplospory (i.e., cell division without recombination and reduction) and parthenogenesis (i.e., embryo development without fertilization) during embryogenesis (Tas and Van Dijk, 1999; Van Dijk et al., 1999, 2020). Recently, the *PARTHENOGENESIS* (*PAR*) gene has been cloned from the apomictic dandelion (Underwood et al., 2022), and of which I am a co-author on the paper (but is not included in this thesis). CRISPR-Cas9 was applied in this study to validate the mechanism of its novel regulation and later enable the heterologous expression in lettuce, showing the potential of *PAR* in lettuce breeding. Besides novel variation, *T. officinale* is a phylogenetically close relative to the *Lactuca* genus (both in the Cichorioideae subtribe). Therefore, it can serve as an outgroup for the trait evolution study among *Lactuca* species, such as the loss of seed shattering and bitter flavor in domesticated lettuce (Sessa et al., 2000; Wei et al., 2021). Before, only contigs of PAR locus were assembled for sexual and asexual dandelions with a focus on apomixis (Underwood et al., 2022). In the future, a complete genomic map is required to gain the comprehensive regulatory network of *PAR* and facilitate the phylogenetic studies of genes related to other essential traits. Because of the complex of triploid, the sexual diploid type is a better sequencing target for the genome construction of *T. officinale*.

5. Asteraceae: a model family for evolutionary biology

5.1 A brief overview of the remarkable diversity in Asteraceae

Asteraceae, also known as Compositae, is one of the largest families of more than 25,000 species, nearly 10% of all extant flowering plants (Funk et al., 2005; Anderberg et al., 2007; Mandel et al., 2019). It has 16 subfamilies, including two large groups: Asteroideae (e.g., sunflower) and Cichorioideae (e.g., lettuce and dandelion) according to current classification (Mandel et al., 2019; Vijverberg et al., 2021). There are about 1,627 genera, and many of them rank top (> 500 species) in angiosperm (Christenhusz et al., 2017). For example, the *Taraxacum* is the third largest genus (Frodin, 2004). Morphologically, Asteraceae are also enormously diverse, of which the most iconic trait is the flower and its capitulate inflorescence with varying forms and sizes illustrated by Figure 5 (Elomaa *et al.*, 2018; Mandel *et al.*, 2019). The floral trait is believed to be one of the main factors contributing to its evolutionary success (Panero and Funk, 2008), which remains widely uncharted. Asteraceae are also globally widespread and are habitat in all continents at an entire range of altitudes for land plants (Mandel et al., 2019). These species from different environments make ideal models for advancing our understanding of ecological biology.

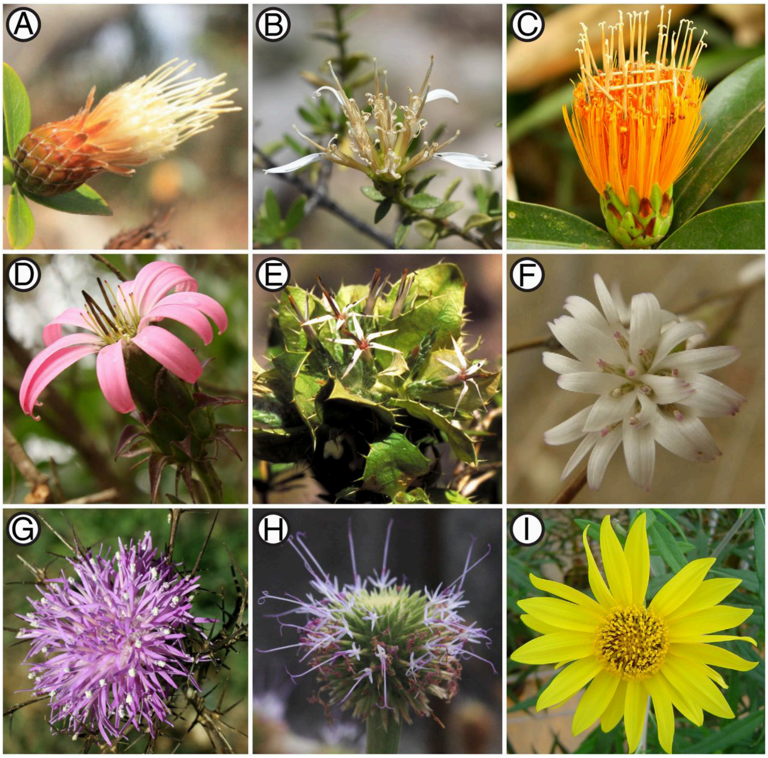


Figure 5. Photos of nine tribes showing the floral diversity in Asteraceae family. **A**, Barnadesieae; **B**, Famatinantheae; **C**, Stifftieae; **D**, Mutisieae; **E**, Hecastocleideae; F, Pertyeae; **G**, Cardueae; **H**, Vernonieae; and **I**, Heliantheae. Photos of A, C, D, and H were provided by C.M.S.; B, image courtesy of J. Mauricio Bonifacino (photographer); E,by V.A.F.; F, image courtesy of Tiangang Gao (photographer); G, image courtesy of Alfonso Susanna (photographer); and I, by J.R.M. This figure is retrieved from Figure 1 of page 2 from Mandel *et al.* (2019).

5.2 Polyploidization and trait evolution of Asteraceae origin and radiation

Asteraceae originated ~83 MYA during the late Cretaceous. It went through a streak of species explosions in the middle to late Eocene, which generated ~95% of existing species (Mandel et al., 2019). Most of these radiation events arose after mass extinction periods when an ecological gap was ready to be occupied (reviewed in Palazzesi *et al.*, 2022). In addition, paleopolyploidy events were also recognized near the origin and divergence of Asteraceae, and the stem nodes of major tribes preceding its rapid radiation (Barker et al., 2008, 2016; Huang et al., 2016). Besides, lineage-specific whole-genome duplication (WGD) events are also found in many sub-groups (Huang et al., 2016; Shen et al., 2021). Subsequently, polyploidy events cause extremely high genomic diversity among Asteraceae species from genome size, polyploidy level and genome organization. These features make them prime models for biological studies of genome evolution and ancient polyploidy.

Plants commonly tolerate more gene dosage imbalances and have shorter regulatory (promoter) regions than animal genes, allowing for greater genome plasticity after WGDs (Pandey et al., 2002; Loidl, 2004). Subsequently, polyploidization plays a critical role in the rapid diversification of Asteraceae via sub- or neofunctionalization on duplicated genes of important traits, like flower morphology (Barker et al., 2008). These retained duplicates derived from ancient polyploidy events are predominantly fundamental for structural components or cellular organization or critical for regulation and development (Barker et al., 2008). For example, the gene family of CYC transcription factors has experienced a significant expansion and positive selection toward specialization in Asteraceae species compared to Arabidopsis, which is involved in the regulation of capitula development (Elomaa et al., 2018; Chen et al., 2018). Studies of other gene retention and evolution of Asteraceae species will further help to comprehend the relationship between trait and diversification.

5.3 Incorporating synteny into phylogenomic analysis of Asteraceae

To assess the effect of gene retention in Asteraceae, sequence-similarity-based clustering combined with phylogenetic analyses can be powerful to distinguish the paralogous homolog (i.e., via duplication) from the orthologous homolog (i.e., via speciation). However, this is limited to slow-evolving conserved genes but cannot solve fast-evolving genes (Natsidis et al., 2021). Hence, the interest in adding synteny to phylogenomic analyses to (re)solve orthology relationships is growing. Synteny represents the conserved chromosome regions between genomes where homologous genes are in a shared order from their common ancestor (Tang et al., 2008). Reportedly, synteny can therefore reliably determine orthologs and paralogs (Liu et al., 2018). Moreover, it can hint at genes with regulatory novelty via the identification of genomic context change between Asteraceae and other flowering plants caused by the different chromatin architectures (e.g., chromatin loop) across plant species (Kadauke and Blobel, 2009; Dong et al., 2017). Thus, synteny is a valuable addition to phylogeny for trait evolution study. To summarize, genome collinearity (i.e., synteny) holds great promise for understanding the evolutionary history of genes and genomes and, ultimately, traits in Asteraceae and species across broad phylogenetic groups and divergence times.

To date, synteny is largely unexplored for the Asteraceae family. Reported studies are mainly performed via pair-wise comparisons, for example, Timms *et al.* (2006) compared *Arabidopsis thaliana* and the Asteraceae crops, lettuce and sunflower. However, such a small-scale comparison has limited power to represent the whole Asteraceae for studies of family-based trait evolution. The lack of family-wise synteny hinders our understanding of Asteraceae traits and genome evolution. In order to perform a broad range of synteny analysis, the appropriate tools and quality genome assemblies with high continuity are needed: i) There are available pipelines that can perform large-scale synteny analysis.

For example, since 2017, the synteny network analysis (SynNet) pipeline was developed to transcend the border of the previously limited number of genomes in comparative genomics and enables a visualized network for vast syntenic relationships thereafter (Zhao and Schranz, 2017; Gamboa-Tuz et al., 2022). SynNet facilitated the identification of lineage-specific syntenies of *MADS-box* genes in Brassicaceae (Zhao et al., 2017), which is possibly also applicable to Asteraceae. ii) Nowadays, genomes of 39 Asteraceae species from 18 genera are assembled (Palazzesi et al., 2022), at different assembling levels (i.e., from the chromosome level to scaffold level). Back in the 2015s, there were a few available Asteraceae assemblies. Most were fragmented for species with relatively small genome sizes, such as *Conyza canadensis* (335Mbp; Peng *et al.*, 2014) and *Cynara cardunculus* (1,084Mbp; Scaglione *et al.*, 2016). Only a handful of genomes for economic crops were at the chromosome-level; for example, the lettuce (2.5Gbp; Reyes-Chin-Wo *et al.*, 2017) and sunflower (3.6Gbp; Badouin *et al.*, 2017) assemblies published in 2017. Thus, more genome assemblies are required to establish the cornerstone for phylogenomic analysis to study the evolutionary biology of Asteraceae as a family unit.

6. Ph.D. project

6.1 Thesis aim

Lettuce breeding faces many challenges for which wild relatives provide essential genetic variation for crop improvement. My PhD project aims to deliver annotated genome assemblies of *Lactuca saligna* and *L. virosa*. These two *Lactuca* species will complement the references of *L. sativa* and *L. serriola sequenced* by UC-Davis (https://lgr.genomecenter.ucdavis.edu/Home.php), and collectively covers the spectrum of currently employed germplasm for lettuce crop breeding. Using mentioned assemblies, downstream analyses were performed to achieve the following goals for fundamental research and practical breeding:

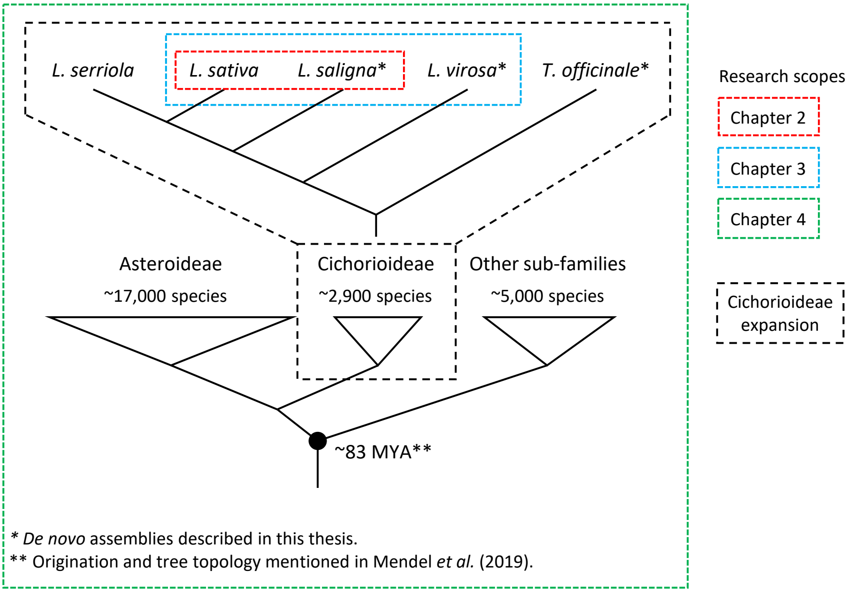
- Depict structure and evolution of genomes between *Lactuca* species. Differences in genomic collinearity directly influence chromosome pairing during meiosis. Therefore, newly gained perspectives of genome structure in this project are valuable for hybridization introgression breeding in lettuce.
- Identify sequence diversity in genes underlying important traits (e.g., resistance) and disclose their genomic context. This knowledge can provide insights into genetic mechanisms that underlie target traits, and identification of genes or alleles will benefit the precision breeding in lettuce.
- Assess sequencing diversity between the accessions of wild lettuce species based on reconstructed genome reference assemblies. My study can help elucidate the genetic variation and population structure, of which shed lights

on evolution and pedigree history. Furthermore, such exploration within wild germplasm can guide material selections to succeed the hybridization and introgression.

In addition to the wild species of crop lettuce, the *de novo* genome assembly of a sexual diploid common dandelion (*T. officinale*) was also constructed. Combined with *Lactuca* genomes, they jointly represent the subfamily Cichorioideae, and enable the study of floral trait evolution for the whole Asteraceae family. Apart from biological research, the strategy for genome reconstruction is also aimed to be evaluated in this thesis. The gained knowledge can contribute to future genome assembling or breeding programs.

6.2 Thesis outline

As summarized in Figure 6, I followed an evolutionary path from the *Lactuca* genus, the Cichorioideae sub-family, finally to the Asteraceae family, and performed scaling-up comparative genomic and phylogenomic analyses on different traits, using three *de novo* genome assemblies, namely, *L. saligna* (**Chapter 2**), *L. virosa* (**Chapter 3**), and *T. officinale* (**Chapter 4**).



* *De novo* assemblies described in this thesis.
** Origination and tree topology mentioned in Mendel *et al.* (2019).

Figure 6. Research scheme contextualized in a collapsed tree of Asteraceae. Studying trait evolution with expanding scopes: from *Lactuca* genus to Asteraceae family. Triangles size are based on the number of species in collapsed groups.

Chapter 2 presents the *de novo* genome assembly of *Lactuca saligna* setting the stage for an in-depth genomic comparison with *L. sativa* and an analysis of non-host resistance to downy mildew *Bremia lactucae*. Based on this *de novo* genome assembly, I coupled the population structure of 15 resequenced *L. saligna* accessions across Eurasia with their geographical location. Genome-wide synteny detection between *L. saligna* and *L. sativa* to reveal the structural variant. A comprehensive inventory for immune genes encoding nucleotide-leucine rich repeat and receptor-like kinase proteins is provided. I pinpointed the hot spots of aforementioned genes and associated them with non-host resistance intervals identified by Giesbers *et al*. (2019). A transcriptomic analysis for *Bremia* infection bioassay discovers a list of immune genes on mentioned resistance regions.

Another representative wild lettuce is *L. virosa*, which has been used in lettuce breeding and contributed genetic content to modern crisphead cultivars. In **Chapter 3**, a *de novo* genome assembly of *L. virosa* is used in three-way comparative genomics analyses among *Lactuca* species to study the genome and gene evolution within genus. Homolog grouping of *L. sativa*, *L. saligna* and *L. virosa* distinguishes the lineage-specific genes from core genes and enables structural variation detection via synteny. Combining assembly annotation, a comparative repeatomics analysis between three *Lactuca* species shows the type of transposable elements driving the genome expansion of *L. virosa*. A similar search for immune genes mentioned in Chapter 2 is described, of which further comparison demonstrates their number difference and evolution history.

The *Lactuca* genus and a representative outgroup (*Taraxacum*) both belong to the Cichoroideae sub-family of the Asteraceae, which is the most successful flowering family. For Asteraceae, the specialization of capitulum and floret are its most distinctive characteristics. In **Chapter 4**, I conducted a broad range of comparative genomic studies of floral gene evolution in the Asteraceae by analyzing 33 plant genomes. A *de novo* assembly of *T. officinale* was included in the comparative genomic analysis together with the two *Lactuca* assemblies mentioned in Chapter 2- Chapter 4. The phylogenomic analysis of *MADS-box* and *TCP* gene families among the selected species is described to anticipate the effect of gene duplication and transposition on flower trait evolution. Transcriptomic data proved the expression of identified *MADS-box* and *TCP* with a lineage-specific genomic context or from an Asteraceae-dominant clade, in different materials and stages during flower development.

Finally, in **Chapter 5**, my research chapters are integrated and analyzed from different angles to forge a final synthesis of the following topics: genome sequencing, model species, trait evolution, gene duplication, and a reflection on genome sequencing projects. Findings of genome reconstructions and genetic diversity for traits are explicitly demonstrated in each chapter, and therefore are discussed with emphasis: 1) While sequencing of different species are described individually, the assessment of strategies is only feasible in this chapter and discussed in the context of current techniques development and contemporary projects. 2) For resistance and floral traits, I will briefly demonstrate some unpresented but highly related results in comparison to ongoing research. Moreover, speculations of potential elements or mechanisms are made to give suggestions for further studies on NHR in *Lactuca* and floral regulation in Asteraceae supported by revising the current dataset. As for the remaining three topics, I focus on my entire thesis. I discuss their importance in this chapter as well: 3) *Lactuca*, and *Taraxacum* (Asteraceae) can all be useful models for different scenarios, in Chapter 5 the exploitation of these plants will be discussed and compared to current and future research. 4) Gene duplication is a major driver for gene evolution, the scattered pieces about gene duplication are now systematically discussed, and finally, 5) an overview reflection moment is taken on this PhD thesis aiming a legacy for future sequencing projects.

7. References

Anderberg, A.A. et al. (2007). Compositae. In: Kubitzki K (ed) The families and genera of vascular plants. VIII. Flowering plants, eudicots, Asterales J.W. Kadereit and C. Jeffrey, eds (Springer: Berlin-Heidelberg-New York).

van der Arend, A.J.M. (2003). The possibility of *Nasonovia ribisnigri* resistance breaking biotype development due to plant host resistance: a literature study. Eucarpia leafy Veg.: 75–81.

Auton, A. et al. (2015). A global reference for human genetic variation. Nature 526: 68–74.

Badouin, H. et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature 546: 148–152.

Barker, M.S., Kane, N.C., Matvienko, M., Kozik, A., Michelmore, R.W., Knapp, S.J., and Rieseberg, L.H. (2008). Multiple paleopolyploidizations during the evolution of the compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol. Biol. Evol. 25: 2445–2455.

Barker, M.S., Li, Z., Kidder, T.I., Reardon, C.R., Lai, Z., Oliveira, L.O., Scascitelli, M., and Rieseberg, L.H. (2016). Most compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the calyceraceae. Am. J. Bot. 103: 1203–1211.

Bonnier, F.J.M., Reinink, K., and Groenwold, R. (1991). New sources of major gene resistance in *Lactuca* to *Bremia lactucae*. Euphytica 61: 203–211.

Boukema, I., Hazekamp, T., and Hintum, T. van (1990). The CGN lettuce collection. Cent. Genet. Resour.

ten Broeke, C.J.M. (2013). Unravelling the resistance mechanism of lettuce against *Nasonovia ribisnigri* (Wageningen University and Research).

ten Broeke, C.J.M., Dicke, M., and van Loon, J.J.A. (2017). The effect of co-infestation by conspecific and heterospecific aphids on the feeding behaviour of *Nasonovia ribisnigri* on resistant and susceptible lettuce cultivars. Arthropod. Plant. Interact. 11: 785–796.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. 31: 1119–1125.

Cao, H. et al. (2014). Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. Gigascience 3: 34.

Chen, J., Shen, C.-Z., Guo, Y.-P., and Rao, G.-Y. (2018). Patterning the Asteraceae Capitulum: Duplications and Differential Expression of the Flower Symmetry *CYC2*-Like Genes. Front. Plant Sci. 9: 551.

Christenhusz, M.J.M., Fay, M.F., and Chase, M.W. (2017). Plants of the world. In Plants of the World (University of Chicago Press).

Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. Nat. Nanotechnol. 4: 265–270.

Craig Venter, J. et al. (2001). The sequence of the human genome. Science 291: 1304–1351.

Crute, I.R. (1992). The role of resistance breeding in the integrated control of downy mildew (Bremia lactucae) in protected lettuce. In (Springer, Dordrecht), pp. 95–102.

Crute, I.R. and Johnson, A.G. (1976). The genetic relationship between races of *Bremiae lactucae* and cultivars of *Lactuca sativa*. Ann. Appl. Biol. 83: 125–137.

Van Dijk, P.J. (2003). Ecological and evolutionary opportunities of apomixis: insights from *Taraxacum* and *Chondrilla*. Philos. Trans. R. Soc. London B Biol. Sci. 358: 1113–1121.

Van Dijk, P.J., Op den Camp, R., and Schauer, S.E. (2020). Genetic dissection of apomixis in dandelions identifies a dominant parthenogenesis locus and highlights the complexity of autonomous endosperm formation. Genes (Basel). 11: 961.

Van Dijk, P.J., Tas, I.C.Q., Falque, M., and Bakx-Schotman, T. (1999). Crosses between sexual and apomictic dandelions (Taraxacum). II. The breakdown of apomixis. Heredity (Edinb). 83: 715–721.

Doležalová, I., Lebeda, A., Janeček, J., Číhalíková, J., Křístková, E., and Vránová, O. (2002). Variation in chromosome numbers and nuclear DNA contents in genetic resources of *Lactuca* L. species (Asteraceae). Genet. Resour. Crop Evol. 49: 383–395.

Dominguez Del Angel, V. et al. (2018). Ten steps to get started in Genome Assembly and Annotation. F1000Research 7: 148.

Dong, P., Tu, X., Chu, P.Y., Lü, P., Zhu, N., Grierson, D., Du, B., Li, P., and Zhong, S. (2017). 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments. Mol. Plant 10: 1497–1509.

Eenink, A.H. and Dieleman, F.L. (1983). Inheritance of resistance to the leaf aphid *Nasonovia ribis-nigri* in the wild lettuce species *Lactuca virosa*. Euphytica 32: 691–695.

Eenink, A.H., Groenwold, R., and Dieleman, F.L. (1982). Resistance of lettuce (*Lactuca*) to the leaf aphid *Nasonovia ribis nigri*. 1. Transfer of resistance from L. virosa to L. sativa by interspecific crosses and selection of resistant breeding lines. Euphytica 31: 291–299.

Eid, J. et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science (80-.). 323: 133–138.

Elomaa, P., Zhao, Y., and Zhang, T. (2018). Flower heads in Asteraceae—recruitment of conserved developmental regulators to control the flower-like inflorescence architecture. Hortic. Res. 5: 36.

FAOSTAT (2020). FAOSTAT. Food Agric. Organ. United Nations.

Farrara, B. and Michelmore, R. (1987). Identification of new sources of resistance to downy mildew in *Lactuca* spp. HortScience 22: 647–649.

Frodin, D.G. (2004). History and concepts of big plant genera. Taxon 53: 753–776.

Funk, V.A., Watson, L., Gemeinholzer, B., Schilling, E., Susanna, A., and Jansen, R.K. (2005). Everywhere but Antarctica : Using a supertree to understand the diversity and distribution of the Compositae.

Gamboa-Tuz, S.D., Pereira-Santana, A., Zhao, T., and Schranz, M.E. (2022). Applying synteny networks (SynNet) to study genomic arrangements of protein-coding genes in plants. In Methods in Molecular Biology (Humana Press Inc.), pp. 199–215.

Giesbers, A.K.J., Pelgrom, A.J.E., Visser, R.G.F., Niks, R.E., Van den Ackerveken, G., and Jeuken, M.J.W. (2017). Effector-mediated discovery of a novel resistance gene against *Bremia lactucae* in a nonhost lettuce species. New Phytol. 216: 915–926.

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M.C., and McCombie, W.R. (2015). Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. Genome Res. 25: 1750–1756.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: Ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17: 333–351.

Heath, M.C. (1981). Nonhost resistance. Plant Dis. Control: 201–217.

Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: An overview. Hum. Immunol. 82: 801–811.

Huang, C.H., Zhang, C., Liu, M., Hu, Y., Gao, T., Qi, J., and Ma, H. (2016). Multiple Polyploidization Events across Asteraceae with Two Nested Events in the Early History Revealed by Nuclear Phylogenomics. Mol. Biol. Evol. 33: 2820–2835.

Huang, T.K. and Puchta, H. (2021). Novel CRISPR/Cas applications in plants: from prime editing to chromosome engineering. Transgenic Res. 30: 529–549.

Jeuken, M. and Lindhout, P. (2002). *Lactuca saligna* , a non-host for lettuce downy mildew (Bremia lactucae), harbors a new race-specific Dm gene and three QTLs for resistance. TAG Theor. Appl. Genet. 105: 384–391.

Jeuken, M.J.W. and Lindhout, P. (2004). The development of lettuce backcross inbred lines (BILs) for exploitation of the *Lactuca saligna* (wild lettuce) germplasm. Theor. Appl. Genet. 109: 394–401.

Jorasch, P. (2019). The global need for plant breeding innovation. Transgenic Res. 28: 81–86.

Kadauke, S. and Blobel, G.A. (2009). Chromatin loops in gene regulation. Biochim. Biophys. Acta - Gene Regul. Mech. 1789: 17–25.

Kaul, S. et al. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.

Koopman, W.J.M., Zevenbergen, M.J., and Van den Berg, R.G. (2001). Species relationships in *Lactuca* s.l. (Lactuceae, Asteraceae) inferred from AFLP fingerprints. Am. J. Bot. 88: 1881–1887.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., and Phillippy, A.M. (2012). Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. Nat. Biotechnol. 30: 693–700.

Křístková, E., Doležalová, I., Lebeda, A., Vinter, V., and Novotná, A. (2008). Description of morphological characters of lettuce (*Lactuca sativa* L.) genetic resources. Hort. Sci. 35: 113–129.

Lander, E.S. et al. (2001). Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., and Studholme, D.J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol. Detect. Quantif. 3: 1–8.

Lebeda, A. (1984). Race-specific factors of resistance to *Bremia lactucae* in the world assortment of lettuce. Sci. Hortic. (Amsterdam). 22: 23–32.

Lebeda, A., Dolezalová, I., Feráková, V., and Astley, D. (2004). Geographical distribution of wild *Lactuca* species (Asteraceae, Lactuceae). Bot. Rev. 70: 328.

Lebeda, A., Doležalová, I., Křístková, E., Kitner, M., Petrželová, I., Mieslerová, B., and Novotná, A. (2009). Wild *Lactuca* germplasm for lettuce breeding: Current status, gaps and challenges. Euphytica 170: 15–34.

Lebeda, A., Křístková, E., Doležalová, I., Kitner, M., and Widrlechner, M.P. (2019). Wild *Lactuca* species in North America. In North American Crop Wild Relatives, Volume 2, pp. 131–194.

Lebeda, A., Křístková, E., Kitner, M., Mieslerová, B., Jemelková, M., and Pink, D.A.C. (2014). Wild *Lactuca* species, their genetic diversity, resistance to diseases and pests, and exploitation in lettuce breeding. Eur. J. Plant Pathol. 138: 597–640.

Lebeda, A., Pink, D.A.C., and Astley, D. (2005). Aspects of the Interactions between Wild *Lactuca* Spp. and Related Genera and Lettuce Downy Mildew (*Bremia Lactucae*). In Advances in Downy Mildew Research, pp. 85–117.

Lebeda, A. and Reinink, K. (1994). Histological characterization of resistance in *Lactuca saligna* to lettuce downy mildew (*Bremiae lactucae*). Physiol. Mol. Plant Pathol. 44: 125–139.

Lebeda, A. and Zinkernagel, V. (2003). Evolution and distribution of virulence in the German population of Bremia lactucae. Plant Pathol. 52: 41–51.

Leebens-Mack, J.H. et al. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574: 679–685.

Lindqvist, K. (1960a). Cytogenetic studies in the serriola group of *Lactuca*. Hereditas 46: 75–151.

Lindqvist, K. (1960b). On the origin of cultivated lettuce. Hereditas 46: 319–350.

Liu, D., Hunt, M., and Tsai, I.J. (2018). Inferring synteny between genome assemblies: A systematic evaluation. BMC Bioinformatics 19: 26.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012: 251364.

Liu, Y.-B. (2004). Distribution and population development of *Nasonovia ribisnigri* (Homoptera: Aphididae) in Iceberg Lettuce. J. Econ. Entomol. 97: 883–890.

Lu, H., Giordano, F., and Ning, Z. (2016). Oxford anopore MinION sequencing and genome assembly. Genomics, Proteomics Bioinforma. 14: 265–279.

Maisonneuve, B. (2003). *Lactuca virosa*, a source of disease resistance genes for lettuce breeding: results and difficulties for gene introgression.

Maisonneuve, B., Chupeau, M.C., Bellec, Y., and Chupeau, Y. (1995). Sexual and somatic hybridization in the genus *Lactuca*. Euphytica 85: 281–285.

Mandel, J.R., Dikow, R.B., Siniscalchi, C.M., Thapa, R., Watson, L.E., and Funk, V.A. (2019). A fully resolved backbone phylogeny reveals numerous dispersals and explosive diversifications throughout the history of Asteraceae. Proc. Natl. Acad. Sci. U. S. A. 116: 14083–14088.

McCreight, J.D. (2008). Potential sources of genetic resistance in *Lactuca* spp. to the lettuce aphid, *Nasanovia ribisnigri* (Mosely) (Homoptera: Aphididae). HortScience 43: 1355–1358.

Mehrotra, S. and Goyal, V. (2014). Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. Genomics, Proteomics Bioinforma. 12: 164–171.

Michelmore, R. and Wong, J. (2008). Classical and molecular genetics of *Bremia lactucae*, cause of lettuce downy mildew. Eur. J. Plant Pathol. 122: 19–30.

Mikel, M.A. (2007). Genealogy of contemporary North American lettuce. HortScience 42: 489–493.

Mikel, M.A. (2013). Genetic composition of contemporary proprietary U.S. lettuce (*Lactuca sativa* L.) cultivars. Genet. Resour. Crop Evol. 60: 89–96.

Modrzejewski, D., Hartung, F., Sprink, T., Krause, D., Kohl, C., and Wilhelm, R. (2019). What is the available evidence for the range of applications of genome-editing as a new tool for plant trait modification and the potential occurrence of associated off-target effects: A systematic map. Environ. Evid. 8: 1–33.

Mogie, M. (1992). The evolution of asexual reproduction in plants (London: Chapman & Hall).

Natsidis, P., Kapli, P., Schiffer, P.H., and Telford, M.J. (2021). Systematic errors in orthology inference and their effects on evolutionary analyses. iScience 24: 102110.

Niks, R.E. (1988). Nonhost plant species as donors for resistance to pathogens with narrow host range. II. Concepts and evidence on the genetic basis of nonhost resistance. Euphytica 37: 89–99.

Niks, R.E. (1987). Nonhost plant species as donors for resistance to pathogens with narrow host range I. Determination of nonhost status. Euphytica 36: 841–852.

Nogler, G.A. (1984). Gametophytic Apomixis. In Embryology of Angiosperms (Springer Berlin Heidelberg), pp. 475–518.

Ozias-Akins, P. and Van Dijk, P.J. (2007). Mendelian genetics of apomixis in plants. Annu. Rev. Genet. 41: 509–537.

Palazzesi, L., Pellicer, J., Barreda, V.D., Loeuille, B., Mandel, J.R., Pokorny, L., Siniscalchi, C.M., Tellería, M.C., Leitch, I.J., and Hidalgo, O. (2022). Asteraceae as a model system for evolutionary studies: from fossils to genomes. Bot. J. Linn. Soc. 200: 143–164.

Panero, J.L. and Funk, V.A. (2008). The value of sampling anomalous taxa in phylogenetic studies: major clades of the Asteraceae revealed. Mol. Phylogenet. Evol. 47: 757–782.

Panstruga, R. and Moscou, M.J. (2020). What is the molecular basis of nonhost resistance? Mol. Plant-Microbe Interact. 33: 1253–1264.

Peng, Y., Lai, Z., Lane, T., Nageswara-Rao, M., Okada, M., Jasieniuk, M., O'Geen, H., Kim, R.W., Sammons, R.D., Rieseberg, L.H., and Stewart, C.N. (2014). *De novo* genome assembly of the economically important weed horseweed using integrated data from multiple sequencing platforms. Plant Physiol. 166: 1241–1254.

Reyes-Chin-Wo, S. et al. (2017). Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. Nat. Commun. 8: 14953.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74: 5463–5467.

Scaglione, D. et al. (2016). The genome sequence of the outbreeding globe artichoke constructed *de novo* incorporating a phase-aware low-pass sequencing strategy of F1 progeny. Sci. Rep. 6: 1–17.

Sedlazeck, F.J., Lee, H., Darby, C.A., and Schatz, M.C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat. Rev. Genet. 19: 329–346.

Sessa, R.A., Bennett, M.H., Lewis, M.J., Mansfield, J.W., and Beale, M.H. (2000). Metabolite Profiling of

Sesquiterpene Lactones from *Lactuca* Species. J. Biol. Chem. 275: 26877–26884.

Shen, C.Z., Zhang, C.J., Chen, J., and Guo, Y.P. (2021). Clarifying recent adaptive diversification of the Chrysanthemum-group on the nasis of an updated multilocus phylogeny of subtribe Artemisiinae (Asteraceae: Anthemideae). Front. Plant Sci. 12: 874.

Shen, R., Wang, L., Liu, X., Wu, J., Jin, W., Zhao, X., Xie, X., Zhu, Q., Tang, H., Li, Q., Chen, L., and Liu, Y.G. (2017). Genomic structural variation-mediated allelic suppression causes hybrid male sterility in rice. Nat. Commun. 8: 1–10.

Sprink, T., Wilhelm, R., and Hartung, F. (2022). Genome editing around the globe: an update on policies and perceptions. Plant Physiol. 190: 1579–1587.

Subbarao, K. V, Davis, R.M., Gilbertson, R.L., Raid, R.N., and others (2017). Compendium of lettuce diseases and pests (Am Phytopath Society).

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. Science (80-.). 320: 486–488.

Tas, I.C.Q. and Van Dijk, P.J. (1999). Crosses between sexual and apomictic dandelions (*Taraxacum*). I. The inheritance of apomixis. Heredity (Edinb). 83: 707–714.

Thabuis, A.P.P., Teekens, K.C., and Van Herwijnen, Z.O. (2011). Lettuce that is resistant to the lettuce aphid *Nasonovia ribisnigri* biotype 1.

Thompson, R.C. and Ryder, E.J. (1961). Description and pedigrees of nine varieties of lettuce. Tech. Bull. Res. Serv. Agric. 1244: 1–19.

Timms, L., Jimenez, R., Chase, M., Lavelle, D., McHale, L., Kozik, A., Lai, Z., Heesacker, A., Knapp, S., Rieseberg, L., Michelmore, R., and Kesseli, R. (2006). Analyses of synteny between *Arabidopsis thaliana* and species in the Asteraceae reveal a complex network of small syntenic segments and major chromosomal rearrangements. Genetics 173: 2227–2235.

van Treuren, R., Coquin, P., and Lohwasser, U. (2012). Genetic resources collections of leafy vegetables (lettuce, spinach, chicory, artichoke, asparagus, lamb's lettuce, rhubarb and rocket salad): Composition and gaps. Genet. Resour. Crop Evol. 59: 981–997.

Trewern, J., Spajic, L., Lieb, T., Thapaliya, P., Quinn, T., Davas-Fahey, R., El-Omrani, O., and Weidgenant, L. (2021). Youth demand political action on healthy sustainable diets. Nat. Food 2: 746–747.

Underwood, C.J. et al. (2022). A PARTHENOGENESIS allele from apomictic dandelion can induce egg cell division without fertilization in lettuce. Nat. Genet. 54: 84–93.

Vijverberg, K., Welten, M., Kraaij, M., van Heuven, B.J., Smets, E., and Gravendeel, B. (2021). Sepal identity of the pappus and floral organ development in the common dandelion (Taraxacum officinale; Asteraceae). Plants 10: 1682.

de Vries, I.M. (1997). Origin and domestication of *Lactuca sativa* L. Genet. Resour. Crop Evol. 44: 165–174.

Walley, P.G. et al. (2017). Towards new sources of resistance to the currant-lettuce aphid (*Nasonovia ribisnigri*). Mol. Breed. 37.

Wei, T. et al. (2021). Whole-genome resequencing of 445 *Lactuca* accessions reveals the domestication history of cultivated lettuce. Nat. Genet. 53: 752–760.

Wei, Z., Zhu, S.X., Van den Berg, R.G., Bakker, F.T., and Schranz, M.E. (2017). Phylogenetic relationships within *Lactuca* L. (Asteraceae), including African species, based on chloroplast DNA sequence comparisons. Genet. Resour. Crop Evol. 64: 55–71.

Whitaker, T.W. (1974). Lettuce: evolution of a weedy cinderella. Hortscience 9: 512–514.

Winson, A. (2010). The demand for healthy eating: supporting a transformative food "Movement." Rural Sociol. 75: 584–600.

Zdravkovic, J., Stankovic, L., and Stevanovic, D. (2001). Possibilities of using wild lettuce forms originating from the spontaneous Yugoslav flora in the selection for virus diseases of *Lactuca sativa* L. In Proceedings of international symposium on sustainable use of plant biodiversity to promote new opportunities for horticultural production development (Antalya, Turkey), pp. 243–245.

Zhang, L. et al. (2017). RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. Nat. Commun. 8: 2264.

Zhang, N.W., Lindhout, P., Niks, R.E., and Jeuken, M.J.W. (2009a). Genetic dissection of *Lactuca saligna* nonhost resistance to downy mildew at various lettuce developmental stages. Plant Pathol. 58: 923–932.

Zhang, N.W., Pelgrom, K., Niks, R.E., Visser, R.G.F.F., and Jeuken, M.J.W.W. (2009b). Three combined quantitative trait loci from nonhost *Lactuca* saligna are sufficient to provide complete resistance of lettuce against *Bremia lactucae*. Mol. Plant. Microbe. Interact. 22: 1160–8.

Zhao, T., Holmer, R., Bruijn, S. de, Angenent, G.C., van den Burg, H.A., and Schranz, M.E. (2017). Phylogenomic synteny network analysis of MADS-Box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation. Plant Cell 29: tpc.00312.2017.

Zhao, T. and Schranz, M.E. (2017). Network approaches for plant phylogenomic synteny analysis. Curr. Opin. Plant Biol. 36: 129–134.

Zheng, G.X.Y. et al. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat. Biotechnol. 34: 303–311.

Zohary, D. (1991). The wild genetic resources of cultivated lettuce (*Lactuca sativa* L.). Euphytica 53: 31–35.

Chapter 2

The genome of *Lactuca saligna*, a wild relative of lettuce, provides insight into non-host resistance to the downy mildew *Bremia lactucae*

Wei Xiong[1], Lidija Berke[1,$], Richard Michelmore[2], Dirk-Jan M. van Workum[3],
Frank F.M. Becker[1], Elio Schijlen[4], Linda V. Bakker[4], Sander Peters[4],
Rob van Treuren[5], Marieke Jeuken[6], Klaas Bouwmeester[1], M. Eric Schranz[1]

[1] Biosystematics Group, Wageningen University and Research, Wageningen, The Netherlands.
[2] Genome Center and Department of Plant Sciences, University of California, Davis, United States.
[3] Bioinformatics Group, Wageningen University, Wageningen, the Netherlands.
[4] Bioscience, Wageningen University and Research, Wageningen, The Netherlands.
[5] Centre for Genetic Resources, The Netherlands (CGN), Wageningen University and Research, Wageningen, The Netherlands.
[6] Plant Breeding, Wageningen University and Research, Wageningen, The Netherlands.
[$] Present address: Genetwister Technologies B.V., Wageningen, The Netherlands.

This chapter has been accepted by: The Plant Journal.

Abstract

Lactuca saligna L. is a wild relative of cultivated lettuce (*Lactuca sativa* L.), with which it is partially interfertile. Hybrid progeny suffer from hybrid incompatibilities (HI), resulting in reduced fertility and distorted transmission ratios. *Lactuca saligna* displays broad spectrum resistance against lettuce downy mildew caused by *Bremia lactucae* Regel and is considered a non-host species. This phenomenon of resistance in *L. saligna* is called non-host resistance (NHR). One possible mechanism behind this NHR is through the plant–pathogen interaction triggered by pathogen-recognition receptors, including nucleotide-binding leucin-rich repeats (NLRs) and receptor-like kinases (RLKs). We report a chromosome-level genome assembly of *L. saligna* (accession CGN05327), leading to the identification of two large paracentric inversions (>50 Mb) between *L. saligna* and *L. sativa*. Genome-wide searches delineated the major resistance clusters as regions enriched in *NLR*s and *RLK*s. Three of the enriched regions co-locate with previously identified NHR intervals. RNA-seq analysis of *Bremia* infected lettuce identified several differentially expressed *RLK*s in NHR regions. Three tandem wall-associated kinase-encoding genes (*WAK*s) in the NHR8 interval display particularly high expression changes at an early stage of infection. We propose *RLK*s as strong candidate(s) for determinants for the NHR phenotype of *L. saligna*.

Keywords

Lettuce breeding, downy mildew disease, hybrid incompatibility, non-host resistance, immune genes, *de novo* genome assembly, comparative genomics, population genetics

1. Introduction

Lettuce (*Lactuca sativa* L.) is a leafy vegetable grown in more than 100 countries, with a total yield of over 29 million tons in 2019 (FAOSTAT, 2019). One of the most important goals for lettuce breeding is the introgression of durable resistance against lettuce downy mildew, a destructive disease caused by the oomycete pathogen *Bremia lactucae* Regel (Lebeda et al., 2009). Outbreak of downy mildew disease leads to substantial yield and economic losses.

Wild relatives of lettuce are often used to introgress novel resistances (Lebeda et al., 2014; Parra et al., 2016). *L. saligna*, which belongs to the secondary gene pool of lettuce, is an important donor to enhance resistance to *B. lactucae* in cultivated lettuce (Netzer et al., 1976; Norwood et al., 1981; Bonnier et al., 1991). *L. saligna* is a diploid (2n=2x=18, same as lettuce) and self-pollinating species, which is partially interfertile with *L. sativa* (Lebeda et al., 2007, 2019). It is broadly distributed across Eurasia, from the Mediterranean region towards temperate Europe, and from the Iberian Peninsula to Central Asia (Zohary, 1991; Doležalová et al., 2002; Lebeda et al., 2019). *Lactuca saligna* is of particular interest to lettuce breeders as a potential resistance donor due to its complete resistance to all races of *B. lactucae.* As such, it is considered a non-host species to *B. lactucae* based on the definition: “All genotypes of a species are resistant against all genotypes of a specific pathogen” (Bonnier et al., 1991; Petrželová et al., 2011; Lebeda et al., 2009; Heath, 1981). For convenience, we term this resistance phenotype of *L. saligna* as non-host resistance (NHR), which is defined as strictly phenomenological and does not imply a molecular mechanism (Panstruga and Moscou, 2020). To successfully introgress this NHR in lettuce cultivars the gene(s) underlying the non-host resistance and the reproductive barriers observed in hybrid offspring should be determined.

Although *L. saligna* is crossable with *L. sativa*, the F_1 plants are nearly sterile, and the resulting inbred offspring (F_2 generation) show severely reduced fertility and transmission ratio distortions due to hybrid incompatibility (HI) (Jeuken et al., 2001; Giesbers et al., 2019). Some case of HI can be explained by the deleterious combination of interspecific alleles according the Dobzhansky-Muller (DM) model (Dobzhansky, 1934; Muller, 1942; Bateson, 1909). Many identified and resolved HI loci are explained by a digenic deleterious epistatic interaction and often results in transmission ratio distortion (TRD) (Fishman and Sweigart, 2018; Fishman and McIntosh, 2019). In F_2 offspring and backcross inbred lines (BILs; i.e., single segment introgression lines) of *L. saligna* x *L. sativa*, 11 HI loci were associated with TRD, six of which were nullified by a paired allele from *L. saligna* (Giesbers et al., 2019). HI loci may reduce the efficiency of introgression of NHR genes from *L. saligna* into *L. sativa* when HI- and NHR loci are closely linked (i.e., linkage drag). In addition to HI, an interspecific chromosomal rearrangement, like an inversion, will also hamper the introgression of desired NHR genes via linkage drag caused by reduced recombination (Hoffmann and Rieseberg, 2008; Fishman and Sweigart, 2018).

NHR in plants is suggested to rely on a continuum of layered defenses, including both constitutive and induced resistance mechanisms (Niks and Marcel, 2009; Jones and Dangl, 2006; Bettgenhaeuser et al., 2014). Previous studies propose that induced NHR and host immunity rely on a similar non-self-recognition system comprising two innate immunity layers: i) pattern-triggered immunity (PTI) mediated by extracellular recognition of conserved non-self-molecules − called pathogen-associated molecular patterns (PAMPs) − by cell surface receptors, such as diverse receptor-like kinases (RLKs), and ii) host defense conferred by effector-triggered immunity (ETI) mediated by *R* genes encoding intracellular nucleotide-binding leucine-rich repeat proteins (NLRs) that recognize cognate pathogen-secreted effector molecules (Niks and Marcel, 2009; Schulze-Lefert and Panstruga, 2011; Jones and Dangl, 2006; Chisholm et al., 2006). After host penetration, hyphal growth of *B. lactucae* is quickly halted in *L. saligna* and consequently haustorium formation is impeded (Niks, 1987; Lebeda and Reinink, 1994; Zhang et al., 2009a, 2009b). To identify common loci associated with NHR to *B. lactucae* in lettuce, Giesbers *et al.* (2018) performed mapping studies based on nine *L. saligna* accessions from a broad range of geographic regions via multiple bidirectional backcrosses: i.e., i) BC1 populations in both parental directions (F_1 x host *L. sativa*) and (F_1 x non-host *L. saligna*), and ii) BC1S3 lines with three generations of inbreeding, respectively. These mapping populations facilitated the identification of four epistatic segments accounting for NHR in *L. saligna*: one positioned on Chromosome 4 (NHR4), two on Chromosome 7 (NHR7.1 & 7.2), and another on Chromosome 8 (NHR8) (Giesbers et al., 2018). It is worth noting that the NHR8 interval is closely linked to HI/TRD loci, which potentially limits fine-mapping and introgression of non-host traits into *L. sativa*. The genes and mechanisms underlying these four NHR loci are unresolved. A high-resolution analysis of these regions is needed to unveil the genetic determinants governing NHR in *L. saligna*.

*NLR*s in lettuce were previously identified by Christopoulou *et al.* (2015a) using the L. sativa v6 genome. Identified NLRs were classified into two major groups: TNLs with TOLL/interleukin-1 receptor (TIR) domains and CNLs with coiled-coil (CC) domains, and subsequently into multiple resistance gene candidates (RGC) families (Takken and Goverse, 2012; McHale et al., 2006; Meyers et al., 2003). Almost all identified *NLRs* were found to reside in five major resistance clusters (MRCs) that co-segregate with resistance to diverse pathogens (McHale et al., 2009; Christopoulou et al., 2015b). For example, MRC2 on Chromosome 2 comprises multiple *RGC2* family members, including the downy mildew resistance genes *Dm3*, *Dm14*, *Dm16*, and *Dm18* (Shen et al., 2002; Wroblewski et al., 2007; Christopoulou et al., 2015a). Similar MRCs are suggested to be present in *L. saligna* based on the expected whole-genome synteny, since some qualitative resistance phenotypes from *L. saligna* have been mapped at single loci syntenic to MRCs in *L. sativa* (Giesbers et al., 2017). An *L. saligna* genome reference can facilitate synteny analysis to recognize these anticipated MRCs.

Multiple *RLK* families contain members involved in a wide range of immune responses in plants. Notable examples can be found in the LRR-RLK sub-family, such as FLS2 involved in the perception of bacterial flagellin and IOS1 that contributes towards resistance to the downy mildew *Hyaloperonospora arabidopsidis* (Zipfel et al., 2004; Hok et al., 2011). Previously, Christopoulou *et al.* (2015a) also described *LRR-RLK* encoding genes in lettuce. Nevertheless, a specific inventory of genes encoding other resistance-related *RLK*s in lettuce, such as those encoding lectin receptor kinases (LecRKs) and wall-associated kinases (WAKs), is still lacking, not to mention the *RLK*s of *L. saligna* (Bouwmeester et al., 2011; Hu et al., 2017; Zuo et al., 2015; Hurni et al., 2015; He et al., 1999).

Here, we report a *de novo* genome assembly of *L. saligna* (accession CGN05327) using a variety of sequencing and scaffolding techniques. The assembly was compiled into nine chromosomal pseudo-molecules by genetic mapping. The resulting assembly enabled us to conduct diverse genomic analyses to dissect the genetic determinants underlying non-host resistance in *L. saligna*. The analyses provide insights of evolution into disease resistance and on host-pathogen arms race in lettuce. For breeding, the gained knowledge helps to facilitate the introgression of *Bremia* resistance into cultivated lettuce.

2. Results

2.1 Genome sequencing and assembly

L. saligna accession CGN05327 was used to produce a reference genome for *L. saligna* (Supplemental Note). A combination of PacBio long-read (95.4 Gb; 41X) and Illumina short-read (407.4 Gb; 175X) sequencing was generated to assemble the genome (Supplemental Data 1). Illumina paired-end (125 bp, PE) and mate-pair (300 bp, MP) reads were generated from three libraries of different insert size (200 bp, 500 bp and 550 bp) (Supplemental Data 1A). The *L. saligna* genome size was estimated by K-mer analysis to be 2.27 Gb, which agrees with genome size estimates established by flow cytometry (2.3 Gb; Doležalová et al., 2002; Zohary, 1991). K-mer analysis also revealed that the genome is highly homozygous (estimated heterozygosity = 0.12%) as expected for this inbreeding species (Supplemental Figure 1 and Supplemental Table 1). To construct a high-quality genome of *L. saligna*, we applied a variety of advanced assembly and mapping techniques (Supplemental Figure 2). An initial Canu assembly (v0.5) consisted of 31,431 contigs, and the N50 number and size were 6,957 and 88.0 kb, respectively (Table 1; Supplemental Data 1A). Bionano fingerprinting, 10x Genomics barcoding, and Dovetail Hi-C library data were sequentially applied to construct the version 2 assembly, which refined the assembly to 24 super-scaffolds (largest scaffold = 279.9 Mb; N50 = 146.7 Mb; Supplemental Data 2B-C).

Table 1. Genome assembly summary.

	Sequencing	**Scaffolding**				
Statistics	**PacBio + Illumina**	**Bionano + 10X Genomics**		**Dovetail**	**Genetic mapping**	**Merge unmapped**
Version	v0.5	Unmapped	v1	v2	v3	v4
N50/number	6,957	-	307	5	4	-
N50/size	88.0 Kb	-	1.8 Mb	146.7 Mb	192.1 Mb	-
N90/number	22,020	-	928	13	8	-
N90/size	31.0 Kb	-	0.6 Mb	62.2 Mb	151.4 Mb	-
Largest contig/scaffold	794.0 Kb	1.1 Mb	8.2 Mb	279.9 Mb	279.9 Mb	-
Size of assembly (Gb)	2.03 Gb	0.42 Gb	1.75 Gb	1.75 Gb	1.75 Gb	2.17 Gb
Contig/scaffold number	31,431	6,174	1,376	24	9+7	9+1

2.2 Linkage group anchoring and assembly assessment

To generate chromosomal pseudo-molecules, we combined 417 genetic markers from an F_2 population linkage map (*L. saligna* x *L. sativa*) and 19,027 syntenic markers between *L. saligna* and *L. sativa* (Supplemental Table 2; Supplemental Data 3A-B). This resulted in a chromosome-level assembly (v3) in which 17 out of 24 scaffolds (99.8% bases, 1.75 Gb) were anchored and oriented into nine chromosomes, covering ~77% of the estimated genomic sequence (1.75 of 2.27 Gb) (Supplemental Table 3; Supplemental Data 2D; Supplemental Figure 3). To obtain a more complete reference assembly, un-scaffolded contigs (>1000 bp) were merged to create a virtual "chromosome zero." This eventually led to a final assembly (v4) with nine chromosomes plus chromosome zero, with a complete genome size of 2.17 Gb (Table 1; Supplemental Table 4; Supplemental Data 2E-F). This final assembly contains 91.9% (1,951 out of 2,121) of the expected BUSCO (1,859 single and 92 duplicated copies) eudicot gene models (Supplemental Table 5), and 92% of the 30,696 *L. saligna* expressed sequence tags (ESTs) in NCBI could be aligned to the v4 assembly at 80% identity and 80% coverage (Supplemental Table 6).

2.3 Repeat and non-coding RNA annotation

Our analyses estimated that 77.5% of the *L. saligna* genome consists of transposable elements (TE; Table 2; Supplemental Table 7). Long terminal repeat retrotransposons (LTR-RT) were the most predominant repetitive elements, comprising both Gypsy and

Copia retrotransposons (43.8% and 23.1% of genome, respectively) (Supplemental Table 8-9). TEs were distributed across the genome, and found to be enriched in regions roughly representing the pericentromeric locations (Supplemental Figure 4: track E). Non-coding RNAs involved in mRNA transcription (snRNA), translation (tRNAs and rRNAs), and regulation of gene expression (miRNAs) were also annotated (Table 2; Supplemental Table 10).

2.4 Gene prediction and functional annotation

A combination of *de novo* search and homology support was applied for gene model prediction. Most of the predicted gene models (93%) were well supported (AED > 0.5) by RNA-seq data and gene homology (Supplemental Figure 5; Supplemental Data 4A). In total, 42,908 gene models were retained after filtering based on coding-potential (Table 2). The average coding size and exon number per gene was 1.3 kb and 5.1 respectively (Table 2). We further validated the potential for protein-encoding sequences using domain, ortholog, and homolog databases. By combining all results, 40,730 genes (94.9%) had matches in at least one database (Table 2; Supplemental Table 11; Supplemental Data 4B).

Table 2. Genome annotation summary.

Genome annotation	Metrics	Statistics
Gene prediction	*n* of genes	42,908
	Mean length of CDS	1,117 bp
	Mean exon number	5.1
	n protein-coding genes	40,730 (94.9%)
ncRNA	*n* of rRNAs	4,114
	n of tRNAs	1,857
	n of miRNAs	128
	n of snRNAs	329
Transposable elements	%Retrotransposons	67.8% (1.5 Gb)
	%DNA transposons	3.1% (66.9 Mb)
	%Unclassified repeats	6.6% (143.6 Mb)
	%Total	77.5% (1.7 Gb)

2.5 *Lactuca saligna* population structure and diversity

To explore the genetic diversity and population structure of *L. saligna*, we re-sequenced 15 accessions representing the distribution across its native range (Supplemental Table 12-13). SNPs were first called on the *L. saligna* genome assembly and then filtered on missing rate (<10 %) and minor allele frequency (>0.05), yielding 5,170,479 SNPs for downstream analysis (Supplemental Table 14-15). After pruning the SNP dataset, we applied three complementary methods to explore the structure of *L. saligna*: neighbor-joining tree building, principal component analysis (PCA), and ancestry history inference. The neighbor-joining tree revealed that the *L. saligna* population can be subdivided into three major clades that are largely congruent with the geographical origins of the selected accessions (Figure 1A). This finding was recapitulated by PCA (Figure 1B) and ADMIXTURE analysis (Figure 1C). These analyses also uncovered the geographical origins of two accessions that were previously unknown. Accession CGN05271 is implicated to be of European origin, whereas CGN05282 groups with multiple accessions from the Middle East (Figure 1D). It is noteworthy to mention that accession CGN05271, now found to be of European origin, has been extensively utilized in many in-depth genetic studies on resistance to downy mildew or reproductive barriers (Jeuken and Lindhout, 2002; den Boer, 2014; Giesbers et al., 2017, 2018; Jeuken et al., 2001; Giesbers et al., 2019). Our sequenced reference CGN05327 is genetically clustered with CGN05271. Finally, the leaf morphology of each accession was also found in line with the *L. saligna* population genetic structure (Supplemental Figure 6).

2.6 Synteny between *L. saligna* and *L. sativa*

Duplication events and structural variation were identified between the *L. saligna* and *L. sativa* genomes by syntenic alignments. Intra-species collinearity revealed a 3:1 syntenic pattern in all nine chromosomes for both species, confirming the known shared whole-genome triplication event within the Asteraceae (Reyes-Chin-Wo et al., 2017; Iorizzo et al., 2016) (Supplemental Figure 7). Inter-species syntenic analysis revealed a high level of genome-wide collinearity between both *Lactuca* species (Figure2A), except for two large inversions (> 50 Mb) on Chromosomes 5 and 8 (Figure 2B-C; Supplemental Table 16). The observed gene density (~ 20 genes per Mb) within these two inverted regions in both species suggests that they are not close to the centromere, i.e., paracentric inversions (Supplemental Table 16). The ranges and positions of inversions were estimated using syntenic genes at the inversion borders (Supplemental Table 17). To confirm these inversions, we mapped markers derived from an interspecific F_2 population to the *L. saligna* genome and compared their genetic and genomic positions. This showed that the genetic position plateaued while the genomic position kept increasing over the inverted region, which reflects the suppressed recombination due to inversion (Supplemental Figure 8). These inversions encompass of a diversity of genes, some of which encode proteins known to play key roles in various biological processes, such as a methyltransferase involved in Vitamin E biosynthesis and a phosphatase regulating cell wall integrity (Supplemental Table 18-19) (Cheng et al., 2003; Franck et al., 2018).

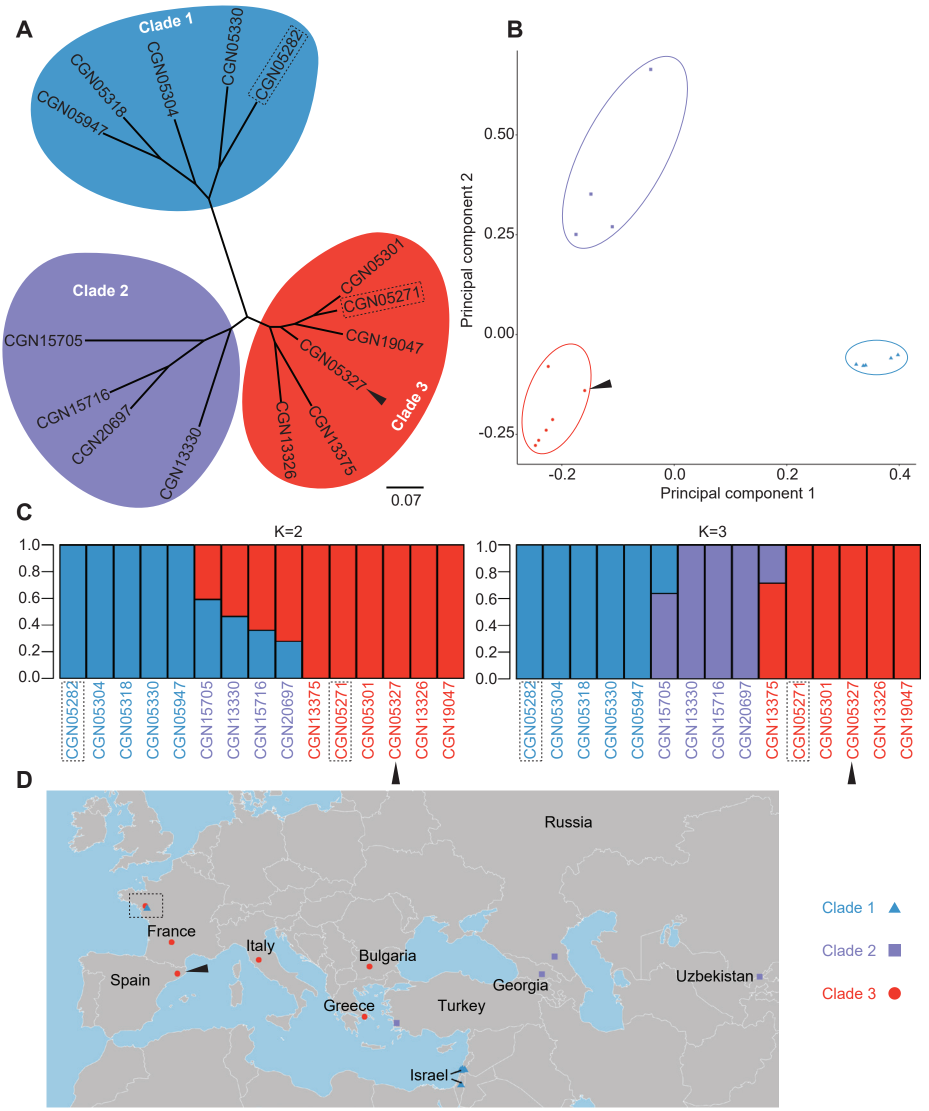


Figure 1. Resequencing of 15 accessions illustrates the *L. saligna* population structure. **A**, Neighbor-joining tree of 15 re-sequenced *L. saligna* accessions based on called SNPs. Accessions were clustered into three clades (colored in red, blue, and purple). Two accessions with unknown origins obtained from a French botanical garden are labelled by dashed lines. The black arrow indicates reference accession CGN05327 used for *de novo* sequencing. **B**, Principal component analysis plot of the top two-components illustrating the *L. saligna* population structure. Colors and shapes correspond to clades 1, 2, and 3. **C**, Genetic ancestry estimation with presumed populations (K=2 and K=3) indicating the population number and evolution. Red and blue represents the two ancestral populations and the purple indicates an intermediate population between the two ancestors. **D**, Geographic locations of *L. saligna* accessions, colored and shaped based on population structure.

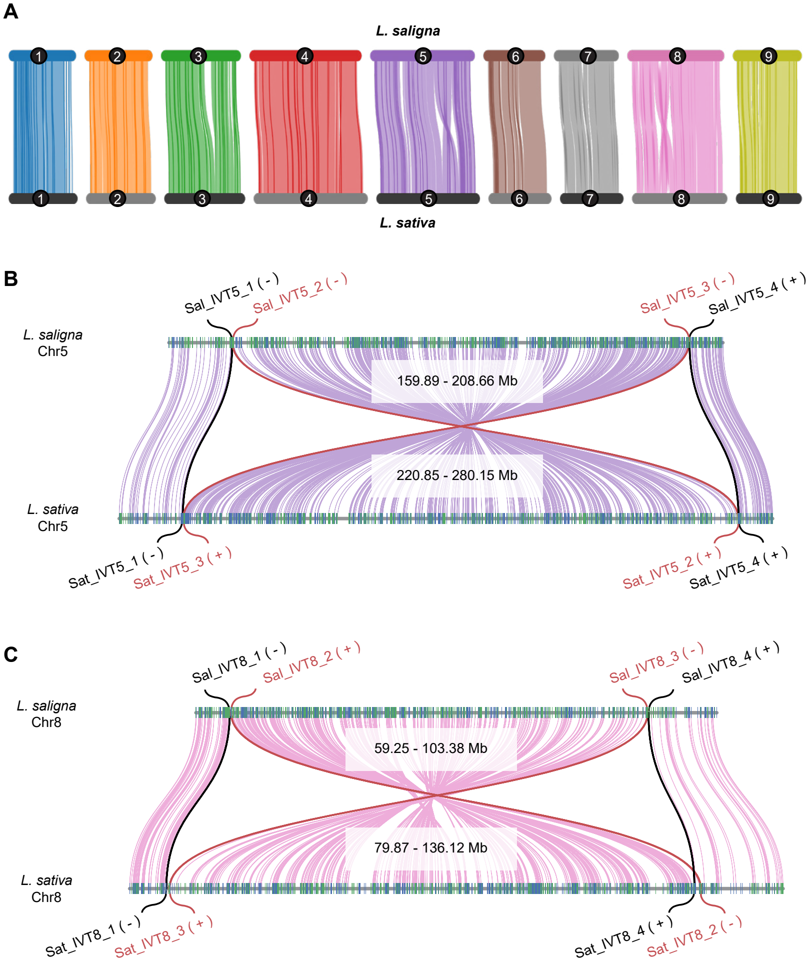


Figure 2. Synteny reveals two large inversions on chromosomes 5 and 8 between *L. saligna* and *L. sativa*. **A**, Synteny of best orthologs for each chromosome between the two *Lactuca* species. Each chromosome is represented by a different color. **B**-**C**, Inverted synteny regions on chromosome 5 (purple) and 8 (pink) with 50 flanking genes shown at borders, respectively. Relative to *L. saligna*, the red lines link the first and last homologous gene pairs within inverted synteny, while the black lines indicate the first homologous pairs outside of the inversion.

2.7 Comparison of *NLR* content and distribution between *L. saligna* and *L. sativa*

To explore variation in the *NLR* gene family, HMMER and BLAST searches were conducted against the proteomes of *L. saligna* and *L. sativa* (Supplemental Data 5A-B). Retrieved amino acid sequences were first classified based on their N-terminal TIR or CC domain (TNLs and CNLs, respectively) and thereafter subdivided to Resistance Gene Candidate (RGC) families by phylogenetic analyses (Supplemental Figure 9; Supplemental Data 5C-D). This resulted in the identification of 323 *NLRs* in *L. saligna* and 364 *NLRs* in *L.*

sativa. *Lactuca saligna* and *L. sativa* were found to contain a similar content of both TNL- and CNL-type, i.e., 184 versus 202 (57.0%, 55.5%), and 139 versus 162 (43.0%, 44.5%), respectively (Table 3; Supplemental Table 20). Genomic positions of MRCs previously identified in *L. sativa* were identified in the *L. saligna* genome assembly using *L. sativa* orthologs (Supplemental Table 21; Supplemental Data 5D). We additionally defined two *NLR*-enriched clusters (NCs) in *L. saligna* on Chromosomes 4 (38.55 – 40.68 Mb) and 7 (44.01 – 44.48 Mb), hereafter named NC4 and NC7 (Supplemental Table 22). These two NCs were also identified in *L. sativa,* but were not previously labeled as MRCs due to the absence of resistance phenotypes (Christopoulou et al., 2015b). In total, 41 RGC families were identified. Seven RGC families (six singletons and one multigene family) present in *L. sativa* were found missing in *L. saligna*, which might be caused by the reconstructed phylogeny or they may be unique to *L. sativa* (Supplemental Table 23). While *L. saligna* has a similar amount of NLRs compared to *L. sativa* in most RGCs, we defined significant size change by count and percentage difference. In this way, we observed that six and three RGC families were contracted (i.e. RGC1, 4, 8, 9, 14, and 21) and expanded (i.e. RGC 16, 20, and 29), respectively, in this accession of *L. saligna* compared to the reference genome of *L. sativa* (Supplemental Figure 9; Supplemental Table 23).

Table 3. Identification and classification of *NLR*s and *RLK*s for *L. saligna and L. sativa*.

Immune genes		**Species**	
Family	**Classification**	***L. saligna***	***L. sativa***
NLR	CNL[a]	139	162
	TNL	184	202
	Total	323	364
RLK[b]	LRR-RK	213	245
	G-LecRK	79	128
	Malectin-RK	55	50
	WAK	48	53
	CRK	35	36
	L-LecRK	29	35
	LysM-RK	12	12
	Rcc1-RK	5	5
	C-LecRK	1	1
	Other[c]	1	1
	Total	478	566

[a] Including RPW8 and Rx_N type of CNLs.

[b] Extracellular domain architecture via HMMER (Supplemental Table 24; Supplemental Dataset 5A-C).

[c] According to iTAK classification the other RLKs are a RLK-Pelle_DLSV (*L. sativa*) and a RLK-Pelle_PERK-1 (*L. saligna*) (Supplemental Dataset 5A-B).

2.8 Comparison of RLK genes between *L. saligna* and *L. sativa*

To identify genes encoding RLK proteins, we performed in-depth HMMER searches against the predicted *L. saligna* and *L. sativa* proteomes. This resulted in the identification of 478 and 566 *RLK* encoding genes in *L. saligna* and *L. sativa,* respectively (Supplemental Table 24; Supplemental Data 6A). Sliding window analysis revealed that *RLKs* are distributed on all chromosomes, with their density elevated at the chromosomal ends (Figure 3: track B). RLKs were further classified into nine subfamilies based on their extracellular domains using HMMER (Table 3; Supplemental Data 6B). In both species, LRR-RLKs and G-type LecRKs (G-LecRKs) formed the largest subfamilies. The major difference in total *RLK*s was also largely accounted by these two subfamilies − with an additional 32 G-LecRKs and 48 LRR-RLKs in *L. sativa*. The other *RLK* subfamilies were found to be of similar size in these accessions of both *L. saligna* and *L. sativa*.

2.9 Mapping HI and NHR loci on *L. saligna* genome

To precisely characterize the HI and NHR regions, markers of these loci were mapped to the *L. saligna* assembly (Supplemental Table 25-26). For HI, one locus was positioned on Chromosome 8 (33.15–138.07 Mb) and contains the inversion identified on Chromosome 8, 59–103 Mb (Figure 3: track A and D; Supplemental Figure 8) (Giesbers et al., 2019). This HI region and inversion region on Chromosome 8 was also adjacent to the resistance-related regions NHR8, MRC8B, and MRC8C (Figure 3: track C-D; Supplemental Figure 8). For NHR, three out of four intervals either overlapped with *NLR* or *RLK* hotspots. NHR7.1 was found to co-segregate with the NC7 region encoding 13 *NLR*s, whereas the other three NHR intervals consist of no or only one *NLR* gene (Supplemental Data 7). Moreover, mapping revealed that both NHR4 and NHR8 co-locate with regions enriched in *RLK* genes (NHR4: 20 *RLKs* in 34.21 Mb; NHR8: 14 *RLKs* in 13.26 Mb). Especially for region NHR8, the *RLK* density (1.06/Mb) was five-times higher than the genome-wide average (0.24/Mb, 422 in 1,745 Mb, excluding Chromosome 0). Mapping revealed a close relationship between NHR regions and resistance gene hotspots, making NLRs/RLKs potential determinants of NHR in *L. saligna*. In addition, NHR8 is also positioned near an HI segment, which may prevent the introgression of the candidate resistance genes to cultivated lettuce, impacting breeding for resistance.

2.10 RNA-seq time-course analysis of *L. saligna* transcriptome in response to *Bremia*

To detect genes with differential expression after infection, we performed a *Bremia* infection assay on leaves of *L. saligna* to generate transcriptomic data and subsequently conducted a differential expression (DE) analysis. Treated and control samples were collected at 8- and 24-hours post-infection (hpi). Statistical analysis of quantified RNA-seq reads count identified a total of 1,268 and 1,688 differentially expressed genes (DEGs) (padj < 0.5 and log2FC > 1) at 8 hpi and 24 hpi, respectively (Supplemental Table 27; Supplemental Data 7). For both time points, the majority of DEGs were up-regulated in expression, i.e. 1,222 up-regulated versus 46 down-regulated genes at 8 hpi, and

1,362 up-regulated versus 326 down-regulated genes at 24 hpi (Supplemental Table 28). One of the most representative DEGs is Lsal_1_v1_gn_1_00001954, showing the largest induction in expression (log2FC=11.72), is a homolog of the penetration resistance gene *PEN1*, which encodes a syntaxin involved in vesicle assembly for non-host resistance against powdery mildew penetration in *Arabidopsis* (Collins et al., 2003).

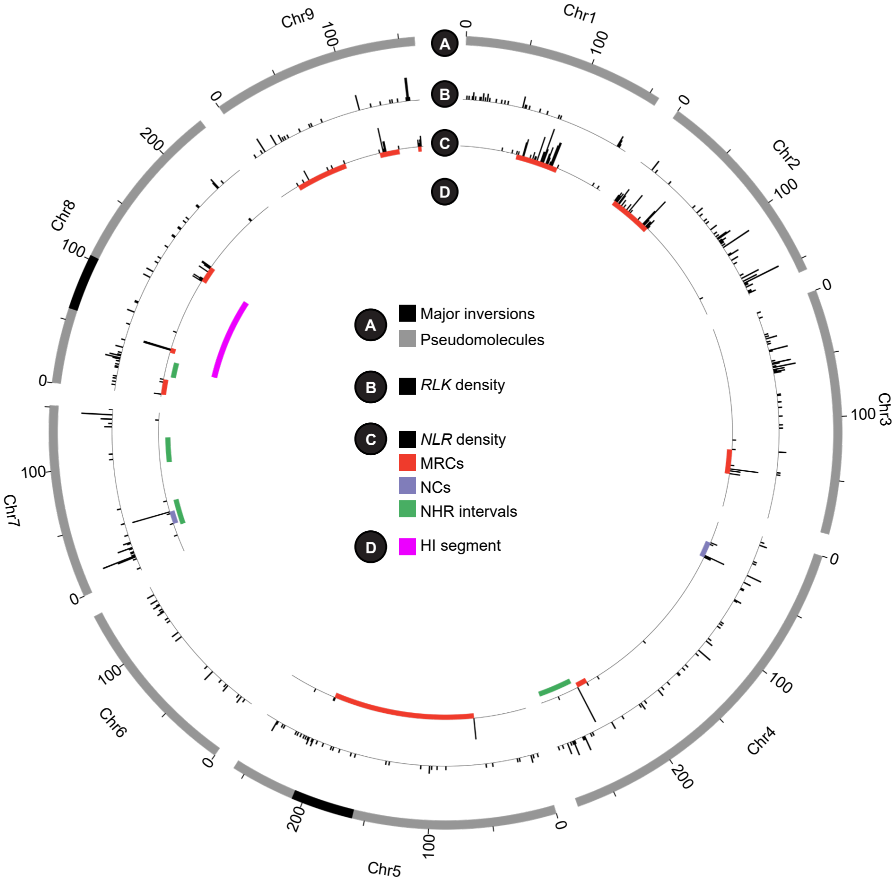


Figure 3. Phenotype mapping associates immune gene hotspots with NHR and HI regions. Track **A**, Circular ideogram of the nine pseudo-chromosomes (Mb) of the *L. saligna* assembly indicating two major inverted regions between *L. saligna* and *L. sativa.* Track **B**, Histogram of *RLK* density (1Mb window). Track **C**, *NLR* density (1Mb window) and tiles related to disease-resistance gene cluster intervals: i.e. major resistance clusters (MRCs), *NLR* clusters (NCs) with elevated density, and previously identified NHR interval. Track **D**, HI segment found on chromosome 8 using backcross inbred lines (*L. saligna* x *L. sativa*).

2.11 Enrichment analysis of identified DEGs in *L. saligna*

Subsequently, we applied gene ontology enrichment analysis of DEGs to explore functional-related biological processes and pathways. Figure 4 shows the 20 most significantly enriched terms related to DEGs at 8 hpi or 24 hpi. Sixteen out of 20 ontology terms were identified at both time points. Most clusters were mainly associated with resistance responses, like stress perception (GO:0009620), signal transduction (GO:0046777), and cell death (GO:0008219). In general, 8 hpi showed a greater enrichment than 24 hpi for most top terms (Figure 4B). In contrast, three unique biological clusters were found for the 24 hpi timepoint, all of which were related to ribosome biogenesis (GO:0042254, GO:0042273, and ath03010) (Figure 4A-B). In addition to the top 20 terms, many up-regulated genes were found to be involved in plant defense, in particular in response to oomycetes, illustrating the immune response of *L. saligna* upon *Bremia* infection (Supplemental Figure 10; Supplemental Table 28-29). For example, these include Lsal_1_v1_gn_9_00004094, a homolog of the lectin receptor gene *LecRK-IX.1* conferring resistance to *Phytophthora* spp. (another oomycete pathogen); Lsal_1_v1_gn_8_00004656 (*SARD1*) and Lsal_1_v1_gn_2_00003439 (*UGT76B1*), encoding two key regulators of salicylic acid (SA) synthesis and SA mediated signaling for stress response (Wang et al., 2015; Ding et al., 2016; Mohnike et al., 2021; Bauer et al., 2021). Our enrichment analysis detected that DEGs at both time points post-inoculation with *Bremia* were enriched in resistance-related biological processes: 8 hpi showed a stronger signal of early immune response and 24 hpi showed a shift of enriched terms to extra post-transcriptional response.

2.12 Differentially expressed genes in NHR regions at 8 hpi

Based on above mapping and DE analysis results, we inspected the statistics of the up-regulated genes in NHR intervals at 8 hpi to further identify candidates for resistance to lettuce downy mildew. First, we calculated the DEG density per million base-pair of four NHR loci and the whole genome (Supplemental Table 30). As baseline, the DEG density for the whole genome was 0.70 per Mb. The NHR8 locus had the highest DEG density (1.54/Mb) among all NHR intervals and was greater than two-times the average density of the entire genome. Moreover, 11 DEGs located in the overlapping region of NHR8 and HI may also inhibit the ability to overcome the hybrid barrier. Secondly, we examined the percentage of up-regulated *RLKs* and *NLRs* (up-regulated number / total number) for each NHR interval (Supplemental Table 30). The percentage of differentially expressed *NLRs* was low (4.6%) across the whole genome. None of the *NLR*s within the two NHR loci were differential expressed. In contrast, more than 22.7% of the *RLKs* (96) were up-regulated genome-wide after *Bremia* inoculation. NHR8 also displayed a high percentage of up-regulated *RLKs* (50%, seven out of 14). Furthermore, we counted the number of DEGs with a large degree of change (log2FC > 3) in NHR regions of interest (Supplemental Table 30). Again, NHR8 was found to contain more highly expressed genes (nine) than the other three NHR regions. Thus, out of four NHR loci, the statistics of DEGs strongly suggests that genes on NHR8 seemed to play a critical role in the resistance to *Bremia*, especially the *RLK*s.

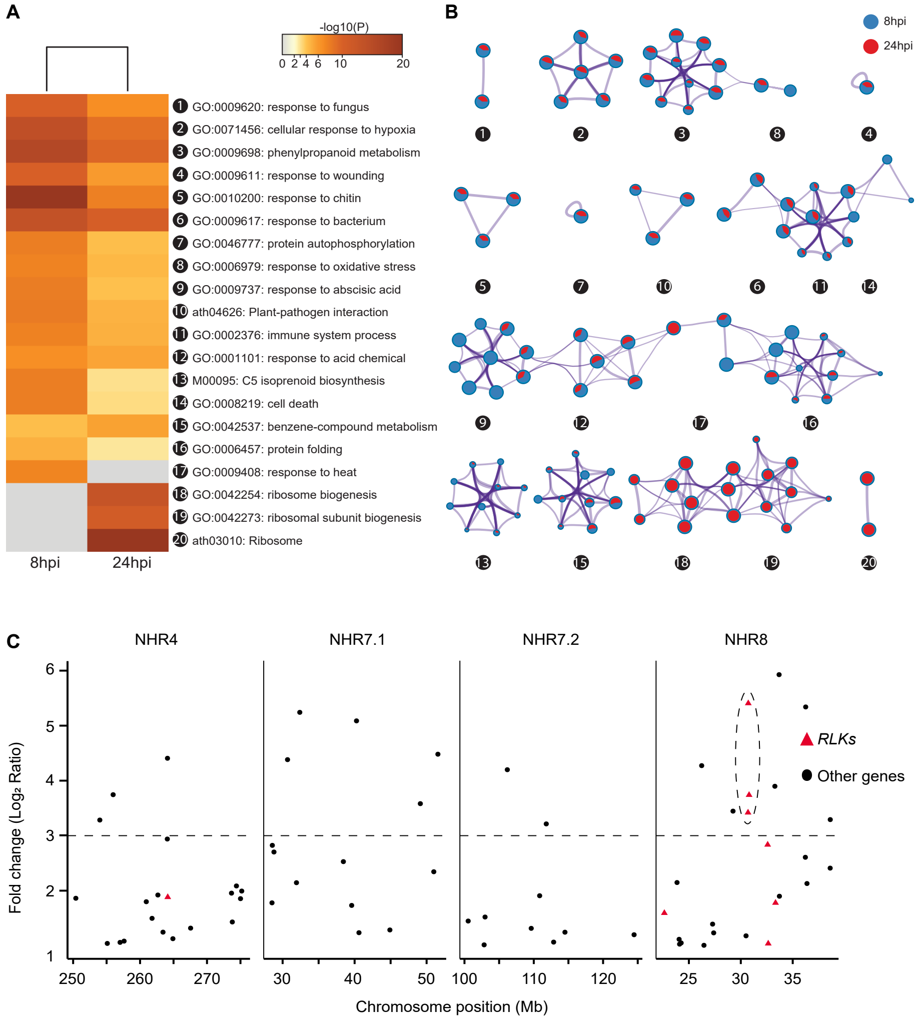


Figure 4. Enrichment analysis and expression levels of DEGs in *L. saligna* upon *Bremia* infection. **A**, Heatmap of the top 20 ontology groups at 8 and 24 hpi. Each group comprises multiple ontology terms and is represented by the term with the best p-value. Groups are hierarchically clustered and heatmap cells are colored according transformed p-values [-$\log_{10}$(p-value)]. Grey cells indicate a lack of enrichment for that term in the corresponding gene list. **B**, Networks of representative terms for the top 20 groups. Each term is displayed by a pie chart node to illustrate the proportional number of up-regulated genes at 8 hpi (blue) and 24 hpi (red). Some groups are interconnected and form a larger network. **C**, Distribution of up-regulated genes across the four identified NHR regions in *L. saligna* at 8 hpi. The horizontal dashed line (y=3) indicates the cutoff for up-regulated genes ($\log_2$FC > 3). Receptor-like kinases (*RLKs*) are indicated by red triangles, and other genes are black circles. The dashed ellipse line points out the three tandem arrayed *WAKs* on NHR8.

Based on these observations, we pinpointed eight DEGs located in NHR8 as candidates for downy mildew resistance in *L. saligna* (Supplemental Table 31). One of the candidate genes encodes a plant U-box type E3 ubiquitin ligase (PUB), of which family members have been reported to play essential roles in plant defense and disease resistance (González-Lamothe et al., 2006). The other candidate genes all encode receptor-like kinases, i.e., one LysM-containing receptor-like kinase (LysM-RK), three G-type lectin receptor kinases (G-LecRKs), and three WAKs. It is noteworthy to mention that the three *WAK*s were tandem-arrayed, of which two were highly up-regulated (log2FC > 3; Figure 4C).

3. Discussion

3.1 *L. saligna* reference genome and population structure

In this study, we report on the *de novo* genome assembly of *L. saligna* based on long- and short-read sequencing together with advanced scaffolding techniques. The genome size of our *L. saligna* assembly (2.17Gb) is in line with the previously reported C-value (2.3Gb) (Doležalová et al., 2002). The genomic content, such as gene space and repeat content of the genome (~77%), is comparable to cultivated lettuce (Reyes-Chin-Wo et al., 2017). Using SNPs called on the reference genome, population genetic analysis identified three *L. saligna* sub-groups that are consistent with geography Figure 1). We also inferred the graphical origin of two genotypes derived from the Jardin Botanique de Nantes, a French botanical garden, including the accession CGN05271 (found to be of European origin) and accession CGN05282 (found to be of Middle Eastern origin). The obtained population genetics structure is in agreement with a previous clustering based on AFLP markers (Giesbers et al., 2018).

3.2 Inversions and HI may hamper breeding with the NHR8 resistance locus

Comparative genomic analysis identified two large inversions (>50 Mb) on Chromosomes 5 and 8 between *L. saligna* and *L. sativa* (Figure 2). We also found that the inversion on Chromosome 8 co-segregated with an HI region (Giesbers et al., 2019). Genic incompatibilities associated with hybrid necrosis are often linked to immune genes (Bomblies and Weigel, 2007; Fishman and Sweigart, 2018). A well-described example of hybrid necrosis for lettuce is the digenic interaction between the *L. saligna* allele of *Rin4*, encoding a putative negative regulator of basal plant defense, and the resistance gene *Dm*39 from cultivated lettuce (Jeuken et al., 2009). The HI locus on Chromosome 8 was not found to be associated with the hybrid necrosis phenotype (Giesbers et al., 2019). Therefore, immune gene(s) are likely not causal to HI on Chromosome 8, even though we found several resistance loci (MRC and NHR) close to the HI regions located on the inverted regions (Figure 3; Supplemental Figure 8). If the HI/TRD locus indeed resides in the inversion, then further fine mapping and introgression of loci associated with HI, and potential immune genes underlying NHR for that matter, will not be feasible due to the lack of recombination caused by inversion.

3.3 *L. sativa* contains more immune genes than the non-host *L. saligna*

Previous studies in *L. saligna* by genetic mapping have detected multiple loci containing *NLR*s associated with its resistance phenotype, for example, the R locus (*Dm39*) interacts with *Rin4*, and the R locus responds to the effector *BLR31,* which are suggested not to govern the NHR phenotype (Jeuken et al., 2009; Giesbers et al., 2017). The lack of knowledge on genome-wide variation in resistance genes has hindered the identification of NHR determinant(s). In this paper, we comprehensively inventoried *NLR* and *RLK* genes in *L. saligna* and *L. sativa* (Table 3). Our results show that *L. sativa* has more *NLRs* (364 / 323 = 1.13) and *RLK*s (566 / 478 = 1.18) than *L. saligna*. This difference could possibly be due to the genome size differences between *L. sativa* and *L. saligna* (2.5Gb / 2.3Gb = 1.09; Doležalová et al., 2002) or the incomplete sequencing and annotation. Immune genes, like *NLRs* or *RLK*s, are known as the most variable genes in plants, including lettuce and its wild relatives (Karasov et al., 2014; Parra et al., 2016). Due to allelic and copy number variation, the genome assembly alone cannot fully capture the complete spectrum of *R* genes (Barragan and Weigel, 2021). Therefore, the genetic determinant of NHR might not be identified by the genome-wide searches using these reference assemblies. Target sequencing of *NLR*s and *RLK*s (e.g. RenSeq and RLKSeq) can be applied to collect a more complete spectrum of resistance genes (Witek et al., 2016; Lin et al., 2020).

3.4 *RLK* and *NLR* genes associated with NHR against *Bremia*

To further understand the relationship between *RLK*/*NLRs* and NHR in *L. saligna*, we mapped the four NHR loci to the *L. saligna* reference genome. Of the four NHR intervals, we found that three have either elevated densities of *RLKs* (NHR4 & NHR8) or *NLR*s (NHR7.1). Moreover, *RLKs* and *NLRs* do not co-occur with each other in analyzed *Lactuca* species, as illustrated by NHR8 (14 *RLKs* vs zero *NLRs*) and NHR7.1 (zero *RLK*s vs 13 *NLRs*) (Supplemental Table 30), which suggests that *NLR*s and *RLK*s act as epistatic genes explaining NHR (Giesbers et al., 2018). Although *RLKs* and *NLRs* elicit PTI and ETI respectively (Jones and Dangl, 2006), there is increasing evidence that PTI and ETI are not separate phenomena and mutually strengthen each other’s immune response (Yuan et al., 2021; Ngou et al., 2021). This could explain why the identified NHR loci in *L. saligna* involves a combination of PTI and ETI.

3.5 RNA-seq highlights a crucial role of *RLKs*

RNA-seq analysis of *L. saligna* leaves inoculated with *Bremia* enabled us to identify DEGs related to NHR-associated plant defense responses. Multiple DEGs with high levels of induced expression were found to be involved in salicylic acid (SA) synthesis (SARD1 and UGT76B1) or SA-dependent penetration resistance (PEN1 and PEN3) contributing to NHR in *Arabidopsis* (Supplemental Table 27-28) (Zhang et al., 2010; Collins et al., 2003; Assaad et al., 2004; Mohnike et al., 2021; Bauer et al., 2021). Various studies have shown that SA increases *RLK* expression in different plants (Ohtake et al., 2000; Coqueiro et al., 2015). Transcriptome analysis also revealed that a large portion of the DEGs at 8 hpi function in early recognition and defense signaling activity, whereas DEGs at 24 hpi were found to be responsible for post-transcription activity. For expression of immune genes in NHR regions, no *NLR* genes were differentially expressed. This is consistent with expectations, as NLR-encoding genes generally are lowly expressed after *Bremia* infection (Wroblewski et al., 2007). In addition, a large amount of *RLK*s was differentially transcribed, which is similar as described for the interaction between lettuce and the fungal pathogen *Botrytis cinerea* (De Cremer et al., 2013).

3.6 NHR8 contains *WAK* genes highly upregulated upon *Bremia* infection

Among the four NHR regions, NHR8 has the highest number of differentially expressed *RLK*s (Supplemental Table 30). Within it, three closely clustered wall-associated kinases (WAKs) were of special interest because of their significant expression change (Figure 4C). These three WAK paralogs were homologs of *Arabidopsis WAK2,* which is highly expressed in leaves and can be up-regulated in expression upon pathogen infection and SA application (He et al., 1999). Various studies illustrated that WAKs provide quantitative resistance against various diseases in crops such as maize and rice (Zuo et al., 2015; Hurni et al., 2015; Hu et al., 2017). For *L. saligna* infected by *Bremia*, oligogalacturonides derived from damaged cell walls could be perceived by WAKs to trigger PTI (Raaymakers and Van den Ackerveken, 2016; Ferrari et al., 2013; Brutus et al., 2010). WAKs have also been implied in cell wall reinforcement. In rice, Xa4 strengthens the cell wall by promoting cellulose synthesis and suppressing cell wall loosening, thereby enhancing resistance to bacterial infection by *Xanthomonas oryzae* (Hu et al., 2017). Hence, WAKs located on NHR8 seem to hold potential in *L. saligna* resistance. Nevertheless, we cannot rule out the possibility that other genes/factors play roles in NHR in *L. saligna*, and the expressions level of *WAK*s along with other genes mentioned in this paper need to be further compared to their homologs in susceptible lettuce cultivars or resistant introgression lines. Future fine mapping and knock down/out experiments are needed to further pinpoint key factors underlying the NHR in *L. saligna* using the reference genome assembly presented in this paper.

3.7 A model for NHR in *L. saligna* against lettuce downy mildew

Based on our findings and previous research, we propose an NHR model for *L. saligna* with the following three elements: i) The host status of *L. sativa* and *L. saligna* is partly determined by the variation in orthologous RLKs involved in immunity. A specific ortholog in *L. saligna* can effectively enhance resistance to colonization by *B. lactucae*.

A comparable role of orthologous RLKs has been observed in the interaction between barley and leaf rust fungi, in which a LecRK of wild barley quantitatively enhances resistance (Wang et al., 2019). ii) After non-self-recognition by RLKs, cell wall-plasma membrane interactions are strengthened (Wolf, 2017), restricting intercellular hyphal growth. This is in line with the reduced hyphae formation found in infected *L. saligna* (Zhang et al., 2009b). In case of successful penetration, NHR to powdery mildew in barley is often backed up by NLR-mediated hypersensitive response (HR) (reviewed in Niks and Marcel, 2009). As for the observed NHR in *L. saligna*, this might also be the case.

4. Materials and Methods

4.1 Plant materials and DNA isolation

L. saligna accessions selected for whole-genome sequencing and resequencing were obtained from the lettuce germplasm collection of the Centre for Genetic Resources, The Netherlands (CGN) (Supplemental Table 12). Accession CGN05327 was collected from Gerona, Spain, of which a Single Seed Descendant (SSD) was used for *de novo* reference genome sequencing and assembly. Re-sequencing data of 15 Single Seed Decent (SSD) lines derived from *L. saligna* accessions (Supplemental Table 12) were selected to represent the *L. saligna* germplasm. Seeds were stratified at 4°C for three days to improve germination. Seedlings were subsequently grown in a growth chamber at 17–19°C with LED light under a 16 h photoperiod and a relative humidity of 75–78%. After eight weeks, plants were transplanted to larger pots containing potting soil and grown under greenhouse conditions. Images of leaves (third mature leaf counted from the base) of 10 accessions belonging to different subgroups were taken from 15-week-old plants, which were grown in triplicate (Supplemental Table 32). Tissue sampling was performed when plants were close to bolting, and DNA was extracted using the protocol as in Ferguson *et al.* (2020).

4.2 Genome sequencing

A *de novo* genome assembly of *L. saligna* CGN05327 (Supplemental Figure 1) was assembled using a ~21-fold coverage of long-read data generated by PacBio Sequel technology (4,083,751 reads; N50 read length=16,581 bp; subread length=8,514 bp), and a ~175-fold coverage of Paired-end (PE) reads obtained by Illumina mate pair sequencing. The mate pair library was prepared using different insert sizes and read lengths: HiSeq (200 bp insert size, 125 bp PE), HiSeq (500 bp insert size, 125 bp PE), and MiSeq (550 bp insert size, 300 bp PE).

4.3 Genome assembling and scaffolding

PacBio reads were assembled using Canu and polished with Pilon (v1.20) using Illumina data (Koren et al., 2017; Walker et al., 2014). Subsequently, multiple techniques were applied to elevate the contiguity of the assembly. A follow-up assembly (version 1) was scaffolded using 10x Genomics Chromium barcoding data (ARC pipeline) and ~130-fold coverage BioNano optical mapping data (Yeo et al., 2017). A Hi-C library produced by Dovetail Genomics providing ~2,553-fold coverage of sequence data (429 million 2x150 bp read pairs) was used for *in vitro* proximity ligation. Mis-joins in assembled contigs were corrected using the HiRise pipeline, resulting into genome assembly v2 (Putnam et al., 2016).

4.4 Assembly reconstruction by syntenic and genetic makers

ALLMAPS was applied to reconstruct scaffolds of the *L. saligna* v2 assembly to chromosomal linkage groups using two types of markers: 417 genetic markers (weight = 2) derived from F_2 (*L. saligna* CGN05271 x *L. sativa* cv. Olof), and syntenic markers (weight = 1) derived from the reciprocal best hits between *L. saligna* v2 and *L. sativa* v8 (Jeuken et al., 2001; Giesbers et al., 2019; Tang et al., 2015b). Contigs (>1 kb) not clustered in chromosomes were concatenated by JCVI with 100 N-content gaps to generate a virtual “chromosome zero” storing left genetic content (Tang et al., 2015a).

4.5 Genome size estimation

Two paired-end Illumina libraries of *L. saligna* were used for genome size estimation (~117 Gb pairs; ~932 million reads) using a k-mer count size of 23 (Supplemental Table 1). Jellyfish v2.3.0 was used to count the k-mer frequency (Marçais and Kingsford, 2011). Jellyfish output was used by GenomeScope (v2.0) to estimate haploid genome length, percentage of repetitive DNA, and heterozygosity of the *L. saligna* genome using the histogram file (Vurture et al., 2017).

4.6 Genome completeness assessment

Completeness of the *L. saligna* genome assembly was evaluated using multiple approaches. BUSCO (v3.0.2) assessment was conducted using the eudicotyledons_odb10 database (Simão et al., 2015). In addition, 226,910 ESTs of diverse *Lactuca* species (retrieved on July 2019 by NCBI) were aligned to the genome using GMAP (version 2019-06-10) (Wu and Watanabe, 2005). For GMAP alignment, presence/absence of ESTs was determined after filtering the alignments by identity and coverage at different levels of stringency using custom scripts.

4.7 Repeat annotation

Tandem Repeats Finder v4.04 was used to detect tandem repeats using the following parameters: Match=2, Mismatch=7, Delta= 7, PM=80, PI=10, Minscore=50, and MaxPeriod=2000 (Benson, 1999). TEs were searched using RepeatMasker v4.0.7 against ortholog and *de novo* databases in a serial order (Smit et al., 2019): i.e., by using orthology data from Repbase and Dfam (version 20170127), and *de novo* TEs library generated by RepeatModeler v2.0 and MITE-Hunter (Han and Wessler, 2010; Jurka et al., 2005; Hubley et al., 2015; Price et al., 2005). Perl tool "One code to find them all" was used to parse and quantify the number and position of predicted repeat elements (Bailly-Bechet et al., 2014).

4.8 Non-coding RNA annotation

Non-coding RNA (ncRNA) loci were annotated according to different types. tRNAscan-SE v2.0.4 was used to annotate tRNAs using eukaryote parameters (Lowe and Eddy, 1997). In addition, rRNA was annotated using RNAmmer v1.2 (Lagesen et al., 2007). INFERNAL v1.1.2 was used to search against the Rfam database (release 14.1) to detect additional miRNA, snRNA, tRNA, rRNA, and snoRNA sequences (Kalvari et al., 2018; Nawrocki and Eddy, 2013). Annotations predicted by different tools were merged and condensed using GenomicRanges in R v3.6 (Lawrence et al., 2013).

4.9 Infection assays

Leaves of three-week old *L. saligna* plants (accession CGN05327) were spray-inoculated with a spore suspension of *B. lactucae* race Bl:21 (2.0*10^5 conidiospores/mL) or with sterile water. Treated plants were first kept in the dark for 4 h to maximize spore germination, and then incubated in a growth chamber at 15°C and a 16/8 h (day/night) photoperiod. Leaf samples of inoculated and mock-treated leaves were collected at 8 and 24 hpi. Leaf samples from the three biological replicates were immediately frozen in liquid nitrogen and stored in-80°C until further use.

4.10 RNA library preparation and sequencing

Total RNA was isolated from 12 infection assay samples and one pooled sample consisting of root and flower bud material (pooled from different floral stages) using a Direct Zol RNA Miniprep Plus kit (Zymo Research) followed by DNAse treatment. RNA was purified by ethanol precipitation. Concentration and purity of RNA samples was measured with a Nanodrop 2000c spectrophotometer and a Qubit 4.0 fluorometer using a RNA Broad Range assay (Thermo Fisher Scientific). Paired-End sequencing (2 x 125 bp) was performed on an Illumina HiSeq2500 platform using two flow cell lanes.

4.11 Gene prediction

Gene models for protein-coding genes were annotated by combining *ab initio* prediction and homology-based annotation. First, BRAKER was used to train an Augustus model with RNA-seq data to predict genes *ab initio* (Hoff et al., 2016). Thereafter, MAKER was applied to integrate the *ab initio* prediction with extrinsic evidence: i.e., *de novo* transcripts assembled by Trinity and protein homology data (Holt and Yandell, 2011; Grabherr et al., 2011). Annotation-edit-distance (AED) calculated by MAKER was used to examine the quality of the genome annotation. Coding-potential was calculated by CPC2 (Kang et al., 2017) to further filter out non-coding transcripts (Supplemental Data 4).

4.12 Functional annotation

Potential biological function of proteins was inferred using three criteria: i) best-hit matches in SwissProt, TrEMBL, and *A. thaliana* Araport11 databases using BLAST v2.2.31 and DIAMOND (Buchfink et al., 2015) (E-value cut-off = 1e-5); ii) protein domains/motifs identified by InterProscan against the Pfam protein database (El-Gebali et al., 2018; Zdobnov and Apweiler, 2001); and iii) gene ontology (GO) based on InterPro entries. Orthology searches for pathway analysis were conducted with Kofamscan (Aramaki et al., 2019) using a customized HMM database of KEGG Orthologs (Kanehisa, 2000).

4.13 Resequencing and SNPs calling

Libraries of PE reads (2 x 150 bp, insert size distribution peaks at 190 bp) were constructed and sequenced. Re-sequencing reads were mapped to the *de novo* genome assembly using the BWA alignment tool (Li, 2013). After mapping, the alignment output in SAM format was translated to the BAM format using SAMtools (Li et al., 2009). Duplicated reads were marked, and read groups were assigned to the remaining reads using the tools built into GATK v4.0.8.1 (Van der Auwera et al., 2013). Subsequently, HaplotypeCaller and GentypeGVCFs were applied to call variants (SNPs and indels, respectively) per sample and used to perform joint genotyping. These results were used to generate a vcf file containing all raw SNPs and indels. SelectVariants and VariantFiltration tools in GATK were used to extract biallelic SNPs, which were subjected to hard-filtering for low-quality SNPs based on several scores (QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, and SOR > 3.0). The distribution of each of the quality scores and their cut-offs was visualized in R (Supplemental Figure 11). Subsequently, SNPs were further filtered by Minor Allele Frequency (MAF) and the missing rate for each SNP (MAF < 0.05, missing rate > 0.1) for downstream analysis. Lastly, the filtered SNP call-set was annotated with SnpEff v4.3 using default settings to predict the nucleotide change effect of every SNP (Cingolani et al., 2012). Generated read data have been deposited in the European Nucleotide Achieve (ENA) under reference number PRJEB36060.

4.14 Population structure analysis

PLINK2 was used to prune the SNP dataset to reduce the redundancy caused by linkage disequilibrium (LD) analysis for different downstream analyzes (Purcell et al., 2007). Firstly, SNPhylo was used to construct a maximum likelihood (ML) phylogenetic tree using 210,358 SNPs with default settings and 1,000 bootstrap replicates (window size = 50 SNPs, sliding size = 10, LD < 0.1) (Lee et al., 2014). Secondly, a PCA was conducted using PLINK2 on the pruned dataset of 904,930 SNPs (window size = 50 SNPs, sliding size = 10 SNPs, LD < 0.5). K-means clustering was performed via Eigen decomposition for the

PCA and visualized in R. ADMIXTURE (v1.3.0) was used to deduce ancestral history and population structure using 96,804 SNPs (window size = 50 SNPs, sliding size = 10 SNPs, LD < 0.05) (Alexander and Lange, 2011). ADMIXTURE was utilized to determine the best number of ancestral populations (K = 1 to 4) by cross-validation errors (Supplemental Figure 12), and then was run again with the best K value with 1,000 bootstrap replicates to infer population structure. Population structure results were summarized using GISCO geographical information and visualized in R (Wickham, 2016; Eurostat GISCO, 2006).

4.15 Comparative genomic analysis

The longest representative transcripts were selected from *L. saligna* and *L. sativa* as the basis for synteny analysis. BLAST (v2.2.31) was used to search homologous gene pairs between both species. MCScanX was employed to detect syntenic blocks (E-value cut-off = 1e-5, collinear block size ≥ 5) between the two *Lactuca* species using the top five alignment hits, which were visualized using SynVisio and JCVI (Tang et al., 2015a; Wang et al., 2012; Bandi and Gutwin, 2020). A separate synteny plot was created using the best-matching hit to remove noise from polyploidy and translocation events. Genetic markers on Chromosomes 5 and 8 were collected, and dot plots coordinated by genetic and physical positions were visualized with R v3.6.1 to validate inversions detected by synteny analysis. To assess the influence of genomic inversions, syntenic gene pairs located at inversion borders were searched against the *A. thaliana* protein database Araport11 (https://www.arabidopsis.org).

4.16 NLR identification and classification

Genome-wide searches to identify *NLRs* were conducted using the genomes of *L. saligna* (v4) and *L. sativa* (v8) (downloaded from the CoGe website; gid35223). HMMER was used to search Hidden Markov Models (HMMs) for structural domains of NLRs (E-value cut-off = 1e-10). The Pfam models used were PF00931.23 and NBS_712. hmm for the NB domain, PF01582.20 and PF13676.6 for TIR, PF18052.1 for CC, and eight HMMs for the LRR domain (PF00560.33, PF07723.13, PF07725.13, PF12799.7, PF13306.6, PF13516.6, PF13855.6, PF14580.6). NB domains identified by InterProScan (see Functional annotation section) and CC motifs predicted by Paircoil2 (McDonnell et al., 2006) (P scores < 0.025) were integrated with the HMMER output. *NLRs* of *L. saligna* were classified into different categories (TNL/CNL and RGC families) by phylogeny clustering using *NLR*s previously identified in the *L. sativa* v8 genome (Christopoulou et al., 2015b). RGC families with a >3 count difference and 1.5 ratio between two species were selected as families with major differences. Amino acid sequences of NB domains were aligned with HmmerAlign (Finn et al., 2011). The alignment was trimmed by trimAl using the '-gappyout' algorithm, retaining 1,367 residues for phylogeny construction (Capella-Gutiérrez et al., 2009). The best-hit model of evolution, Blosum62+F+R10, was first selected by IQ-TREE v1.6.12 and ML trees were inferred with IQ-TREE (Nguyen et al., 2015). IQTREE (-pers 0.1, -nm 500) was run independently 10 times with 1,000 ultrafast bootstrap (UFBoot) replicates. Finally, the 10 best ML trees inferring the tree with highest log-likelihood was selected for NLR classification. Phylogenetic trees were visualized and annotated using iTOL v6 (Letunic and Bork, 2021).

4.17 Identification of *NLR* clusters

Annotated *NLR* genes were used to determine gene intervals of MRCs on the *L. saligna* genome. The syntenic regions of MRCs in *L. saligna* were named lsal-MRCs to distinguish them from those detected in the *L. sativa* genome. An additional sliding window search was performed to identify *NLR* clusters (NCs) containing more than five *NLR*s (maximum10-genes gap). Identified MRCs and NCs were visualized on the *L. saligna* genome using Circos (Krzywinski et al., 2009).

4.18 RLK identification and classification

Sequence similarity searches against primary protein sequences were performed with HMMER v3.1 using the PKinase alignment file (PF00069; E-value cut-off = 1e-10). Obtained protein sequences were subsequently scanned for the presence of extracellular domains using HMMER (E-value cut-off = 1e-3; Supplemental Table 23). TMHMM2.0 and SCAMPI2 were used to detect transmembrane regions (Krogh et al., 2001; Peters et al., 2016).

4.19 Mapping NHR and HI regions

Genetic markers previously used in assembly reconstruction were aligned to genome assembly via BLAST v2.2.31 to locate the HI and NHRs regions in *L. saligna*. The genomic positions of one HI and four NHR regions were subsequently plotted on the *L. saligna* genome using Circos.

4.20 RNA-seq analysis

Raw RNA-seq reads were quantified on *L. saligna* transcripts using Kallisto (v.0.44.0) to gain normalized transcript per million (TPM). Transcripts with a TPM value below 0.1 were considered not expressed. Then, DESeq2 was used to normalize the read count for each gene (total read count > 3) and execute statistical analyses to determine the DEGs with padj < 0.05 and $\log_2$FC > 1 (Love et al., 2014). Next, the read count mean and SD of infected and mock samples were calculated for all DEGs. Metascape was used for enrichment analysis of up-regulated genes and to render protein–protein interaction networks in Cytoscape (Zhou et al., 2019; Shannon et al., 2003). To identify potential candidate genes in the four NHR regions, additional counting for DE *RLK* and *NLR* genes, and highly regulated genes (|Log2FC| > 3) were counted separately.

5. Data availability

The genome assembly described in this paper, *L. saligna* v4, is available under the BioProject PRJEB56287. All raw sequencing reads have been deposited in the ENA database under BioProject PRJEB56288. This includes the Illumina, PacBio, 10x Genomics, Bionano and Hi-C whole-genome sequences as well as RNA sequencing data for genome annotation and statistical analysis of *Bremia*-infection assay. The resequencing data for 15 *L. saligna* accessions are deposited under the BioProject PRJEB36060, which contains data of 100 *Lactuca* accessions derived from the TKI-100 project.

6. Acknowledgements

This research was supported by a grant of the International Lettuce Genomics Consortium (ILGC) funded by the Top Consortium for Knowledge and Innovation Horticultural and Starting Materials (grant number 1406-039). W.X. was financially supported by a fellowship by the China Scholarship Council (CSC). We thank Rens Holmer, Zhaodong Hao, Deedi Sogbohossou, Nora Walden, Sandra Smit, and Henri van de Geest for their support with bioinformatic analysis and genome sequencing data. We would also like to thank Wenhao Li for his help with statistics and R programming. We also thank Elizabeth Marie Georgian for writing support.

7. References

Alexander, D.H. and Lange, K. (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12: 246.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2019) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics 36: 2251–2252.

Assaad, F.F., Qiu, J.L., Youngs, H., Ehrhardt, D., Zimmerli, L., Kalde, M., Wanner, G., Peck, S.C., Edwards, H., Ramonell, K., Somerville, C.R., and Thordal-Christensen, H. (2004) The PEN1 syntaxin defines a novel cellular compartment upon fungal attack and is required for the timely assembly of papillae. Mol. Biol. Cell 15: 5118–5129.

Van der Auwera, G.A. et al. (2013) From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr. Protoc. Bioinforma. 43: 11.10.1-11.10.33.

Bailly-Bechet, M., Haudry, A., and Lerat, E. (2014) “One code to find them all”: A perl tool to conveniently parse RepeatMasker output files. Mob. DNA 5: 1–15.

Bandi, V. and Gutwin, C. (2020) Interactive exploration of genomic conservation. In Proceedings of the 46th Graphics Interface Conference on Proceedings of Graphics Interface 2020 (GI’20) (Waterloo).

Barragan, A.C. and Weigel, D. (2021) Plant NLR diversity: The known unknowns of pan-NLRomes. Plant Cell 33: 814–831.

Bateson, W. (1909) Heredity and variation in modern Lights. In Darwin and Modern Science, A.C. Seward, ed (Cambridge University Press: Cambridge), pp. 85–101.

Bauer, S., Mekonnen, D.W., Hartmann, M., Yildiz, I., Janowski, R., Lange, B., Geist, B., Zeier, J., and Schäffner, A.R. (2021) UGT76B1, a promiscuous hub of small molecule-based immune signaling, glucosylates N-hydroxypipecolic acid, and balances plant immunity. Plant Cell 33: 714–734.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27: 573–580.

Bettgenhaeuser, J., Gilbert, B., Ayliffe, M., and Moscou, M.J. (2014) Nonhost resistance to rust pathogens – A continuation of continua. Front. Plant Sci. 5: 664.

den Boer, E., Pelgrom, K.T.B., Zhang, N.W., Visser, R.G.F., Niks, R.E., and Jeuken, M.J.W. (2014) Effects of stacked quantitative resistances to downy mildew in lettuce do not simply add up. Theor. Appl. Genet. 127: 1805–1816.

Bonnier, F.J.M., Reinink, K., and Groenwold, R. (1991) New sources of major gene resistance in *Lactuca* to *Bremia lactucae*. Euphytica 61: 203–211.

Brutus, A., Sicilia, F., Macone, A., Cervone, F., and De Lorenzo, G. (2010) A domain swap approach reveals a role of the plant wall-associated kinase 1 (WAK1) as a receptor of oligogalacturonides. Proc. Natl. Acad. Sci. 107: 9452–9457.

Buchfink, B., Xie, C., and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12: 59–60.

Coqueiro, D.S.O., de Souza, A.A., Takita, M.A., Rodrigues, C.M., Kishi, L.T., and Machado, M.A. (2015). Transcriptional profile of sweet orange in response to chitosan and salicylic acid. BMC Genomics 16: 1–14.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009) TrimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.

Cheng, Z., Sattler, S., Maeda, H., Sakuragi, Y., Bryant, D.A., and DellaPenna, D. (2003) Highly divergent methyltransferases catalyze a conserved reaction in tocopherol and plastoquinone synthesis in cyanobacteria and photosynthetic eukaryotes. Plant Cell 15: 2343–2356.

Chisholm, S.T., Coaker, G., Day, B., and Staskawicz, B.J. (2006) Host-microbe interactions: Shaping the evolution of the plant immune response. Cell 124: 803–814.

Christopoulou, M., McHale, L.K., Kozik, A., Wo, S.R.C., Wroblewski, T., and Michelmore, R.W. (2015a) Dissection of two complex clusters of resistance genes in lettuce (*Lactuca sativa*). Mol. Plant-Microbe Interact. 28: 751–765.

Christopoulou, M., Wo, S.R.C., Kozik, A., McHale, L.K., Truco, M.J., Wroblewski, T., and Michelmore, R.W. (2015b) Genome-wide architecture of disease resistance genes in lettuce. G3 Genes, Genomes, Genet. 5: 2655–2669.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 6: 80–92.

Collins, N.C., Thordal-Christensen, H., Lipka, V., Bau, S., Kombrink, E., Qiu, J.-L., Hückelhoven, R., Stein, M., Freialdenhoven, A., Somerville, S.C., and Schulze-Lefert, P. (2003) SNARE-protein-mediated disease resistance at the plant cell wall. Nature 425: 973–977.

De Cremer, K., Mathys, J., Vos, C., Froenicke, L., Michelmore, R.W., Cammue, B.P.A., and De Coninck, B. (2013) RNAseq-based transcriptome analysis of *Lactuca sativa* infected by the fungal necrotroph *Botrytis cinerea*. Plant, Cell Environ. 36: 1992–2007.

Ding, P., Rekhter, D., Ding, Y., Feussner, K., Busta, L., Haroth, S., Xu, S., Li, X., Jetter, R., Feussner, I., and Zhang, Y. (2016) Characterization of a pipecolic acid biosynthesis pathway required for systemic acquired resistance. Plant Cell 28: 2603–2615.

Dobzhansky, T. (1934) Studies on hybrid sterility- I. Spermatogenesis in pure and hybrid *Drosophila pseudoobscura*. Zeitschrift für Zellforsch. und Mikroskopische Anat. 21: 169–223.

Doležalová, I., Lebeda, A., Janeček, J., Číhalíková, J., Křístková, E., and Vránová, O. (2002) Variation in chromosome numbers and nuclear DNA contents in genetic resources of *Lactuca* L. species (Asteraceae). Genet. Resour. Crop Evol. 49: 383–395.

El-Gebali, S. et al. (2018) The Pfam protein families database in 2019. Nucleic Acids Res. 47: D427–D432.

Eurostat GISCO (2006) Administrative boundaries of NUTS 2006 and COAS 2006.

FAOSTAT (2019) Food and Agriculture Organization (FAO) of the United Nations. http://www.fao.org/faostat/en/#data (May 18, 2019)

Ferrari, S., Savatin, D.V., Sicilia, F., Gramegna, G., Cervone, F., and De Lorenzo, G. (2013) Oligogalacturonides: Plant damage-associated molecular patterns and regulators of growth and development. Front. Plant Sci. 4: 49.

Finn, R.D., Clements, J., and Eddy, S.R. (2011) HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res. 39: W29–W37.

Fishman, L. and McIntosh, M. (2019) Standard deviations: The biological bases of transmission ratio distortion. Annu. Rev. Genet. 53: 347–372.

Fishman, L. and Sweigart, A.L. (2018) When two rights make a wrong: The evolutionary genetics of plant hybrid incompatibilities. Annu. Rev. Plant Biol. 69: 707–731.

Franck, C.M., Westermann, J., Bürssner, S., Lentz, R., Lituiev, D.S., and Boisson-Dernier, A. (2018) The protein phosphatases ATUNIS1 and ATUNIS2 regulate cell wall integrity in tip-growing cells. Plant Cell 30: 1906–1923.

Giesbers, A.K.J., den Boer, E., Braspenning, D.N.J., Bouten, T.P.H., Specken, J.W., van Kaauwen, M.P.W., Visser, R.G.F., Niks, R.E., and Jeuken, M.J.W. (2018) Bidirectional backcrosses between wild and cultivated lettuce identify loci involved in nonhost resistance to downy mildew. Theor. Appl. Genet. 131: 1761–1776.

Giesbers, A.K.J., den Boer, E., Ulen, J.J.W.E.H., van Kaauwen, M.P.W., Visser, R.G.F., Niks, R.E., and Jeuken, M.J.W. (2019) Patterns of transmission ratio distortion in interspecific lettuce hybrids reveal a sex-independent gametophytic barrier. Genetics 211: 263–276.

Giesbers, A.K.J., Pelgrom, A.J.E., Visser, R.G.F., Niks, R.E., van den Ackerveken, G., and Jeuken, M.J.W. (2017) Effector-mediated discovery of a novel resistance gene against *Bremia lactucae* in a nonhost lettuce species. New Phytol. 216: 915–926.

González-Lamothe, R., Tsitsigiannis, D.I., Ludwig, A.A., Panicot, M., Shirasu, K., and Jones, J.D.G. (2006) The U-box protein CMPG1 is required for efficient activation of defense mechanisms triggered by multiple resistance genes in tobacco and tomato. Plant Cell 18: 1067–1083.

Grabherr, M.G. et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29: 644–652.

Han, Y. and Wessler, S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. 38: e199–e199.

He, Z-H, Cheeseman, I., He, D and Kohorn, B. (1999) A cluster of five cell wall-associated receptor kinase genes, Wak1–5, are expressed. Plant Mol. Biol. 39: 1189–1196.

Heath, M.C. (1981) Nonhost resistance. Plant Dis. Control: 201–217.

Hok, S., Danchin, E.G.J., Allasia, V., Panabiéres, F., Attard, A., and Keller, H. (2011). An *Arabidopsis* (malectin-like) leucine-rich repeat receptor-like kinase contributes to downy mildew disease. Plant. Cell Environ. 34: 1944–1957.

Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016) BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32: 767–769.

Hoffmann, A.A. and Rieseberg, L.H. (2008) Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? Annu. Rev. Ecol. Evol. Syst. 39: 21-42.

Holt, C. and Yandell, M. (2011) MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12: 491.

Hu, K. et al. (2017) Improvement of multiple agronomic traits by a disease resistance gene via cell wall reinforcement. Nat. Plants 3: 17009.

Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F.A., and Wheeler, T.J. (2015) The Dfam database of repetitive DNA families. Nucleic Acids Res. 44: D81–D89.

Hurni, S., Scheuermann, D., Krattinger, S.G., Kessel, B., Wicker, T., Herren, G., Fitze, M.N., Breen, J., Presterl, T., Ouzunova, M., and Keller, B. (2015) The maize disease resistance gene *Htn1* against northern corn leaf blight encodes a wall-associated receptor-like kinase. Proc. Natl. Acad. Sci. U. S. A. 112: 8780–5.

Iorizzo, M. et al. (2016) A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat. Genet. 48: 657–666.

Jeuken, M. and Lindhout, P. (2002) *Lactuca saligna*, a non-host for lettuce downy mildew (*Bremia lactucae*), harbors a new race-specific *Dm* gene and three QTLs for resistance. TAG Theor. Appl. Genet. 105: 384–391.

Jeuken, M., van Wijk, R., Peleman, J., and Lindhout, P. (2001) An integrated interspecific AFLP map of lettuce (*Lactuca*) based on two *L. sativa* × *L. saligna* F2 populations. Theor. Appl. Genet. 103: 638–647.

Jeuken, M.J.W. and Lindhout, P. (2004) The development of lettuce backcross inbred lines (BILs) for exploitation of the *Lactuca saligna* (wild lettuce) germplasm. Theor. Appl. Genet. 109: 394–401.

Jeuken, M.J.W., Zhang, N.W., McHale, L.K., Pelgrom, K., Den Boer, E., Lindhout, P., Michelmore, R.W., Visser, R.G.F., and Niks, R.E. (2009) *Rin4* causes hybrid necrosis and race-specific resistance in an interspecific lettuce hybrid. Plant Cell 21: 3368–3378.

Jones, J.D.G. and Dangl, J.L. (2006) The plant immune system. Nature 444: 323–329.

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

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., and Petrov, A.I. (2018) Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 46: D335–D342.

Kanehisa, M. (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28: 27–30.

Kang, Y.J., Yang, D.C., Kong, L., Hou, M., Meng, Y.Q., Wei, L., and Gao, G. (2017) CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 45: W12–W16.

Karasov, T.L., Horton, M.W., and Bergelson, J. (2014) Genomic variability as a driver of plant–pathogen coevolution? Curr. Opin. Plant Biol. 18: 24–30.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017) Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. 27: 722–736.

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J. Mol. Biol. 305: 567–580.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009) Circos: An information aesthetic for comparative genomics. Genome Res. 19: 1639–1645.

Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.-H.H., Rognes, T., and Ussery, D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35: 3100–3108.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013) Software for computing and annotating genomic ranges. PLoS Comput. Biol. 9: e1003118.

Lebeda, A., Doležalová, I., Křístková, E., Kitner, M., Petrželová, I., Mieslerová, B., and Novotná, A. (2009) Wild *Lactuca* germplasm for lettuce breeding: Current status, gaps and challenges. Euphytica 170: 15–34.

Lebeda, A., Křístková, E., Doležalová, I., Kitner, M., and Widrlechner, M.P. (2019) Wild *Lactuca* species in North America. In North American Crop Wild Relatives, Volume 2, pp. 131–194.

Lebeda, A., Křístková, E., Kitner, M., Mieslerová, B., Jemelková, M., and Pink, D.A.C. (2014) Wild *Lactuca* species, their genetic diversity, resistance to diseases and pests, and exploitation in lettuce breeding. Eur. J. Plant Pathol. 138: 597–640.

Lebeda, A. and Reinink, K. (1994) Histological characterization of resistance in *Lactuca saligna* to lettuce downy mildew (*Bremia lactucae*). Physiol. Mol. Plant Pathol. 44: 125–139.

Lebeda, A., Ryder, E.J., Grube, R., Doležalová, I., and Křístková, E. (2007) Lettuce (Asteraceae; *Lactuca* spp). In Genetic Resources, Chromosome Engineering, and Crop Improvement: Vegetable Crops, R. J. Singh, ed (CRC), pp. 377–472.

Lee, T.H., Guo, H., Wang, X., Kim, C., and Paterson, A.H. (2014) SNPhylo: A pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 15: 1–6.

Letunic, I. and Bork, P. (2021) Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 49: W293–W296.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr. arXiv: 1303.3997.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1000 Genome Project Data Processing (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Lin, X., Armstrong, M., Baker, K., Wouters, D., Visser, R.G.F., Wolters, P.J., Hein, I., and Vleeshouwers, V.G.A.A. (2020) RLP/K enrichment sequencing; a novel method to identify receptor-like protein (*RLP*) and receptor-like kinase (*RLK*) genes. New Phytol. 227: 1264–1276.

Love, M.I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15: 1–21.

Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: A Program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25: 955–964.

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

McDonnell, A. V, Jiang, T., Keating, A.E., and Berger, B. (2006) Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics 22: 356–358.

McHale, L., Tan, X., Koehl, P., and Michelmore, R.W. (2006) Plant NBS-LRR proteins: Adaptable guards. Genome Biol. 7: 212.

McHale, L.K., Truco, M.J., Kozik, A., Wroblewski, T., Ochoa, O.E., Lahre, K.A., Knapp, S.J., and Michelmore, R.W.

(2009) The genomic architecture of disease resistance in lettuce. Theor. Appl. Genet. 118: 565–580.

Meyers, B.C., Kozik, A., Griego, A., Kuang, H., and Michelmore, R.W. (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant Cell 15: 809–834.

Mohnike, L., Rekhter, D., Huang, W., Feussner, K., Tian, H., Herrfurth, C., Zhang, Y., and Feussner, I. (2021) The glycosyltransferase UGT76B1 modulates N-hydroxy-pipecolic acid homeostasis and plant immunity. Plant Cell 33: 735–749.

Muller, H.J. (1942) Isolating mechanisms, evolution, and temperature. Biol. Symp. 6: 71–125.

Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29: 2933–2935.

Netzer, D., Globerson, D., and Sacks, J. (1976). *Lactuca saligna* L., a new source of resistance to downy mildew (*Bremia lactucae* Regel). HortScience 11: 612–613.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32: 268–274.

Niks, R.E. (1987) Nonhost plant species as donors for resistance to pathogens with narrow host range I. Determination of nonhost status. Euphytica 36: 841–852.

Niks, R.E. and Marcel, T.C. (2009) Nonhost and basal resistance: How to explain specificity? New Phytol. 182: 817–828.

Norwood, J.M., Johnson, A.G., and Crute, I.R. (1981). The utilization of novel sources of resistance to *Bremia lactucae* from wild *Lactuca* species. Euphytica 30: 659–668.

Ohtake, Y., Takahashi, T., and Komeda, Y. (2000). Salicylic acid induces the expression of a number of receptor-like kinase genes in *Arabidopsis thaliana*. Plant Cell Physiol. 41: 1038–1044.

Panstruga, R. and Moscou, M.J. (2020) What is the molecular basis of nonhost resistance? Mol. Plant-Microbe Interact. 33: 1253–1264.

Parra, L., Maisonneuve, B., Lebeda, A., Schut, J., Christopoulou, M., Jeuken, M., McHale, L., Truco, M.J., Crute, I., and Michelmore, R. (2016) Rationalization of genes for resistance to *Bremia lactucae* in lettuce. Euphytica 210: 309–326.

Peters, C., Tsirigos, K.D., Shu, N., and Elofsson, A. (2016) Improved topology prediction using the terminal hydrophobic helices rule. Bioinformatics 32: 1158–1162.

Petrželová, I., Lebeda, A., and Beharav, A. (2011) Resistance to *Bremia lactucae* in natural populations of *Lactuca saligna* from some Middle Eastern countries and France. Ann. Appl. Biol. 159: 442–455.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005) *De novo* identification of repeat families in large genomes. Bioinformatics 21: i351–i358.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.

Putnam, N.H., Connell, B.O., Stites, J.C., Rice, B.J., Hartley, P.D., Sugnet, C.W., Haussler, D., and Rokhsar, D.S. (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26: 342–350.

R Core Team (2013) R: A Language and environment for statistical computing.

Raaymakers, T.M. and Van den Ackerveken, G. (2016) Extracellular recognition of oomycetes during biotrophic infection of plants. Front. Plant Sci. 7: 906.

Reyes-Chin-Wo, S. et al. (2017) Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. Nat. Commun. 8: 14953.

Schulze-Lefert, P. and Panstruga, R. (2011) A molecular evolutionary concept connecting nonhost resistance, pathogen host range, and pathogen speciation. Trends Plant Sci. 16: 117–125.

Sedlářová, M., Luhová, L., Petřivalský, M., and Lebeda, A. (2007) Localisation and metabolism of reactive oxygen species during *Bremia lactucae* pathogenesis in *Lactuca sativa* and wild *Lactuca* spp. Plant Physiol. Biochem. 45: 607–616.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 13: 2498–2504.

Shen, K.A., Chin, D.B., Arroyo-Garcia, R., Ochoa, O.E., Lavelle, D.O., Wroblewski, T., Meyers, B.C., and Michelmore, R.W. (2002) *Dm3* is one member of a large constitutively expressed family of nucleotide binding site-leucine-rich repeat encoding genes. Mol. Plant-Microbe Interact. 15: 251–261.

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

Smit, A., Hubley, R., and Green, P. (2019) 2013–2015. RepeatMasker Open-4.0.

Takken, F.L.W. and Goverse, A. (2012) How to build a pathogen detector: Structural basis of NB-LRR function. Curr. Opin. Plant Biol. 15: 375–384.

Tang, H., Krishnakumar, V., and Li, J. (2015a) jcvi: JCVI utility libraries. Zenodo.

Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J.C., Schnable, P.S., Lyons, E., and Lu, J. (2015b) ALLMAPS: Robust scaffold ordering based on multiple maps. Genome Biol. 16: 3.

Telenius, Håk., Ponder, B.A.J., Tunnacliffe, A., Pelmear, A.H., Carter, N.P., Ferguson-Smith, M.A., Behmel, A., Nordenskjöld, M., and Pfragner, R. (1992) Cytogenetic analysis by chromosome painting using dop-pcr amplified flow-sorted chromosomes. Genes, Chromosom. Cancer 4: 257–263.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., and Schatz, M.C. (2017) GenomeScope: Fast reference-free genome profiling from short reads. Bioinformatics 33: 2202–2204.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014) Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963.

Wang, Y., Cordewener, J.H.G., America, A.H.P., Shan, W., Bouwmeester, K., and Govers, F. (2015). Arabidopsis lectin receptor kinases LecRK-IX.1 and LecRK-IX.2 are functional analogs in regulating *Phytophthora* resistance and plant cell death. Mol. Plant-Microbe Interact. 28: 1032–1048.

Wang, Y., Subedi, S., de Vries, H., Doornenbal, P., Vels, A., Hensel, G., Kumlehn, J., Johnston, P.A., Qi, X., Blilou, I., Niks, R.E., and Krattinger, S.G. (2019) Orthologous receptor kinases quantitatively affect the host status of barley to leaf rust fungi. Nat. Plants 5: 1129–1135.

Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., Guo, H., Kissinger, J.C., and Paterson, A.H. (2012) MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40: e49–e49.

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).

Witek, K., Jupe, F., Witek, A.I., Baker, D., Clark, M.D., and Jones, J.D.G. (2016) Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing. Nat. Biotechnol. 34: 656–660.

Wolf, S. (2017) Plant cell wall signalling and receptor-like kinases. Biochem. J. 474: 471–492.

Wroblewski, T., Piskurewicz, U., Tomczak, A., Ochoa, O., and Michelmore, R.W. (2007) Silencing of the major family of NBS-LRR-encoding genes in lettuce results in the loss of multiple resistance specificities. Plant J. 51: 803–818.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859–1875.

Yuan, M., Jiang, Z., Bi, G., Nomura, K., Liu, M., Wang, Y., Cai, B., Zhou, J.M., He, S.Y., and Xin, X.F. (2021). Pattern-recognition receptors are required for NLR-mediated plant immunity. Nature 592: 105–109.

Yeo, S., Coombe, L., Warren, R.L., Chu, J., and Birol, I. (2017) ARCS: Scaffolding genome drafts with linked reads. Bioinformatics 34: 725–731.

Zdobnov, E.M. and Apweiler, R. (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847–848.

Zhang, N.W., Lindhout, P., Niks, R.E., and Jeuken, M.J.W. (2009a) Genetic dissection of *Lactuca saligna* nonhost resistance to downy mildew at various lettuce developmental stages. Plant Pathol. 58: 923–932.

Zhang, N.W., Pelgrom, K., Niks, R.E., Visser, R.G.F.F., and Jeuken, M.J.W. (2009b) Three combined quantitative trait loci from nonhost *Lactuca saligna* are sufficient to provide complete resistance of lettuce against *Bremia lactucae*. Mol. Plant. Microbe. Interact. 22: 1160–8.

Zhang, Y., Xu, S., Ding, P., Wang, D., Cheng, Y.T., He, J., Gao, M., Xu, F., Li, Y., Zhu, Z., Li, X., and Zhang, Y. (2010) Control of salicylic acid synthesis and systemic acquired resistance by two members of a plant-specific family of transcription factors. Proc. Natl. Acad. Sci. 107: 18220–18225.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. 10: 1–10.

Zipfel, C., Robatzek, S., Navarro, L., Oakeley, E.J., Jones, J.D.G., Felix, G., and Boller, T. (2004) Bacterial disease resistance in *Arabidopsis* through flagellin perception. Nature 428: 764–767.

Zohary, D. (1991) The wild genetic resources of cultivated lettuce (*Lactuca sativa* L.). Euphytica 53: 31–35.

Zuo, W. et al. (2015) A maize wall-associated kinase confers quantitative resistance to head smut. Nat. Genet. 47: 151–157.

8. Supplementary materials

Data is available via: 10.4121/21803436

Supplemental Figures

Supplemental Figure 1. Sequencing and assembly workflow to construct the *L. saligna* reference genome.
Supplemental Figure 2. Genome size estimation of *L. saligna* by GenomeScope.
Supplemental Figure 3. ALLMAPS re-scaffolding for *L. saligna* pseudo-chromosomes using genetic and syntenic map.
Supplemental Figure 4. Genomic features of the *Lactuca saligna* genome.
Supplemental Figure 5. Annotation Edit Distance (AED) cumulative fraction genome *de novo* annotation.
Supplemental Figure 6. Diversity of leaf shape of 10 resequenced *L. saligna* accessions.
Supplemental Figure 7. Syntenic path dot plot of *L. sativa* versus *L. saligna* highlighting two large inversions on chromosomes 5 and 8.
Supplemental Figure 8. Genetic distance versus physical position on chromosome 5 and 8 supporting the presence of large inversions.
Supplemental Figure 9. Circular tree of *L .saligna* nucleotide binding-leucine rich repeat receptors (NLR) generated by IQTREE.
Supplemental Figure 10. Top 100 enriched biological clusters of up-regulated DE genes at 8 and 24 hpi.
Supplemental Figure 11. Hard-filtering for biallelic SNPs variant from resequencing analysis of 15 *L. saligna* accessions.
Supplemental Figure 12. Cross-validation estimates the best K (population numbers) of ADMIXTURE for 15 re-sequenced *L. saligna* accessions.

Supplemental Tables

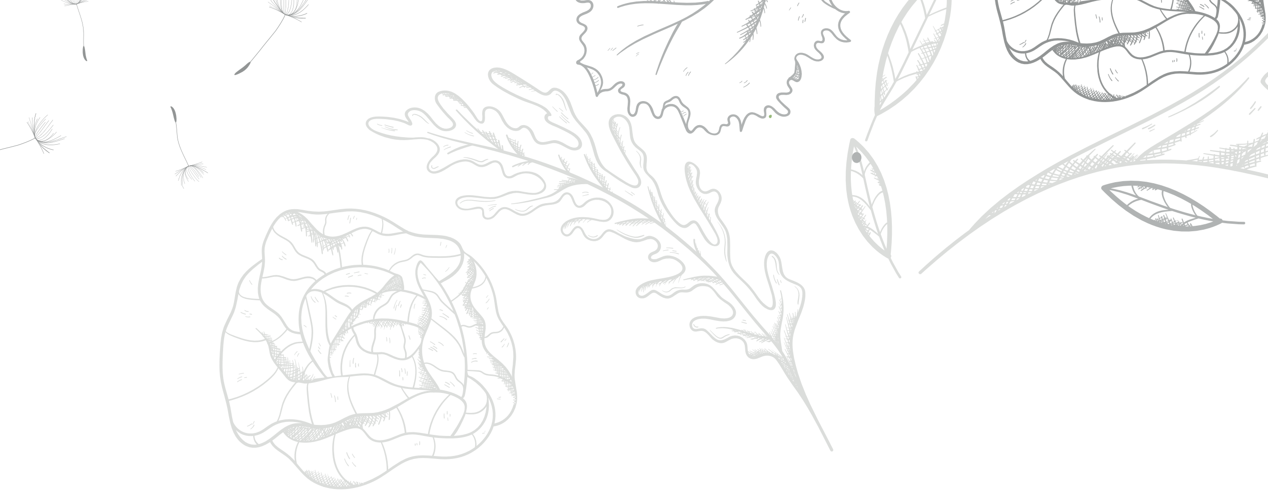
Supplemental Table 1. Resequencing data for genome size estimation.
Supplemental Table 2. ALLMAPS genome reconstruction by genetic and syntenic mapping.
Supplemental Table 3. ALLMAPS summary for the consensus map.
Supplemental Table 4. Chromosomal length of *L. saligna* (v4).
Supplemental Table 5. BUSCO assessment.
Supplemental Table 6. Mapping ESTs of different *Latuca* species to *L. saligna* genome.
Supplemental Table 7. Repeat annotation summary.
Supplemental Table 8. Categories of TEs predicted in the *L. saligna* genome.
Supplemental Table 9. Subcategories of TEs predicted in the *L. saligna* genome.
Supplemental Table 10. ncRNA prediction by different tools.
Supplemental Table 11. Functional gene annotation.
Supplemental Table 12. Passport of 15 *L. saligna* accessions.
Supplemental Table 13. Summary of *L. saligna* re-sequencing.
Supplemental Table 14. Summary of filtered SNPs for 15 TKI lines against *L. saligna* reference genome (v4).
Supplemental Table 15. Annotation of filtered SNPs by SnpEff.
Supplemental Table 16. Chromosomal position and gene count of inversions.
Supplemental Table 17. Genes at the border of putative breaking regions.
Supplemental Table 18. Genes flanking interspecific inversions between *L. saligna* and *L. sativa.*
Supplemental Table 19. ID conversion of genes flanking interspecific inversion.
Supplemental Table 20. *NLR* gene number in *L. saligna* and *L. sativa.*
Supplemental Table 21. Genomic position of Major Resistance Clusters (MRCs) in *Lactuca saligna.*
Supplemental Table 22. *NLR* clusters (NCs) in the *Lactuca saligna* v4 assembly.
Supplemental Table 23. Number of *NLR*s per RGC family in *L. sativa* and *L. saligna.*
Supplemental Table 24. Pfam HMM motifs used for *RLK* classification.
Supplemental Table 25. Hybrid incompatibility (HI) regions mapping in *L. saligna v4 assembly.*
Supplemental Table 26. NHR intervals and *R* gene locus mapping in *L. saligna* v4 assembly.
Supplemental Table 27. Number of differentially expressed genes in *L. saligna* upon *Bremia* infection.
Supplemental Table 28. DEGs in oomycete related ontology of Bremia-infected *L. saligna*.

Supplemental Table 29. Expression level of DEGs in oomycete related ontology of Bremia-infected *L. saligna.*
Supplemental Table 30. Statistics of DEGs in four NHR regions of *L. saligna.*
Supplemental Table 31. Candidate genes in NHR8 with potential *L. saligna* nonhost resistance.

Supplemental Data

Supplemental Dataset 1
Supplemental Datasets 2A-F
Supplemental Datasets 3A-B
Supplemental Datasets 4A-B
Supplemental Datasets 5A-D
Supplemental Datasets 6A-B
Supplemental Dataset 7
Supplemental note

Chapter 3



Genome assembly and analysis of *Lactuca virosa*: implications for lettuce breeding

Wei Xiong[1, +], Dirk-Jan M. van Workum[2, +], Lidija Berke[1, $], Linda V. Bakker[3, ‡], Elio Schijlen[3], Frank F.M. Becker[1, 4], Henri van de Geest[3, #], Sander Peters[4], Richard Michelmore[5], Rob van Treuren[6], Marieke Jeuken[7], Sandra Smit[2], M. Eric Schranz[1]

[1] Biosystematics Group, Wageningen University, P.O. Box 16, 6700 AP Wageningen, The Netherlands.
[2] Bioinformatics Group, Wageningen University, P.O. Box 633, 6700 AP Wageningen, The Netherlands.
[3] Bioscience, Wageningen University and Research, P.O. Box 16, 6700 AA Wageningen, The Netherlands.
[4] Laboratory of Genetics, Wageningen University, P.O. Box 16, 6700 AA Wageningen, The Netherlands.
[5] The Genome Center, Genome & Biomedical Sciences Facility, 451 East Health Sciences Drive, University of California, Davis, CA 95616-8816, United States.
[6] Centre for Genetic Resources, the Netherlands (CGN), P.O. Box 16, 6700 AA Wageningen, The Netherlands.
[7] Plant Breeding, Wageningen University, P.O. Box 386, 6700 AJ Wageningen, The Netherlands.
[$] Present address: Genetwister Technologies B.V., Wageningen, The Netherlands.
[‡] Present address: Hubrecht Institute, Developmental Biology and Stem Cell Research, Utrecht, The Netherlands.
[#] Present address: Hudson River Biotechnology, Wageningen, The Netherlands.
[+] Contributed equally.

This chapter has been submitted and is currently under review.

Abstract

Lettuce (*Lactuca sativa* L.) is a leafy vegetable crop with ongoing breeding efforts related to quality, resilience, and innovative production systems. Genetic variation of important traits in close relatives is necessary to meet lettuce breeding goals. *Lactuca virosa* (2x=2n=18), a wild relative assigned to the tertiary lettuce gene pool, has a much larger genome (3.7 Gbp) than *Lactuca sativa* (2.5 Gbp). It has been used in interspecific crosses and is a donor to modern crisphead lettuce cultivars. Here, we present a *de novo* reference assembly of *L. virosa* with high continuity and complete gene space. This assembly facilitated comparisons to the genome of *L. sativa* and to that of the wild species *L. saligna*, a representative of the secondary lettuce gene pool. To assess the diversity in gene content, we classified the genes of the three *Lactuca* species as core, accessory and unique. In addition, we identified three interspecific chromosomal inversions compared to *L. sativa*, which each may cause recombination suppression and thus hamper future introgression breeding. Using three-way comparisons in both reference-based and reference-free manners, we show that the proliferation of long-terminal repeat elements has driven the genome expansion of *L. virosa*. Further, we performed a genome-wide comparison of immune genes, nucleotide-binding leucine-rich repeat, and receptor-like kinases among *Lactuca* spp. and indicate the evolutionary patterns and mechanisms behind their expansions. These genome analyses greatly facilitate the understanding of genetic variation in *L. virosa*, which is beneficial for the breeding of improved lettuce varieties.

Keywords

Lettuce, genome assembly, comparative genomics, transposable elements (TEs), immune genes

1. Introduction

Lettuce (*Lactuca sativa* L.) is an agronomic crop with an economic value of ~3 billion USD per year (Food and Agriculture Organization of the United Nations, 2019). To breed better lettuce cultivars, breeders often search for novel genetic variation in wild relatives of lettuce. *Lactuca virosa* (2x = 2n = 18, biennial) is an important species in lettuce gene pool, for instance it is a donor for resistance to different viruses, such as beet western yellows virus, potyviruses lettuce Italian necrotic virus and Lettuce mosaic virus (Maisonneuve *et al.*, 2022; Maisonneuve *et al.*, 2018). The exploitation of *L. virosa* for lettuce breeding has had both challenges and successes. For example, despite reproductive barriers for direct intercrossing of the two species, breeders and scientists were able to execute interspecific hybridization bridged by *Lactuca serriola* or by use of *in vitro* embryo rescue and protoplast fusion to introduce traits such as robust root architecture and resistance to the currant-lettuce aphid and viruses (Thompson and Ryder, 1961; Eenink *et al.*, 1982; Maisonneuve *et al.*, 1995). Such interspecific crosses are part of the breeding pedigrees of well-known cultivars such as Vanguard and Salinas, representing modern crisphead lettuce cultivars (Mikel, 2007; Mikel, 2013). Novel introgressions of important genes and traits from *L. virosa* into lettuce could be accelerated through improved understanding of the lettuce genome and that of its wild relatives.

With the development of molecular markers and sequencing techniques, breeders can select traits with greater precision. For example, marker-assisted selection (MAS) has been used in lettuce breeding to accelerate selection by identifying gene/alleles in offspring (Simko, 2013). Genome-wide association studies (GWAS) have been performed to identify single-nucleotide polymorphisms (SNPs; Walley *et al.*, 2017) associated with various traits in lettuce for breeding (Sthapit Kandel *et al.*, 2020; Simko *et al.*, 2022) using the assembled lettuce (*L. sativa*) reference genome (Reyes-Chin-Wo *et al.*, 2017). A reference genome is essential for GWAS to reveal the loci of interesting traits of *L. virosa*. In addition, an assembly of *L. virosa* will enable the study of inter- and intra-species variation at the gene and genome level. For example, a whole-genome screening could be conducted to search for interesting genes such as immune genes that can trigger plant resistance response. An assembly of *L. virosa* could also be used to detect genome rearrangements between *L. virosa* and other *Lactuca* species via comparative genomics. However, the generation of a high-quality genome assembly for *L. virosa* is challenging because it has a large and highly repetitive genome. *Lactuca virosa* is a *diploid* species with 2n=2x=18 chromosomes, similar to *L. sativa* and *L. saligna* (Maisonneuve, 2003); however, *L. virosa* (3.7 Gbp) has a considerably larger genome size compared to *L. sativa* (2.5 Gbp), *L. serriola* (2.6 Gbp), and *L. saligna* (2.3 Gbp) (Doležalová *et al.*, 2002), probably due to variation in transposable elements (TEs). TEs are usually responsible for the variable and large genome sizes of plants (Wendel *et al.*, 2016). To date, there is only a single available genome assembly of *L. virosa* (CGN04683) (Wei *et al.*, 2021), which is a short-read based and highly fragmented assembly (3,694,810 scaffolds; N50=4,910bp) with a relatively high completeness (BUSCO=92.7%). The use of short-read technology combined with the high repeat content in *L. virosa* probably caused assembly inaccuracies and low continuity. Long-read sequencing could highly improve the accuracy and continuity of an *L. virosa* genome assembly.

Here, we present a near chromosome-level *de novo* assembly of *L. virosa* (CGN04683) using a combination of long-read and short-read sequencing plus Bionano and Dovetail scaffolding. We contextualize the *L. virosa* genome within the lettuce gene pool compared with the *L. sativa* and *L. saligna* (Xiong *et al.*, 2022) genomes. First, we show shared and specific homolog groups across the three species. Based on homologs, we show interspecific collinearity with an emphasis on structural variants (SVs) in different chromosomes. Next, we demonstrate that the proliferation of long-terminal repeat (LTR) superfamilies (Gypsy and Copia) explains the genome expansion of *L. virosa*. Finally, we describe a well-classified inventory of the two important resistance-related gene types encoding nucleotide-binding leucine-rich repeat (NLR) receptors and receptor-like kinase (RLK).

2. Materials and Methods

2.1 DNA and RNA sequencing

L. virosa accession CGN04683, also known as IVT280 and resistant to *Nasonovia ribisnigri* (currant-lettuce aphid). Single seed descent of IVT280 (obtained from a breeding company) was grown for whole-genome sequencing. The seeds were stratified at 4°C for three days to improve germination. Subsequently, seedlings were grown in a growth chamber at 18–21°C and a relative humidity of 75–78%. After eight weeks, plants were transplanted to larger pots containing potting soil and grown under greenhouse conditions. Tissue sampling was performed when plants were close to bolting, and DNA was extracted using the same protocol mentioned in Xiong *et al.* (2022). DNA material was used to prepare appropriate libraries and to produce a 20-fold coverage of long-read data generated by PacBio Sequel technology and a 69-fold coverage of paired-end (PE) reads obtained by Illumina sequencing. An optical mapping library of 130X coverage was produced by Bionano sequencing for hybrid scaffolding. A Hi-C library produced by Dovetail Genomics provided 10,492X physical coverage of the genome (10 kbp – 10 Mbp pairs) for in vitro proximity ligation (Supplementary Table 1). As additional evidence for gene prediction, RNA was isolated from pooled samples of leaf, root, and flower tissues (pooled from different floral stages) using a Direct Zol RNA Miniprep Plus kit (Zymo Research) followed by treatment with DNAse. RNA was purified by ethanol precipitation. The concentration and purity of RNA samples were measured with a Nanodrop 2000c spectrophotometer and a Qubit 4.0 fluorometer using an RNA Broad Range assay (Thermo Fisher Scientific). PE sequencing (2 x 125 bp) was performed on an Illumina HiSeq2500 platform (Supplementary Table 1).

2.2 Genome size estimation

After trimming, PE Illumina reads of *L. virosa* were used for genome size estimation (~1,590 million reads; ~183 Gb). Jellyfish v2.3.0 was used with a k-mer size of 21 to count k-mer frequencies (maximum 1 million count) (Marçais and Kingsford, 2011). The Jellyfish output was used by GenomeScope (v2.0) to estimate haploid genome length, percentage of repetitive DNA, and heterozygosity of the *L. virosa* genome (Ranallo-Benavidez *et al.*, 2020).

2.3 Assessment of genome completeness

Genome and proteome (annotation) completeness were assessed using BUSCO v5.2.0 with the 'eudicots_odb10' dataset (Manni *et al.*, 2021). K-mer completeness was assessed with KAT v2.4.1 with a k-mer value of 31 (Mapleson *et al.*, 2016).

2.4 Genome assembly

PacBio reads were assembled using Canu and then polished by Pilon (v1.20) using Illumina data (Koren *et al.*, 2017; Walker *et al.*, 2014). This assembly was corrected and improved using Bionano optical mapping data. Mis-joins in assembled contigs were corrected using the HiRise pipeline (Putnam *et al.*, 2016). Since the resulting assembly of Hi-C scaffolding was only 75.2% BUSCO complete, the publicly available—but highly fragmented—assembly for *L. virosa* (Wei *et al.*, 2021) was used to augment the completeness of our assembly. PE Illumina reads were trimmed before use with Trimmomatic v0.39 ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 (Bolger *et al.*, 2014), and the barcodes of the 10X mate pairs were stripped with longranger v2.2.2 basic. Before combining the assemblies, we first polished our assembly for a second time with the PE Illumina reads and the 10X mate pair reads (treated as single-end reads) using Pilon v1.24--changes--diploid--fix all (Walker *et al.*, 2014). Mapping of sequencing reads for combining these two assemblies was performed with bwa-mem2 v2.2.1 (Vasimuddin *et al.*, 2019). Next, we combined our assembly with all sequences >1 kb in the Wei *et al.* [17] assembly. We then aligned all PE Illumina and 10X data to the combined genome. The coverage of this data was used to get the best haplotype representation of the complete genome with purge_haplotigs v1.1.1 (cut-offs were 10, 85, 180) (Roach *et al.*, 2018). Since the number of sequences in the resulting assembly increased from 29 to 54,814, we applied several filtering steps to reduce the number of small, uninformative sequences. We filtered out possible mitochondrial and plastid sequences by blasting all sequences to the mitochondrial and plastid NCBI databases. We filtered out non-Viridiplantae sequences as identified by a blastn search against the NCBI database. Then, we polished the newly added sequences using the same method we used to polish our original genome assembly before (with Pilon v1.24). Based on coverage of PE Illumina and 10X data, we used purge_haplotigs to check whether any duplications were introduced, but since this was not the case, we did not apply purge_haplotigs a second time. For scaffolding the newly added sequences, we mapped the original PacBio data to the genome with minimap2 v2.21-cxmap-pb (Li, 2018). Scaffolding was done with LRScaf v1.12 -misl 3 -t mm (Qin *et al.*, 2019). To keep only potential gene coding sequences, we mapped the RNA-seq data with STAR v2.7.7a (Dobin *et al.*, 2013) and removed all sequences lacking a single alignment. Finally, we also removed all sequences smaller than 5 kb.

2.5 Repeat annotation

To annotate the repetitive elements in the *L. virosa* genome, a custom library was created by combining different sources: a *de novo* library of TEs created by RepeatModeler (v2.0.1) with -LTRStruct parameter, a *de novo* library of miniature inverted-repeat transposable elements (MITEs) searched by MITE-hunter, and a specific database for the genus *Lactuca* extracted from a combined database of Dfam (20170127) and Repbase (20170127) (Flynn *et al.*, 2020; Han and Wessler, 2010; Hubley *et al.*, 2015; Bao *et al.*, 2015). Then, RepeatMasker (v4.0.7) was used to soft mask the *L. virosa* genome assembly (Smit *et al.*, 2019). The same pipeline was also applied to create a TE library and mask the genome assembly of *L. saligna* version 4 (PRJEB35809) and *L. sativa* version 7 (GCF_002870075.2), which were used in reference-based repeatome comparison. The three generated TE libraries were used for a reference-free approach for TE classification (see Individual and comparative clustering analysis of repetitive elements below). The RepeatMasker outputs were further processed to summarize the different categories of repeat elements. Moreover, the LTR elements were extracted from the cross_match output of RepeatMasker and compared to the genome using bedmap.

2.6 Gene prediction

Protein encoding genes in the nuclear assembly were annotated using MAKER2, which combines *de novo* gene prediction and homology prediction (Holt and Yandell, 2011). rRNA reads were filtered out by SortMeRNA version 4.3.4 (Kopylova *et al.*, 2012) using all databases to remove non-coding rRNA. Subsequently, HISAT2 (v2.2.1) was applied to map the remaining reads to the final genome assembly, which includes nuclear sequences, the mitochondrion assembly of CGN013357 (MZ159960.1) and plastid assembly of TKI-404/CGN04683 (CNP0000335 on CNGB) (Kim *et al.*, 2015; Wei *et al.*, 2021; Fertet *et al.*, 2021). The alignment to the nuclear sequences was used as input to BRAKER (version 2) and Stringtie (v2.1.6) to conduct *de novo* gene prediction and transcriptome assembly, respectively, both with default settings (Hoff *et al.*, 2016; Pertea *et al.*, 2015). The protein alignment was done by BLAST in MAKER2 during the integration with protein databases of *A. thaliana* (Araport11), *L. sativa*, *Helianthus annuus* (HA412. v1), and Uniprot (SwissProt set only: release-2019_10). The predicted transcripts were then filtered using the following criteria: eAED > 0.9 (computed by MAKER2), protein length < 50, identical isoforms, and missing start and stop codon.

2.7 Functional annotation

Potential biological function of proteins was inferred using three criteria: i) best-hit matches in SwissProt, TrEMBL using DIAMOND version 2.0.14 at E-value cut-off of 1e-5 (Buchfink *et al.*, 2015); ii) protein domains/structure identified by InterProscan 5.53-87.0 against the Pfam, Coils, Gene3D, PANTHER, SUPERFAMILY, ModiDBLite, and TIGRFAM databases (El-Gebali *et al.*, 2018; Zdobnov and Apweiler, 2001); and iii) orthology searches for pathway information were conducted by Kofamscan (Aramaki *et al.*, 2019) using a customized HMM database of KEGG orthologs (Kanehisa, 2000) with an E-value cut-off of 1e-5.

2.8 Gene space analysis

To enable a comparison between *L. virosa*, *L. saligna*, and *L. sativa*, we used PanTools v3.4.0 (Jonkheer *et al.*, 2022) to calculate homologous relationships in a panproteome of these three species. We used the longest isoform for each gene. Based on an optimal distribution of BUSCO genes, we decided to use-rn 2 for homology grouping. Subsequent gene classification of the homolog groups was also done with PanTools. The number of shared groups were visualized with ComplexUpset (Krassowski, 2020). Functional enrichment analyses were performed and visualized for the unique sets of genes with ClusterProfiler v3.18.1 (Yu *et al.*, 2012).

2.9 Synteny detection

MCScanX (Wang *et al.*, 2012) was utilized to detect syntenic blocks (default settings) among the three *Lactuca* species using the calculated homolog groups from PanTools. The interspecific collinearities were visualized using SynVisio (Bandi and Gutwin, 2020). MCScanX was run a second time to detect the tandem arrayed genes using DIAMOND (version 2.0.14) on proteomes for each species.

2.10 Individual and comparative clustering analysis of repetitive elements

RepeatExplorer2 on a Galaxy server was used (https://repeatexplorer-elixir.cerit-sc.cz/) to conduct individual and comparative clustering of Illumina PE reads for three *Lactuca* species (*L. sativa*, *L. saligna*, and *L. virosa*) (Novák *et al.*, 2020). Re-sequencing data of these three *Lactuca* species were retrieved from ENA database (PRJEB36060). Trimmed FASTQ reads were converted to FASTA format and interlaced prior to the clustering analysis. In addition, a four-letter prefix identity code was added to each sample dataset (i.e., Lsat for *L. sativa*, Lsal for *L. saligna*, and Lvir for *L. virosa*). After a preliminary round, each set of reads was randomly subsampled with a same proportion to maximize the repeat detection and annotation accuracy. For individual analysis, reads representing 20% of the genome size were separately clustered for each *Lactuca* species (i.e., genome proportion = 0.2x, *L. sativa* = 4,166,668 reads, *L. saligna* = 3,833,334 reads, and *L. virosa*= 6,166,668 reads). For comparative analysis, a mixed dataset of reads equal to 0.07x depth for all species was clustered at once (i.e., genome proportion = 0.07x, *L. sativa* = 1,307,006 reads, *L. saligna* = 1,420,966 reads, and *L. virosa*= 2,103,018 reads). For both analyses, the reads were clustered based on the default settings (90% similarity, 55% coverage), and clusters containing more than 0.01% reads were classified at a supercluster level.

After clustering, repeat reads were annotated based on a similarity search to REXdb (protein domain in retrotransposons) using BLAST on a Galaxy server (Neumann *et al.*, 2019). Additionally, the custom libraries previously created by reference-based searches were utilized to further annotate the repeat clusters (see previous section: Repeat annotation). After annotation, clusters from plastid and mitochondrial origins were identified and excluded for downstream analysis. Next, we quantified different TE categories based on clusters and their connections to superclusters. To characterize the interspecific difference, the clusters resulting from comparative analysis were sorted via hierarchical clustering (ward.D2) using transformed read number [$\log_2$(count + 1)] in each cluster for every species.

2.11 Analysis of immune gene repertoire

NLRs were searched for in the proteomes of *L. virosa* and *L. sativa,* and retrieved from the *L. saligna* genome (Xiong *et al.*, 2022). HMMER was used to search Hidden Markov Models (HMMs) profiles obtained from Pfam or the UC Davis database (https://niblrrs.ucdavis.edu/At_Rgenes/HMM_Model) for structural domains of NLR proteins (E-value cut-off = 1e-10): PF00931.23 and NBS_712.hmm for the nucleotide-binding (NB) domain; PF01582.20 and PF13676.6 for TIR (TOLL/interleukin-1 receptor); PF05659.11 and PF18052.1 for CC (coiled-coil); and eight HMMs for the LRR (leucine-rich repeat) domain (PF00560.33, PF07723.13, PF07725.13, PF12799.7, PF13306.6, PF13516.6, PF13855.6, PF14580.6). Furthermore, NB and LRR domains identified by InterProScan (see Functional annotation), and CC motifs predicted by Paircoil2 (P scores < 0.025) were combined with the HMMER output (McDonnell *et al.*, 2006; Zdobnov and Apweiler, 2001). The identified NLRs were classified as TNL or CNL based on the presence of either the TIR or CC domain, respectively. To further solve the unclassified NLRs (TNL or CNL), a phylogenetic tree for amino-acid (aa) sequences with NB domains was constructed. First, aa sequences were aligned using HmmerAlign (Finn *et al.*, 2011). The alignment was then trimmed by trimAl using -automated1 mode and retained 727 residues for phylogenetic construction (Capella-Gutiérrez *et al.*, 2009). A maximum-likelihood (ML) tree was inferred by IQTREE version 1.6.12 (-m PMB+F+R10) with 1,000 ultrafasta bootstrap (UFBoot) replicates (Nguyen *et al.*, 2015). The phylogenetic tree was visualized and annotated using iTOL v6 (Letunic and Bork, 2021). An Inventory of RLKs was also performed for *L. virosa* and *L. sativa*. First HMMER (v3.3.2) was used to search the Pkinase domain (PF00069; E-value cut-off = 1e-10). Then, proteins containing Pkinase were examined for the existence of extracellular domains using HMMER (E-value cut-off = 1e-3) and transmembrane regions using TMHMM (v2.0) and SCAMPI (v2) (Krogh *et al.*, 2001; Peters *et al.*, 2016).

3. Results and Discussion

3.1 Genome assembly and annotation

We created a complete and structurally informative genome assembly for *L. virosa* with a total length of 3.45 Gbp (Table 1). Based on a k-mer analysis of Illumina data, we estimated the genome size to be 3.3 Gbp with 73% repeat content (Supplementary Figure 1). This predicted size was lower than the previously measured C-value (3.7 Gbp) (Doležalová *et al.*, 2002), which might be caused by the large repeat content of *L. virosa* (Ranallo-Benavidez *et al.*, 2020). The genome assembly resulted from a non-redundant combination of a novel long-read assembly and an existing short-read assembly from Wei *et al.* (Wei *et al.*, 2021) (Supplementary Table 3). The long-read assembly was based on PacBio and Illumina data, and scaffolded using Bionano and Hi-C data. The longest 12 scaffolds out of the 29 scaffolds comprised 99.8% of the total length (3.3 Gbp) of this first assembly, yet not all chromosomes were reconstructed in full. The BUSCO completeness score (75.2%) indicated some coding regions of the genome were missing from this *L. virosa* genome assembly, which was also confirmed by the k-mer analysis with Illumina data (Supplementary Figure 2A). Thus, we completed the assembly through additional polishing and leveraging the fragmented, short-read based genome assembly of the same *L. virosa* accession (Wei *et al.*, 2021) (Supplementary Data 1), which did have a high BUSCO and k-mer completeness (Supplementary Figure 2B and 2D). The final combined assembly consisted of 5,855 contigs spanning a total of 3.45 Gbp with an N90 (L90) score of 116,478,781 (10) (Supplementary Figure 2C; Supplementary Table 2). The BUSCO completeness score was 96.2% (the duplication score was 4.5%; Supplementary Table 3; Supplementary Figure 2D).

Table 1. Summary of assemblies for *Lactuca* spp. in this paper.

Characteristic	*L. sativa*	*L. saligna*	*L. virosa*
Accession ID	GCF_002870075.2	PRJEB35809	PRJEB50301
source	RefSeq (NCBI)	ENA	This study
Assembly size (Gb)	2.39	2.17	3.45
# seq	8,325	10	5,855
N50	257.9 Mb	238.6 Mb	316.9 Mb
L50	4	4	5
Assembly complete BUSCO	97.8% (2,273)	92.4% (2,147)	96.2% (2,236)
# protein-coding genes	36,136	42,908	39,887
# transcripts	46,867	45,476	42,791
Proteome complete BUSCO	98.5% (2,291)	88.8% (2,065)	90.2% (2,096)

Based on both expression and orthology evidence, 39,887 protein-coding genes with a total of 42,791 transcripts were annotated. We mapped RNA-seq data from root, leaf, and flower tissue to the genome assembly to support *de novo* gene prediction. Next, we aligned protein sequences of model plant species to the genome and used MAKER for merging all gene predictions. We filtered the predicted genes to only retain annotations that were in accordance with the provided evidence. The final BUSCO score of the proteome was 90.2%, indicating a high level of completeness. Furthermore, we were able to predict functional domains in 93% (37,106) of the genes for various databases (Supplementary Table 4; Supplementary Data 2A - B). This structural and functional annotation is vital for the biological interpretation of *L. virosa* data.

3.2 Homolog grouping of three representative *Lactuca* spp.

Even though the genome size of *L. virosa* is substantially larger than *L. sativa* and *L. saligna*, the number of genes annotated across species was similar (Table 1). A comparison of *L. virosa* with *L. saligna* and *L. sativa* showed that about half of the homologous groups are shared across *Lactuca* (Figure 1; Supplementary Data 2C). These 17,741 homolog groups in *Lactuca* contained 19,270 *L. virosa* genes, meaning that about half of the *L. virosa* genes are part of the core *Lactuca* genome.

This is comparable to what was found in other interspecies comparisons. For example, in rice ~62% of core genes were reported between two species (Zhao *et al.*, 2018), and in *Raphanus*, ~50% of core genes were reported among 11 accessions belonging to two species (Zhang *et al.*, 2021). Both *L. virosa* and *L. saligna* share fewer homologous genes with each other than with *L. sativa*. This stresses the importance of wild species in breeding as they contain a large pool of novel genes. The large, unique genomes of both *L. virosa* and *L. saligna* indicate that these wild species are rich sources of genetic diversity that thus far has been unexploited for lettuce breeding. We performed a functional annotation for the proteomes of the three species with InterProScan to perform functional enrichment for the unique content of *L. virosa* (15,048 genes; Supplementary Data 2D). The InterProScan domain enrichment found disease resistance proteins to be among the set of significantly enriched domains (Supplementary Figure 3). Therefore, the genome of *L. virosa* is a resource for potential novel genes needed for resilience breeding in lettuce.

Furthermore, it will be relevant to sequence and produce high-quality assemblies of other wild relatives of lettuce, such as *L. georgica*, *L. serriola*, and *L. aculeata,* to obtain an overview of the entire *Lactuca* gene space (Guo *et al.*, 2022; Wei *et al.*, 2021). Using high-quality genetic resources will enable the construction of a comprehensive pangenome that covers the variation in the genus *Lactuca*.

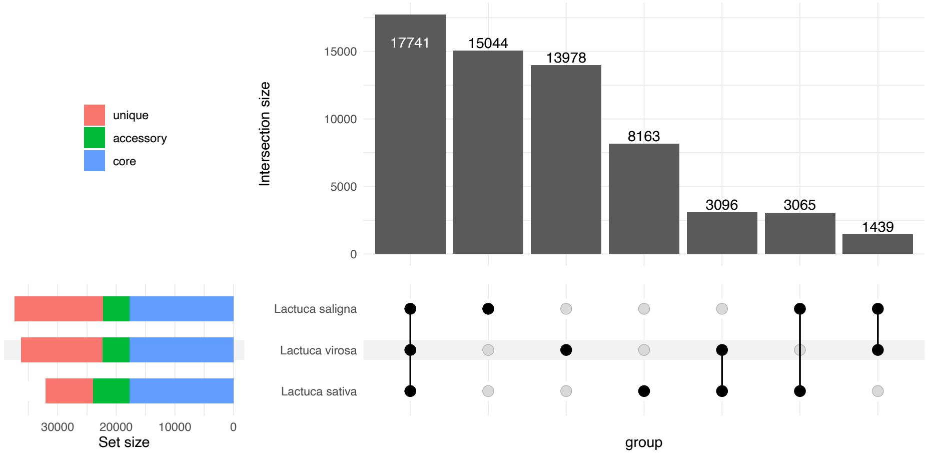


Figure 1. Overview of homolog grouping for *L. sativa*, *L. saligna*, and *L. virosa* in an upset plot. The numbers are groups of homologous genes. In total, there are 62,526 homologous groups.

3.3 Synteny detection between three *Lactuca* spp. via comparative genomics

By synteny detection of homologous pairs, we identified major chromosomal inversions between the three *Lactuca* genomes. Overall, there was whole-genome collinearity (synteny) among *Lactuca* species (Figure 2D). Based on the collinearity, we determined the major 12 scaffolds that comprised 96% (3.30 Gbp) of the total genome assembly (Supplementary Table 5).

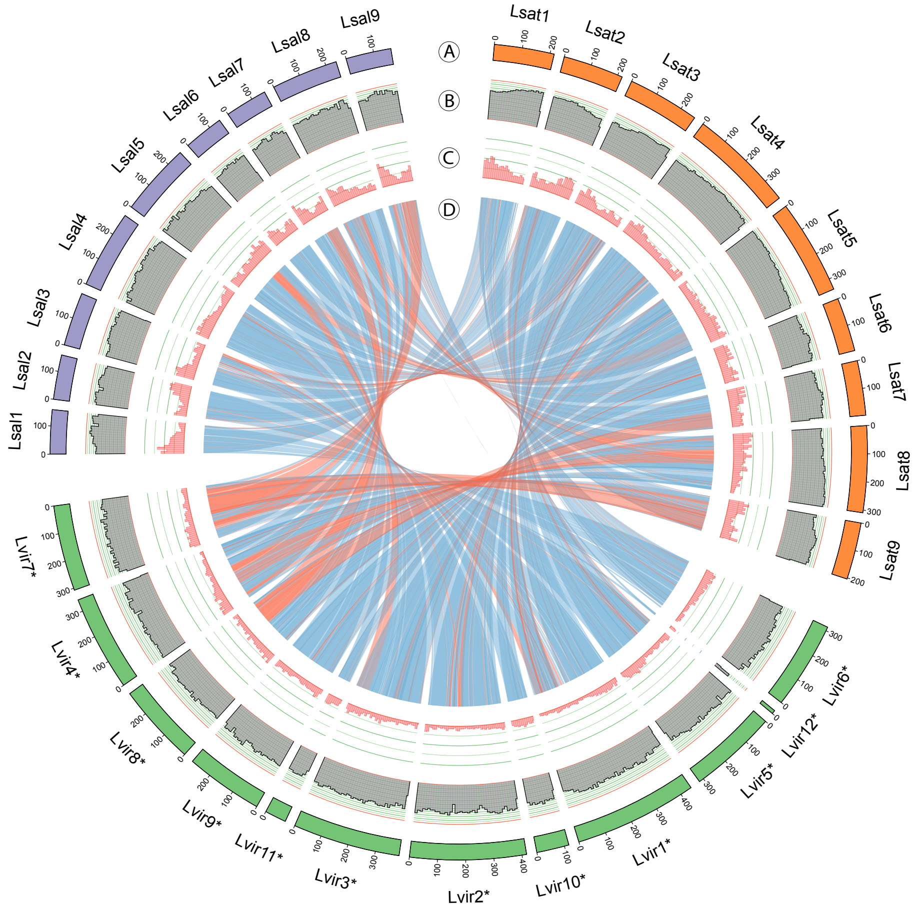


Figure 2. Circos plot of *L. virosa* genome compared with the *L. sativa* and *L. saligna* genomes. **A**, For each of the three genomes (Lsat: *L. sativa*; Lvir: *L. virosa*; Lsal: *L. saligna*), only sequences larger than 1 Mbp are shown. These are only the chromosomes for *L. sativa* and *L. saligna*. Since *L. virosa* is near-chromosome level, we indicated the sequence numbers with an asterisk (*) to make clear that these are not chromosome numbers. Some of the sequences for *L. virosa* are inverted to match the synteny best (**D**); the sequence coordinates (in Mbp) show this. **B**, The repeat density for each sequence is calculated per 10 Mbp and shown here as a fraction. Since the genome assembly for *L. virosa* has more N bases, repeats are more difficult to find than in the other two genomes. The scale goes from 0 to 1. **C**, The gene density for each sequence is calculated per 10 Mbp and shown here as a fraction. As the three genomes contain approximately the same number of genes but their genome sizes differ; *L. virosa* has a lower overall gene density. The scale goes from 0 to 0.2. **D**, Synteny between the three genomes. Inversions are shown in red as opposed to non-inverted syntenic blocks, which are shown in blue.

Compared to the *L. sativa* genome, three lineage-specific inversions were identified on different chromosomes (Chr; Figure 3). Two of the three inversions that were previously described between *L. saligna* and *L. sativa* were validated and further characterized: one is specific to *L. saligna* on Chr5 and one is specific to *L. sativa* on Chr8 (Xiong *et al.*, 2022).

Furthermore, synteny also revealed a large inversion specific to *L. virosa* on tentative Chr7 (Scaffold8) (Figure 3; Supplementary Table 5). These inversions could hamper genetic mapping of interesting traits and further introgression. The syntenic patterns between *L. virosa* Chr9 (scaffold 7) and the other two species showed complicated inverted and translocated regions, which might be due to a reversed-joining from Hi-C scaffolding (Supplementary Figure 4).

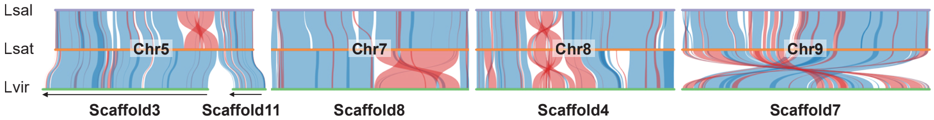


Figure 3. Synteny discloses species-specific inversions across three *Lactuca* species. Through genomic comparison, major interspecific inversions (red) were identified among the reference genomes of three *Lactuca* species. Here, the synteny in four sets of scaffolds/chromosomes reveal species-specific inversions: *L. saligna* (Lsal: purple), *L. sativa* (Lsat: orange), and *L. virosa* (Lvir: green). The chromosome numbers are labelled in the middle. Black arrows at the bottom indicate reversed scaffolds in the *L. virosa* assembly. Supported by Supplementary Table 5.

3.4 Comparative repeatomics between three *Lactuca* spp. via reference-based and reference-free approaches

In the three reference assemblies, we annotated repeat elements and classified them into TEs and other repeats (Supplementary Data 3A). The genomes of all three *Lactuca* species contained a major proportion of TEs, in agreement with previous studies (Supplementary Table 6) [13]. Intriguingly, the TE content of the *L. virosa* (60%) assembly is lower than that of both *L. sativa* (74%) and *L. saligna* (77%), whereas a higher TE content was expected in *L. virosa*. After excluding the N content of each genome, the percentage of TEs for all *Lactuca* genomes exceeded 80% (Supplementary Figure 5; Supplementary Table 6). Moreover, LTR TEs represented more than 50% of the three *Lactuca* genomes (excluding N content). We further characterized LTRs in the *L. virosa* genome by determining their genomic position. Almost all identified LTRs (99%) were located in the intergenic regions and their density gradually decreased nearing a genic region (Supplementary Figure 6). Moreover, the non-repeat regions (i.e., unmasked length, excluding N content) were similar in all species, regardless of the change in repeat length (Supplementary Table 6). To conclude, this reference-based repeat annotation showed that TEs are the most abundant components of *Lactuca* spp. genomes. However, genome incompleteness and N content of the reference genome assemblies hamper a precise estimation of TEs.

In addition to the reference-based repeat annotation, we also classified repeat components and estimated their composition for the three *Lactuca* spp. in a reference-free way by annotating the clusters of repetitive re-sequencing short reads (Supplementary

Table 7). First, we performed an individual analysis for three *Lactuca* species with a read depth of 0.2x. In total, we found that *L. virosa* had the highest percentage of repeated reads assembled as clusters (82%). For top clusters (cluster size > 0.01% analyzed reads), the percentage of repeated reads was in line with the estimated genome size for three *Lactuca* species: *L. virosa* (74%), *L. sativa* (68%), and *L. saligna* (65%). After curation, we calculated the genomic proportion of different types of TEs based on the annotated read clusters (Supplementary Table 8). In all species, more than 60% of repeated sequences were annotated as TEs, which comprised almost 100% of Class I LTRs. Among them, *L. virosa* had the highest amount of LTRs (68.34%) followed by *L. sativa* (61.24%) and *L. saligna* (58.64%). It is likely that the overall genome size expansion of *L. virosa* was driven by TEs.

To explore this further, we identified the major differences in TEs represented by read clusters between three *Lactuca* species via another comparative analysis (RepeatExplorer). This approach used mixed reads at the same depth (0.07x) for every species (Supplementary Table 7). Compared to the individual mode, the clusters resulting from the comparative analysis contained the read number from each species for each cluster, i.e., a cluster matrix for three *Lactuca* species (Supplementary Data 3B). For example, cluster 10 was annotated as LTR/Gypsy and mainly composed of *L. virosa* reads (Supplementary Figure 7). After curation, the total repeat content of the top clusters was 69.45% for mixed data, which was higher than *L. sativa* (63.54%) and *L. saligna* (61.75%) and lower than *L. virosa* from individual analysis (70.40%), indicating that *L. virosa* carries a higher percentage of repeats than the other two *Lactuca* species (Supplementary Table 8).

Based on the comparative analysis, an in-depth cluster analysis revealed that the LTR proliferation in *L. virosa* drove its genome expansion (Supplementary Data 3C). The heatmap of hierarchical clustering shows six groups that were either dominated (D) by one of the three *Lactuca* species: *L. sativa* (Lsat), *L. saligna* (Lsal), or *L. virosa* (Lvir) (Figure 4A: left). The bar plot in Figure 4A (right) further decomposes the read sources for each group. The Lvir_D2 group, dominated by *L. virosa* reads, was also the largest group in the hierarchical clustering analysis. Further investigation of the genomic proportion for clusters in the Lvir_D2 group was carried out. Collectively, the clusters in Lvir_D2 comprised 50.05% of the reads and were significantly larger than the other five groups. This Lvir_D2 group is dominated by *L. virosa*, contributing approximately two times as many reads as *L. sativa* and *L. saligna*. Furthermore, nearly all read clusters in Lvir_D2 were annotated as LTR (48.47% of analyzed reads) and mainly represented by two LTR sub-families: Gypsy (27.31%) and Copia (20.46%) (Supplementary Table 9). Additionally, the subgroups Tekay and Angela were the primary elements for the Gypsy and Copia clusters within the Lvir_D2 group (Figure 4B; Supplementary Table 9).

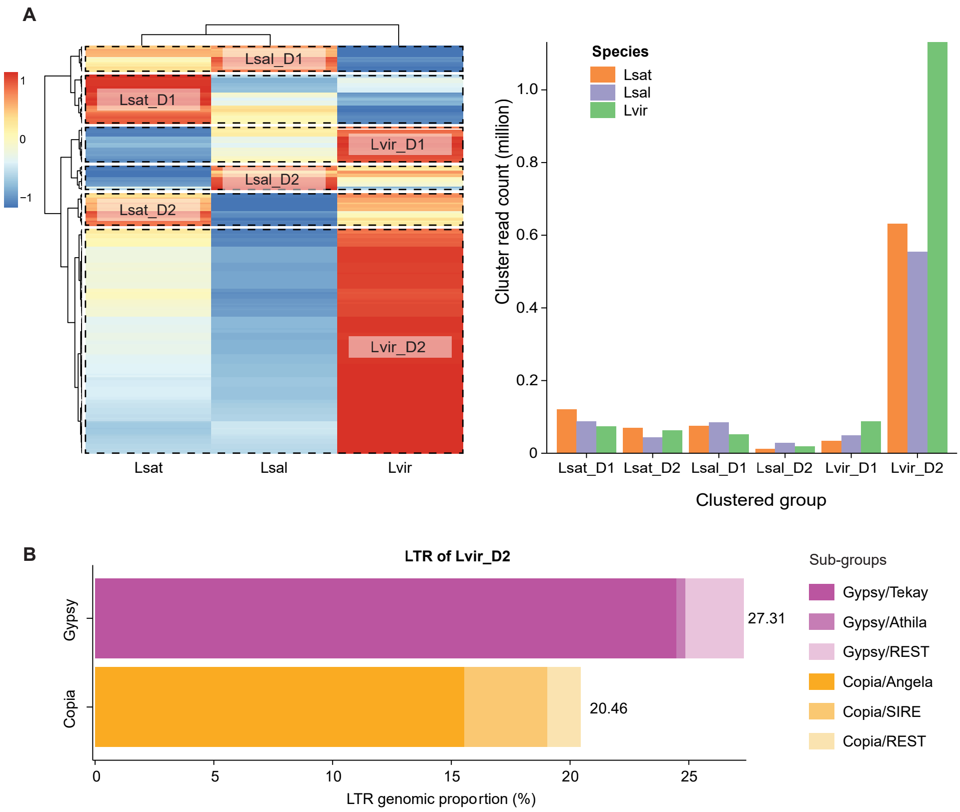


Figure 4. Proliferation of long-terminal repeats (LTR) drives the expansion of the *L. virosa* genome. Read-clusters assembled by RepeatExplorer2 using a mix of re-seq data (coverage = 0.07x) from three *Lactuca* species references to detect the major difference of repeat elements. **A,** Heatmap shows the scaled read-count of individual cluster (row) for each species (column). Clusters and species were sorted by hierarchical clustering. Six groups (squared by dash lines) were dominated by reads either from *L. sativa* (Lsat), *L. saligna* (Lsal), or *L. virosa* (Lvir) and suffixed with a D (dominant). Bar plot shows the size (y-axis) of six clustered groups (x-axis) for each species: *L. sativa* (orange), *L. saligna* (purple), and *L. virosa* (green). **B,** Stacked bar chart shows the composition of sub-groups for the two major LTR superfamilies: Gypsy (gradient purple) and Copia (gradient yellow). Supported by Supplementary Table 9.

L. virosa is estimated to have a significantly larger genome (3.7 Gbp) than *L. sativa* (2.5 Gbp) and *L. saligna* (2.3 Gbp) (Doležalová *et al.*, 2002). TEs have been shown to drive plant genome expansion (Wendel *et al.*, 2016); for example, within the genus of rice (Ma and Bennetzen, 2004; Ammiraju *et al.*, 2007; Piegu *et al.*, 2006). Based on our combined findings, we conclude that the subgroups of transposon LTR, Tekay in Gypsy, and Angela in Copia drove the genome expansion of *L. virosa*.

3.5 Comparison of NLR and RLK genes between three *Lactuca* spp.

Besides the difference within TEs, there is also sizable variation in the number of genes as shown by the homology grouping (accessory/unique genes) among these three *Lactuca* species (Supplementary Figure 3), which might convey resilience to important traits like resistance against various pathogens or pests. In our previous study, an extensive search of resistance genes was performed for lettuce and its wild relative *L. saligna* (Xiong *et al.*, 2022). Using the new *L. virosa* assembly, we identified and classified immunity-related genes encoding NLR and RLK proteins for *L. virosa* and compared them to *L. sativa* and *L. saligna* (Table 2).

Table 2. Identification and classification of immunity related genes for *Lactuca* species.

Immune genes		**Species**		
Family	**Classification**	***L. sativa***	***L. saligna***	***L. virosa***
NLR	CNL*	158	139	148
	TNL	227	184	161
	Total	385	323	309
*RLK***	Rcc1-RK	5	5	2
	WAK	61	48	36
	G-LecRK	132	79	70
	L-LecRK	31	29	21
	C-LecRK	1	1	1
	CRK	41	35	38
	Malectin-RK	55	55	32
	LysM-RK	12	12	11
	LRR-RK	258	213	233
	PERK	1	1	1
	Total	597	478	445

* RPW8 and Rx_N type of CNL included in this study.

** RLK classification based on extracellular domain (Supplementary Table 12 and 13).

The *L. sativa* genome was found to have the highest number of NLRs (385), followed by *L. saligna* (323) and *L. virosa* (309) (Table 2; Supplementary Table 10). In association with the homology grouping, a Venn diagram showed that the NLRs identified in three *Lactuca* spp. are highly diverged, where more than 50% of NLRs in each species belong to specific homologous groups (Figure 5A: left; Supplementary Data 5A). This observation is in line with our enrichment study of homologs specific to *L. virosa*, where InterProScan domains were significantly enriched with terms related to NLR proteins (Supplementary Figure 3). Furthermore, NLR proteins were classified into TNL and CNL types based on the N-terminal domain (TIR or CC domain, respectively) and curated by the phylogeny of a nucleotide-binding (NB) domain alignment (Supplementary Figure 8; Supplementary Data 4A - B). The difference between *L. sativa*, *L. saligna*, and *L. virosa* was mainly contributed by *TNL* genes (227 vs 184 and 180), and the difference between *L. saligna* and *L. virosa* can be explained by the *CNL* type (139 vs 162). Due to the unequal completeness of the proteomes, we applied the ratio of complete BUSCOs for proteomes as a benchmark to anticipate whether *NLR genes* expand or contract between the three *Lactuca* spp.: *L. sativa* (2,291), *L. saligna* (2,065), and *L. virosa* (2,096). The ratio of BUSCOs (1.10 : 1.00 : 1.02) reflects the NLR ratio across species (1.25 : 1.05 : 1.00), where *L. sativa* showed a slight inflation. For different NLR types, the number of CNLs was similar in the examined species *sativa*, *L. saligna*, and *L. virosa* (1.14 : 1.00 : 1.06); however, the ratio of TNL numbers highly deviated from the BUSCO ratio (1.41 : 1.14 : 1.00; Supplementary Table 10). Such comparison suggests an expansion of *NLR*s in *L. sativa*, which is possibly caused by tandem duplication events as in most studied angiosperms (Wu *et al.*, 2021). This hypothesis is supported by whole-genome search of tandem duplicates (TDs) clusters between three *Lactuca* spp. genomes (Supplementary Data 5B). The number of TDs encoding NLRs in *L. sativa* (121) was approximately two-times larger than that in *L. saligna* (61) and *L. virosa* (76), which principally explains the number difference among the three species (Figure 5B: left). In addition to tandem duplication, transposon activities (e.g. LTRs) could also greatly elevate the number of *NLR*s by retroduplication as reported in the chili genome (Kim *et al.*, 2017). The retroduplicated *NLR*s could partially explain the lineage-specific homologs among *Lactuca* species (Figure 5A: left).

As for RLK encoding genes, we identified RLK proteins by searching for the extracellular, transmembrane, and intracellular domains. Then, resulting RLKs were classified into nine types based on their extracellular and kinase domains (Supplementary Table 11-12; Supplementary Data 4C-D). Like NLRs, we found more genes encoding RLK proteins in the *L. sativa* (597) genome assembly than in *L. saligna* (478) or *L. virosa* (445; Table 2). Homology shows that *RLK*s were much more conserved in *Lactuca* spp. compared to NLRs, where 70% of *RLK*s in each *Lactuca* species were homologous to another RLK from at least one sister species (Figure 5A: right). Compared to the BUSCO completeness, the RLK ratio (1.25 : 1.00 : 1.00) showed an increase of *RLKs* in *L. sativa*, suggesting a possible expansion of the *RLK* family. The expansion in *L. sativa* was majorly contributed by *G-LecRK*, followed by *Malectin-RK* and *WAK*, while other types of *RLK*s were either similar in all species or slightly inflated in *L. sativa*. The extra *G-LecRK* and *WAK* copies might confer specific immunity in *L. sativa*. For example, G-LecRK and WAK can both mediate resistance to *Phytophthora spp.* (oomycete) in tobacco and melon plants (Pi *et al.*, 2022; Wang *et al.*, 2020). On contrary, the expansion of *Malectin-RK* might benefit pathogen invasion in *L. sativa*, like the increased susceptibility to *Hyaloperonospora arabidopsidis* (oomycete) observed in *Arabidopsis* (Hok *et al.*, 2011). Similar to *NLR*s, *RLK*s also commonly expand via tandem duplications. For example, a *G-LecRK* expansion was reported in soybean (Rodgers-Melnick *et al.*, 2012; Liu *et al.*, 2018). The number of tandem arrayed *RLK*s in *L. sativa* was 1.5 and 1.9 times that of the *RLK*s in *L. saligna* and *L. virosa*, respectively, which constitutes more than 60% of the difference between

L. sativa and other two species (Figure 5B: right; Supplementary Data 5B). Especially for *G-LecRK*, the number of tandem genes appeared more than doubled in *L. sativa* (Supplementary Data 5B).

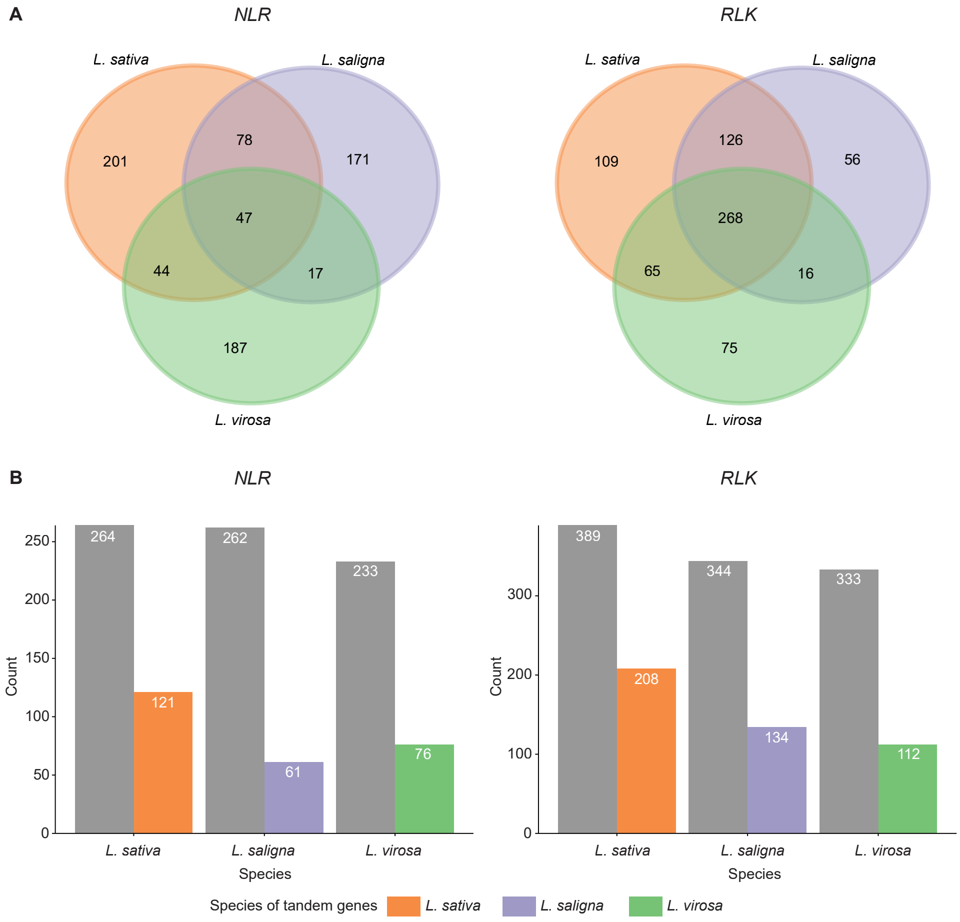


Figure 5. Associating immune genes with their homology and tandem duplication event. **A,** Venn diagrams of homologous groups for nucleotide-binding leucine-rich repeats (*NLRs*) and receptor-like kinases (*RLKs*) in *Lactuca* spp. Homologous grouping was done by PanTools. **B,** Bar plots show the count of tandem and non-tandem *NLRs* and *RLKs* in three *Lactuca* species. Tandem arrayed genes were identified by MCScanX. Supported by Supplementary Data 5.

4. Conclusions

Here, we publish a near chromosome-level genome assembly for *L. virosa* (accession CGN04683) that has a high completeness. As a representative of the tertiary lettuce gene pool, this *L. virosa* genome assembly enables comparisons with *L. sativa* of the primary gene pool and *L. saligna* of the secondary gene pool. For gene content, *L. virosa* harbors a large number of genes absent from *L. saligna* and *L. sativa* and may thus constitute an important source of novel genes for lettuce breeding. Based on synteny, a three-way genome comparison uncovered species-specific major inversions. These inversions should be considered as likely barriers to gene introgression in future breeding. In addition, we demonstrated the genome expansion in *L. virosa* is driven by the proliferation of LTR elements. An assembly-based comparison of *NLR* and *RLK* genes between *Lactuca* spp. found more immune system-related genes in the *L. sativa* genome than in those of the *L. virosa* and *L. saligna* genomes. These findings may contribute to future research on gene expression and regulation in *L. virosa.* Using this novel genome assembly, researchers can subsequently study the genetic variation in *L. virosa* populations to fully release its potential for lettuce breeding.

5. Data availability

The genome assembly of *L. virosa*, is available under the BioProject PRJEB50301. All raw sequencing reads have been deposited in the ENA database under BioProject PRJEB56289. This includes the Illumina, PacBio, Bionano, and Hi-C whole-genome sequences as well as RNA sequencing data for genome annotation.

6. Acknowledgement

This research was supported by a grant of the International Lettuce Genomics Consortium (ILGC) funded by the Top Consortium for Knowledge and Innovation Horticultural and Starting Materials (grant number 1406-039). W.X. was financially supported by a fellowship by the China Scholarship Council (CSC). We thank Jiri Macas and Floris Breman for their help with RepeatExplorer2 analysis. We also thank Elizabeth Marie Georgian for writing support.

7. References

Ammiraju, J.S.S., Zuccolo, A., Yu, Y., et al. (2007) Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus Oryza. Plant J., 52, 342–351.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S. and Ogata, H. (2019) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics, 36, 2251–2252. Available at: https://doi.org/10.1093/bioinformatics/btz859.

Bandi, V. and Gutwin, C. (2020) Interactive exploration of genomic conservation. In In Proceedings of the 46th Graphics Interface Conference on Proceedings of Graphics Interface 2020 (GI'20). Waterloo.

Bao, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA, 6, 11. Available at: http://www.mobilednajournal.com/content/6/1/11 [Accessed May 2, 2020].

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, 30, 2114–2120. Available at: https://academic.oup.com/bioinformatics/article-abstract/30/15/2114/2390096 [Accessed February 11, 2020].

Buchfink, B., Xie, C. and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. Nat. Methods, 12, 59–60. Available at: http://www.nature.com/articles/nmeth.3176 [Accessed May 6, 2019].

Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25, 1972–1973. Available at: https://academic-oup-com.ezproxy.library.wur.nl/bioinformatics/article/25/15/1972/213148 [Accessed December 4, 2021].

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: Ultrafast universal RNA-seq aligner. Bioinformatics, 29, 15–21. Available at: https://academic.oup.com/bioinformatics/article-abstract/29/1/15/272537 [Accessed February 11, 2020].

Doležalová, I., Lebeda, A., Janeček, J., Číhalíková, J., Křístková, E. and Vránová, O. (2002) Variation in chromosome numbers and nuclear DNA contents in genetic resources of *Lactuca* L. species (Asteraceae). Genet. Resour. Crop Evol., 49, 383–395.

Eenink, A.H., Groenwold, R. and Dieleman, F.L. (1982) Resistance of lettuce (*Lactuca*) to the leaf aphid Nasonovia ribis nigri. 1. Transfer of resistance from L. virosa to L. sativa by interspecific crosses and selection of resistant breeding lines. Euphytica, 31, 291–299.

El-Gebali, S., Mistry, J., Bateman, A., et al. (2018) The Pfam protein families database in 2019. Nucleic Acids Res., 47, D427–D432.

Fertet, A., Graindorge, S., Koechler, S., Boer, G.-J. de, Guilloteau-Fonteny, E. and Gualberto, J.M. (2021) Sequence of the Mitochondrial Genome of *Lactuca* virosa Suggests an Unexpected Role in *Lactuca sativa*'s Evolution. Front. Plant Sci., 12, 1565. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2021.697136/full [Accessed August 2, 2022].

Finn, R.D., Clements, J. and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res., 39, W29–W37. Available at: https://doi.org/10.1093/nar/gkr367.

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

Food and Agriculture Organization of the United Nations (2019) FAOSTAT. FAOSTAT. Available at: https://www.fao.org/faostat/en/#home [Accessed May 18, 2019].

Guo, Z., Li, B., Du, J., et al. (2022) LettuceGDB: the community database for lettuce genetics and omics. Plant Commun., 100425.

Han, Y. and Wessler, S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res., 38, e199–e199. Available at: https://doi.org/10.1093/nar/gkq862.

Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M. and Stanke, M. (2016) BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics, 32, 767–769. Available at: https://academic.oup.com/bioinformatics/article/32/5/767/1744611 [Accessed May 2, 2020].

Hok, S., Danchin, E.G.J., Allasia, V., Panabiéres, F., Attard, A. and Keller, H. (2011) An Arabidopsis (malectin-like) leucine-rich repeat receptor-like kinase contributes to downy mildew disease. Plant. Cell Environ., 34, 1944–1957. Available at: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-3040.2011.02390.x [Accessed September 1, 2022].

Holt, C. and Yandell, M. (2011) MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics, 12, 491.

Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F.A. and Wheeler, T.J. (2015) The Dfam database of repetitive DNA families. Nucleic Acids Res., 44, D81–D89. Available at: https://doi.org/10.1093/nar/gkv1272.

Jonkheer, E.M., Workum, D.-J.M. van, Sheikhizadeh Anari, S., Brankovics, B., Haan, J.R. de, Berke, L., Lee, T.A.J. van der, Ridder, D. de and Smit, S. (2022) PanTools v3: functional annotation, classification and phylogenomics T. Marschall, ed. Bioinformatics. Available at: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac506/6647839 [Accessed August 2, 2022].

Kanehisa, M. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res., 28, 27–30.

Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: A fast spliced aligner with low memory requirements. Nat. Methods, 12, 357–360. Available at: http://www.ccb.jhu.edu/ [Accessed July 20, 2020].

Kim, S., Park, J., Yeom, S.I., et al. (2017) New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. Genome Biol., 18, 210. Available at: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1341-9 [Accessed September 20, 2022].

Kopylova, E., Noé, L. and Touzet, H. (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics, 28, 3211–3217. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts611 [Accessed July 31, 2022].

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. (2017) Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res., 27, 722–736.

Krassowski, M. (2020) ComplexUpset. Available at: https://github.com/krassowski/complex-upset [Accessed September 21, 2022].

Krogh, A., Larsson, B., Heijne, G. Von and Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J. Mol. Biol., 305, 567–580.

Letunic, I. and Bork, P. (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res., 49, W293–W296. Available at: https://pubmed.ncbi.nlm.nih.gov/33885785/ [Accessed December 4, 2021].

Li, H. (2018) Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics, 34, 3094–3100. Available at: https://academic.oup.com/bioinformatics/article-abstract/34/18/3094/4994778 [Accessed August 7, 2020].

Liu, P.L., Huang, Y., Shi, P.H., Yu, M., Xie, J.B. and Xie, L. (2018) Duplication and diversification of lectin receptor-like kinases (LecRLK) genes in soybean. Sci. Rep., 8, 1–14.

Ma, J. and Bennetzen, J.L. (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc. Natl. Acad. Sci. U. S. A., 101, 12404–12410. Available at: https://pnas.org/doi/full/10.1073/pnas.0403715101 [Accessed August 4, 2022].

Maisonneuve, B. (2003) *Lactuca* virosa, a source of disease resistance genes for lettuce breeding: results and difficulties for gene introgression, Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.573.6189&rep=rep1&type=pdf [Accessed May 3, 2019].

Maisonneuve, B., Chovelon, V. and Lot, H. (2022) Inheritance of Resistance to Beet Western Yellows Virus in *Lactuca* virosa L. HortScience, 26, 1543–1545.

Maisonneuve, B., Chupeau, M.C., Bellec, Y. and Chupeau, Y. (1995) Sexual and somatic hybridization in the genus *Lactuca*. Euphytica, 85, 281–285.

Maisonneuve, B., Pitrat, M., Gognalons, P. and Moury, B. (2018) Growth stage-dependent resistance to the potyviruses lettuce Italian necrotic virus and Lettuce mosaic virus displayed by *Lactuca sativa* introgression lines carrying the Mo3 locus from L. virosa. Plant Pathol., 67, 2013–2018. Available at: https://onlinelibrary.wiley.com/doi/10.1111/ppa.12909 [Accessed November 21, 2022].

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A. and Zdobnov, E.M. (2021) BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol. Biol. Evol., 38, 4647–4654. Available at: https://academic.oup.com/mbe/article/38/10/4647/6329644 [Accessed March 9, 2022].

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. and Clavijo, B.J. (2016) KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics, 33, btw663. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw663 [Accessed July 31, 2022].

Marçais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 27, 764–770. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr011 [Accessed August 7, 2019].

McDonnell, A. V, Jiang, T., Keating, A.E. and Berger, B. (2006) Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics, 22, 356–358. Available at: https://doi.org/10.1093/bioinformatics/bti797.

Mikel, M.A. (2007) Genealogy of contemporary North American lettuce. HortScience, 42, 489–493. Available at: https://journals.ashs.org/hortsci/view/journals/hortsci/42/3/article-p489.xml [Accessed October 9, 2019].

Mikel, M.A. (2013) Genetic composition of contemporary proprietary U.S. lettuce (*Lactuca sativa* L.) cultivars. Genet. Resour. Crop Evol., 60, 89–96. Available at: http://link.springer.com/10.1007/s10722-012-9818-6 [Accessed August 19, 2019].

Neumann, P., Novák, P., Hoštáková, N. and MacAs, J. (2019) Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mob. DNA, 10, 1–17. Available at: https://mobilednajournal.biomedcentral.com/articles/10.1186/s13100-018-0144-1 [Accessed March 11, 2022].

Nguyen, L.T., Schmidt, H.A., Haeseler, A. Von and Minh, B.Q. (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol., 32, 268–274. Available at: https://pubmed.ncbi.nlm.nih.gov/25371430/ [Accessed October 9, 2020].

Novák, P., Neumann, P. and Macas, J. (2020) Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nat. Protoc. 2020 1511, 15, 3745–3776. Available at: https://www.nature.com/articles/s41596-020-0400-y [Accessed March 11, 2022].

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. and Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol., 33, 290–295.

Peters, C., Tsirigos, K.D., Shu, N. and Elofsson, A. (2016) Improved topology prediction using the terminal hydrophobic helices rule. Bioinformatics, 32, 1158–1162. Available at: https://pubmed.ncbi.nlm.nih.gov/26644416/ [Accessed April 1, 2021].

Pi, L., Yin, Z., Duan, W., Wang, N., Zhang, Y., Wang, J. and Dou, D. (2022) A G-type lectin receptor-like kinase regulates the perception of oomycete apoplastic expansin-like proteins. J. Integr. Plant Biol., 64, 183–201.

Piegu, B., Guyot, R., Picault, N., et al. (2006) Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res., 16, 1262–1269.

Putnam, N.H., Connell, B.O., Stites, J.C., Rice, B.J., Hartley, P.D., Sugnet, C.W., Haussler, D. and Rokhsar, D.S. (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res., 26, 342–350.

Qin, M., Wu, S., Li, A., Zhao, F., Feng, H., Ding, L. and Ruan, J. (2019) LRScaf: Improving draft genomes using long noisy reads. BMC Genomics, 20, 955. Available at: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6337-2 [Accessed September 13, 2021].

Ranallo-Benavidez, T.R., Jaron, K.S. and Schatz, M.C. (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun., 11, 1–10.

Reyes-Chin-Wo, S., Wang, Z., Yang, X., et al. (2017) Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. Nat. Commun., 8, 14953. Available at: http://www.nature.com/doifinder/10.1038/ncomms14953.

Roach, M.J., Schmidt, S.A. and Borneman, A.R. (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics, 19, 460. Available at: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2485-7 [Accessed September 13, 2021].

Rodgers-Melnick, E., Mane, S.P., Dharmawardhana, P., Slavov, G.T., Crasta, O.R., Strauss, S.H., Brunner, A.M. and DiFazio, S.P. (2012) Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus. Genome Res., 22, 95–105.

Simko, I. (2013) Marker-assisted selection for disease resistance in Lettuce. In Translational Genomics for Crop Breeding, Volume I: Biotic Stress. wiley, pp. 267–289. Available at: https://onlinelibrary-wiley-com.ezproxy.library.wur.nl/doi/full/10.1002/9781118728475.ch14 [Accessed June 23, 2021].

Simko, I., Peng, H., Sthapit Kandel, J. and Zhao, R. (2022) Genome-wide association mapping reveals genomic regions frequently associated with lettuce field resistance to downy mildew. Theor. Appl. Genet., 135, 2009–2024.

Smit, A., Hubley, R. and Green, P. (2019) 2013–2015. RepeatMasker Open-4.0.

Sthapit Kandel, J., Peng, H., Hayes, R.J., Mou, B. and Simko, I. (2020) Genome-wide association mapping reveals loci for shelf life and developmental rate of lettuce. Theor. Appl. Genet., 133, 1947–1966.

Thompson, R.C. and Ryder, E.J. (1961) Description and pedigrees of nine varieties of lettuce. Tech. Bull. Res. Serv. Agric., 1244, 1–19.

Vasimuddin, M., Misra, S., Li, H. and Aluru, S. (2019) Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In Proceedings- 2019 IEEE 33rd International Parallel and Distributed Processing Symposium, IPDPS 2019. Institute of Electrical and Electronics Engineers Inc., pp. 314–324.

Walker, B.J., Abeel, T., Shea, T., et al. (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement J. Wang, ed. PLoS One, 9, e112963. Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112963 [Accessed September 13, 2021].

Walley, P.G., Hough, G., Moore, J.D., et al. (2017) Towards new sources of resistance to the currant-lettuce aphid (Nasonovia ribisnigri). Mol. Breed., 37.

Wang, P., Xu, X., Zhao, G., et al. (2020) Genetic mapping and candidate gene analysis for melon resistance to Phytophthora capsici. Sci. Rep., 10, 1–11.

Wang, Y., Tang, H., DeBarry, J.D., et al. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res., 40, e49–e49. Available at: https://doi.org/10.1093/nar/gkr1293.

Wei, T., Treuren, R. van, Liu, Xinjiang, et al. (2021) Whole-genome resequencing of 445 *Lactuca* accessions reveals the domestication history of cultivated lettuce. Nat. Genet., 53, 752–760.

Wendel, J.F., Jackson, S.A., Meyers, B.C. and Wing, R.A. (2016) Evolution of plant genome architecture. Genome Biol., 17, 1–14.

Wu, J.-Y., Xue, J.-Y. and Peer, Y. Van de (2021) Evolution of NLR Resistance Genes in Magnoliids: Dramatic Expansions of CNLs and Multiple Losses of TNLs. Front. Plant Sci., 12, 2998. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2021.777157/full [Accessed September 16, 2022].

Xiong, W., Berke, L., Michelmore, R., et al. (2022) The genome of *Lactuca* saligna, a wild relative of lettuce, provides insight into non-host resistance to the downy mildew Bremia lactucae. bioRxiv, 2022.10.18.512484.

Yu, G., Wang, L.G., Han, Y. and He, Q.Y. (2012) ClusterProfiler: An R package for comparing biological themes among gene clusters. Omi. A J. Integr. Biol., 16, 284–287.

Zdobnov, E.M. and Apweiler, R. (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 17, 847–848.

Zhang, X., Liu, T., Wang, J., et al. (2021) Pan-genome of Raphanus highlights genetic variation and introgression among domesticated, wild, and weedy radishes. Mol. Plant, 14, 2032–2055.

Zhao, Q., Feng, Q., Lu, H., et al. (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat. Genet., 50, 278–284.

8. Supplementary materials

Data is available via:10.4121/21803436

Supplementary Figures

Supplementary Figure 1. Genome size estimation of *L. virosa* by GenomeScope.
Supplementary Figure 2. K-mer and BUSCO completeness plots for the first *L. virosa* assembly.
Supplementary Figure 3. Functional enrichment of genes unique to *L. virosa*.
Supplementary Figure 4. Link density histogram (generated by Dovetail Genomics).
Supplementary Figure 5. Genome composition of three *Lactuca* species.
Supplementary Figure 6. LTR distribution at gene flanking regions (+/- 5 kb) in the *L. virosa* genome.
Supplementary Figure 7. Example cluster of repeat reads from RepeatExplorer.
Supplementary Figure 8. Circular tree of NLRs for *Lactuca* species generated by IQTREE.

Supplementary Tables

Supplementary Table 1. Summary statistics of DNA and RNA sequencing.
Supplementary Table 2. Overview of all statistics of *L. virosa* during the process of improving the assembly.
Supplementary Table 3. Summary of the different *L. virosa* assemblies used in this study.
Supplementary Table 4. Functional annotation summary.
Supplementary Table 5. Tentative linkage groups (LG) of *L. virosa* scaffolds.
Supplementary Table 6. Percent of sequence for different types of repeat elements in *Lactuca* spp. genomes.
Supplementary Table 7. Summary of individual and comparative RepeatExplorer analysis.
Supplementary Table 8. Genomic proportion of annotated clusters for individual and comparative analyses.
Supplementary Table 9. Genomic proportion of six groups after hierarchical clustering for annotated repeat clusters.
Supplementary Table 10. Summary of NLR domain search for three *Lactuca* spp..
Supplementary Table 11. Pfam HMM motifs used for RLK classification.
Supplementary Table 12. RLK classification based on extracellular domain.

Supplementary Data

Supplementary Data 1. Genome assembly and scaffolding.
A, The input assembly scaffolds in the Hirise scaffolds using Hi-C. **B**, The input assembly scaffolds in the final scaffolds (WUR + BGI). **C**, The sequence length of scaffolds in final assembly (WUR + BGI).

Supplementary Data 2. Functional annotation and homology grouping of *L. virosa* transcripts.
A, Detailed match of functional annotation by different approaches for *L. virosa* transcripts for all isoforms. **B**. Detailed match of functional annotation by different approaches for *L. virosa* genes. **C**, Homology groups of the three *Lactuca* species calculated by PanTools. **D**, InterPro enrichment of *L. virosa* specific homologs.

Supplementary Data 3. Repeatome analysis of *L. virosa*, *L. sativa* and *L. saligna*.
A, Summary of RepeatMasker output for the three genomes. **B**, Matrix and cumulative stats of RepeatExplorer clusters. **C**, Curated annotation and genomic proportion of RepeatExplorer clusters excluding organelle reads.

Supplementary Data 4. Identified NLR and RLK proteins in *L. virosa* and *L. saligna*.
A, Identification and classification of NLR (*L. virosa*). **B**, Identification and classification of NLR (*L. sativa*). **C**, Identification and classification of RLK (*L. virosa*). **D**, Identification and classification of RLK (*L. sativa*).

Supplementary Data 5. Overview of homology within NLRs and RLKs in *Lactuca*.
A, Homolog groups of identified NLRs and RLKs for three *Lactuca* species. **B**, Tandem array detection of identified NLRs and RLKs for three *Lactuca* species.

Chapter 4

Phylogenomic analysis of the *de novo* genome and transcriptome of dandelion (*Taraxacum officinale*) provide insights of *MADS-box* and *TCP* gene diversification and floral development of the Asteraceae

Wei Xiong[1,+], Judith Risse[2, 3,+], Lidija Berke[1, 4], Tao Zhao[1, 5], Henri van de Geest [2, 6], Carla Oplaat[1, 7], Marco Busscher[1, 8], Ingrid van der Meer[1, 8], Koen Verhoeven[3], M. Eric Schranz[1], Kitty Vijverberg[1]

[1] Biosystematics Group, Wageningen University and Research, Wageningen, The Netherlands.
[2] Bioinformatics Group, Wageningen University and Research, Wageningen, The Netherlands.
[3] Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands
[4] Present address: Genetwister Technologies B.V., Wageningen, The Netherlands
[5] Present address: State Key Laboratory of Crop Stress Biology for Arid Areas, Northwest A&F University, Yangling, Shaanxi, China
[6] Present address: Hudson River Biotechnology, Wageningen, The Netherlands
[7] Present address: National Plant Protection Organization, National Reference Centre of Plant Health (NVWA), Wageningen, the Netherlands
[8] Bioscience, Wageningen University and Research, Wageningen, The Netherlands
+ Contributed equally.

This chapter has been submitted and is currently under review.

Abstract

The Asteraceae is the largest angiosperm family with more than 25,000 species. Individual studies have shown that MADS-box and TCP transcription factors regulate the development and symmetry of flowers, contributing to their iconic flower-head (capitulum) and floret development. However, a systematic study of *MADS-box* and *TCP* genes across the Asteraceae is lacking. We performed a comparative analysis of genome sequences of 33 angiosperm species (12 Asteraceae) including our *de novo* assembly of diploid sexual dandelion (*Taraxacum officinale*) to investigate the lineage-specific evolution of *MADS-box* and *TCP* genes in the Asteraceae. We compared the phylogenomic results of *MADS-box* and *TCP* genes with their expression in *T. officinale* floral tissues at different stages to demonstrate the regulation of genes with Asteraceae specific attributes. Here, we show that *MADS-box MIKC*^{c} and *TCP-CYC* genes have expanded in the Asteraceae. The phylogenomic analysis identified *AGAMOUS*-Like (*STK*-Like), *SEPALATA*-Like (*SEP3*-Like), and *TCP-PCF*-Like copies with lineage-specific contexts in the Asteraceae or dandelion. Different expression patterns of some of these gene copies hint at functional divergence. We also confirm the presence and revisit the evolutionary history of previously named "*Asteraceae Specific MADS-box* genes (*AS-MADS*)." Specifically, we identify non-Asteraceae homologs, indicating a more ancient origin of this gene clade. Syntenic relatsionships support that *AS-MADS* is paralogous to *FLC* as demonstrated by the shared ancient duplication of *FLC* and *SEP3*.

Key words

Asteraceae, dandelion, *de novo* sequencing, floral development, *MADS-box* gene, phylogenomics, *TCP* gene

1. Introduction

The Asteraceae (Compositae) are one of the largest and most diverse families of angiosperms, with great ecological and economic importance. It contains ~25,000 species, which represents 10% of extant flowering plants (Mandel et al., 2019). The Asteraceae is subdivided into 16 subfamilies, including two large crown-groups, the Asteroideae (e.g., sunflower, daisy) and Cichorioideae (e.g., lettuce, dandelion) (Chase et al., 2016; Stevens, 2017; Susanna et al., 2020). Members of the Asteraceae inhabit an incredible range of ecosystems varying in climates and landscapes, on every continent (Smith and Richardson, 2010; Folk et al., 2020). Their global distribution makes Asteraceae plants interesting models for various questions in ecology and evolution (Shen et al., 2021; Palazzesi et al., 2022). For humans, the Asteraceae are of considerable societal and economic value including ornamentals (e.g., Gerbera, Chrysanthemum), medicines (e.g., sweet wormwood, chamomile) and crops (e.g., sunflower, lettuce); but also includes many well-known weedy species (e.g., groundsel, dandelion). Genome assemblies can facilitate the study of the molecular and evolutionary basis of ecological and economic traits. To date, most sequenced Asteraceae species are ornamentals and crops.

The unique floral and fruit traits of the Asteraceae, including the representative flower heads (capitula) and one-seeded dry fruits (cypsela), often with a hairy or scaly pappus, underlie much of the diversity and evolutionary and ecological success of the group (Panero and Funk, 2008; Mandel et al., 2019). The capitulum is one of the most iconic floral features of the Asteraceae, a highly compressed inflorescence with many closely packed flowers, named 'florets', that resembles a flower (Elomaa et al., 2018). There are three major floret types in Asteraceae: disc (tube), ray (2-3 lobed) and ligulate (5-lobed) (Anderberg et al., 2007), which are discriminatory to the subfamilies, particularly the Asteroideae are characterized by disc florets ± one or more rows of ray florets and the Cichorioideae by ligulate florets (Carlquist, 1976). In addition, the pappus, a highly modified calyx (Vijverberg et al., 2021), is another striking characteristic of the Asteraceae. It assists in seed dispersal and can protect against herbivores and aid in water uptake to facilitate germination (Carlquist, 1976; Stuessy and Garver, 1996; Jana and Mukherjee, 2012). Understanding the genetic basis of capitulum formation and floral and fruit characteristics is, therefore, of large interest to disclose the evolutionary success of the Asteraceae.

Whole-genome duplications (WGDs) have likely played a critical role in boosting the diversity of the Asteraceae (Barker et al., 2008), similar to other angiosperm lineages (Ohno, 1970; De Bodt et al., 2005; Magadum et al., 2013). In the Asteraceae, two paleopolyploid events occurred preceding their major radiation (Barker et al., 2016), and more recent WGDs occurred in major tribes and subfamilies (Huang et al., 2016;

Shen et al., 2021). After WGDs, the additional gene copies may retain their original function (redundant copies) or undergo sub- and neo-functionalization events (Panchy et al., 2016). Moreover, genes in a new genomic context (i.e., gene order disruption) may result in novel (*cis*) gene regulation (Ilic et al., 2003; Langham et al., 2004; Lockton and Gaut, 2005). The *MADS-box* and *TCP* transcription factors are among the most important regulators of floral organ determination and development. Polyploidization has resulted in expanded *MADS-box* and *TCP* gene families, which have been shown to contribute to the evolution of the capitulum, floral and fruit traits in the Asteraceae in different studies (see below). They are both included in this study to reveal the link between their evolution, genome context, expression and phenotype.

The *MADS-box* gene family consists of two major clades: Type I, with a conserved N-terminal MADS DNA binding domain (M), and Type II, containing an M-domain, a less conserved Intervening domain (I), a conserved Keratin-like coiled-coil domain (K-box), and a highly variable, often species specific, C-terminal domain (Theißen et al., 1996; Alvarez-Buylla et al., 2000; Smaczniak et al., 2012). Type II *MADS-box* genes are also known as *MIKC*-genes and can be further subdivided into MIKC^c and MIKC* types (Henschel et al., 2002). *MIKC^c* genes comprise several sub-groups including the well-known ABC(D)E genes crucial for floral organ initiation and development (Becker and Theißen, 2003; Theißen et al., 2016). Research results on Asteraceae floral development mainly come from the classical model gerbera (Mutisioideae; Zhang et al., 2017) and more recently from crops such as lettuce (Cichorioideae; Ning et al., 2019), sunflower (Asteroideae; Dezar, 2003) and chrysanthemum (Asteroideae; Won et al., 2021). For example, in gerbera, eight *SEPALLATA*-like (*SEP*-like; class E) genes were found, whereas Arabidopsis has only four *SEP*-like genes (Zhang et al., 2017). Unlike the redundancy of *SEP* copies in Arabidopsis, the different *SEP*-like genes in gerbera show sub-functionalization in floral organ development and neo-functionalization in the inflorescence meristem (IM) beside conserved functions (Elomaa et al., 2018). Genome-wide analysis of *MADS-box* genes in chrysanthemum and lettuce identified an Asteraceae-specific *MADS-box* clade, named *Asteraceae Specific-MADS* (*AS-MADS*), of which the evolution and function is still unclear (Won et al., 2021).

All *TCP* genes contain a highly conserved *basic HELIX LOOP HELIX* (*bHLH*) domain and on which they are divided into Class I (P) and Class II (C) (Kosugi and Ohashi, 1997; Navaud et al., 2007; Li, 2015). Class I *TCP* genes represent the *PCF* genes, while class II *TCP* genes are divided into the ubiquitous *CIN* genes and angiosperm specific *CYC/TB1* genes (Martín-Trillo and Cubas, 2010; Nath et al., 2003; Luo et al., 1996; Doebley et al., 1997). Among them, *CYC/TB1* genes are closely associated with the regulation of flower symmetry (e.g., in *Antirrhinum majus;* (Luo et al., 1996). Studies of *Senecio* (Asteraceae) showed that the *CYC2*-like genes *RAY*1 and *RAY*2 are involved in the development of ray florets (Kim et al., 2008). A role of different *CYC*2 homologs in the formation of ray and disc florets in distinct Asteraceae lineages visualizes neo-functionalization in *CYC* genes, which underwent several duplications in Asteraceae (Elomaa et al., 2018). An extensive study within the Asteraceae further confirmed that the developmental program of making a ray flower involves functionally divergent *CYC*2-like genes in different lineages within the Asteraceae (Chen et al., 2018). However, the function of *CYC* in the formation of ligulate florets is yet unconfirmed. Another overlooked element of the *TCP* genes is the function of *PCF* genes (Kosugi and Ohashi, 1997), which participate in a wide range of plant growth, including flower development. The increasing number of sequenced genomes presents us with an opportunity to conduct a systematic analysis of these important *MADS-box* and *TCP* gene families in a wide range of Asteraceae species.

To study the evolution of *MADS-box* and *TCP* families and their effects on Asteraceae floral traits, a family-based phylogenomic analysis is required to gain more knowledge about the history of gene retention after Asteraceae radiation-related WGDs. Moreover, the patterns of gene movement (transpositions) could help identify potential sources of regulatory novelty induced by genomic context change. Thus, a broad range of genomic comparisons, like synteny network analysis (Zhao and Schranz, 2017), are valuable to conduct alongside phylogenetic analysis. Because synteny can help determine the orthologous relationships after complex WGDs and identify other genomic positional changes, like ancient tandem duplications and gene transpositions (Dewey, 2011; Zhao et al., 2017).

In this study, we used the common dandelion (*Taraxacum officinale*; Figure 1), a member of the Cichorioideae and taxonomic outgroup of lettuce, as a model. Dandelion is well-studied because of its two reproduction modes that co-occur within its distribution range: sexual diploids (*2n* = *2x* = 16) and asexual, apomict, triploids (*2n* = *3x* = 24) (Van Dijk et al., 1999). For example, studies of molecular genetic basis for different apomixis elements, including diplospory (Vijverberg et al., 2004, 2010) and parthenogenesis (Vijverberg et al., 2019; Van Dijk et al., 2020; Underwood, Vijverberg, Rigola et al., 2022). Apart and connected to this, dandelion has been investigated for its ecological evolution and adaption (Brock et al., 2005); Verhoeven et al. 2018) and more recently for its aforementioned floret and fruit characteristics (Vijverberg et al., 2021). A genome assembly of this interesting model species will provide insights into its gene and genome evolution and serve as an important *reference* for comparative analysis within the Asteraceae, other *Taraxacum* species and genotypes, and related species such as lettuce, and for gene analysis and gene editing purposes.

Here we analyzed the published whole genome sequences of 32 species plus our *de novo* assembly of the *T. officinale* genome. We performed genome-wide searches for *MADS-box* and *TCP* genes of the 33 species analyzed (Figure 1A). We constructed a synteny network of the identified *MADS-box* and *TCP* genes to reveal the lineage-specific context of the genes and ancient tandem duplications, with a focus on the Asteraceae, its subclades Asterioideae and Cichorioideae, and dandelion. We examined the synteny on phylogenetic trees based on MADS-box (Figure 1B) and TCP domain sequences and assessed a possible change in function after gene duplication or genomic context change via comparison to gene expression data in different floral stages and tissues (Figure 1C). We also applied phylogenomic data to characterize the evolution of the *Asteraceae Specific-MADS* (*AS-MADS*) genes and their expression during flower development. Our results provide insights into the evolution of Asteraceae and their *MADS-box* and *TCP* genes, while the wealth of genome and transcriptome data serves as a reference for future comparative analyses and research on floral development in dandelion and beyond.

2. Results

2.1 Genome sequencing and Assembly

The *T. officinale* genome of the sexual diploid plant FCh72 was sequenced with PacBio RSII and 10X Genomics on Illumina HiSeq2500, and optically mapped with BioNano. We obtained an average of 75x coverage of PacBio reads with a mean subread length of 12,259 bp and assembled them using Canu v1.3 (Koren et al., 2017). The assembly was scaffolded with the 10X and BioNano data and polished with the 10X Illumina reads. Haplo-contigs were then collapsed where possible and the assembly was polished and scaffolded multiple times in subsequent rounds (see Materials & Methods). The resulting assembly has a total genome size of 936 Mb (Table 1; Supplementary Table S1), which is slightly larger than the expected 831 Mb based on C-values (cvalues.science.kew.org/search/angiosperm, 2/3th for a diploid sexual plant), and significantly larger than the estimated genome size based on kmer analysis (~614 Mb; Supplementary Figure S1). Blobtools confirmed the absence of contamination (Supplementary Figure S2). The assembly is of draft genome quality, with 4,059 scaffolds, an N50 size of 757 kb and the longest scaffold of ~23 Mb (Supplementary Table S1). The GC content is 37.0%. The mitochondrial genome was assembled in a single scaffold that showed high homology to the mtDNA of the related species lettuce (Supplementary Figure S3), whereas the chloroplast genome has not been recovered, probably as a result of bleaching prior to the harvesting of plant material. Difficulties in assembling were posed by the heterozygosity of the genome, which was estimated at 1.5% with GenomeScope, showing two clear kmer peaks (Supplementary Figure S1; k = 21). BUSCO quality assessment of the genome assembly showed 95.4 % completeness, with 2,219/2,326 (75.6%) complete and single copy and 460/2326 (19.8%) complete and duplicated genes. This indicated that, despite all our efforts, we have not been able to collapse all haploid contigs, conforming to the high heterozygosity of the dandelion genome, leaving ~20% of the genome as allelic.

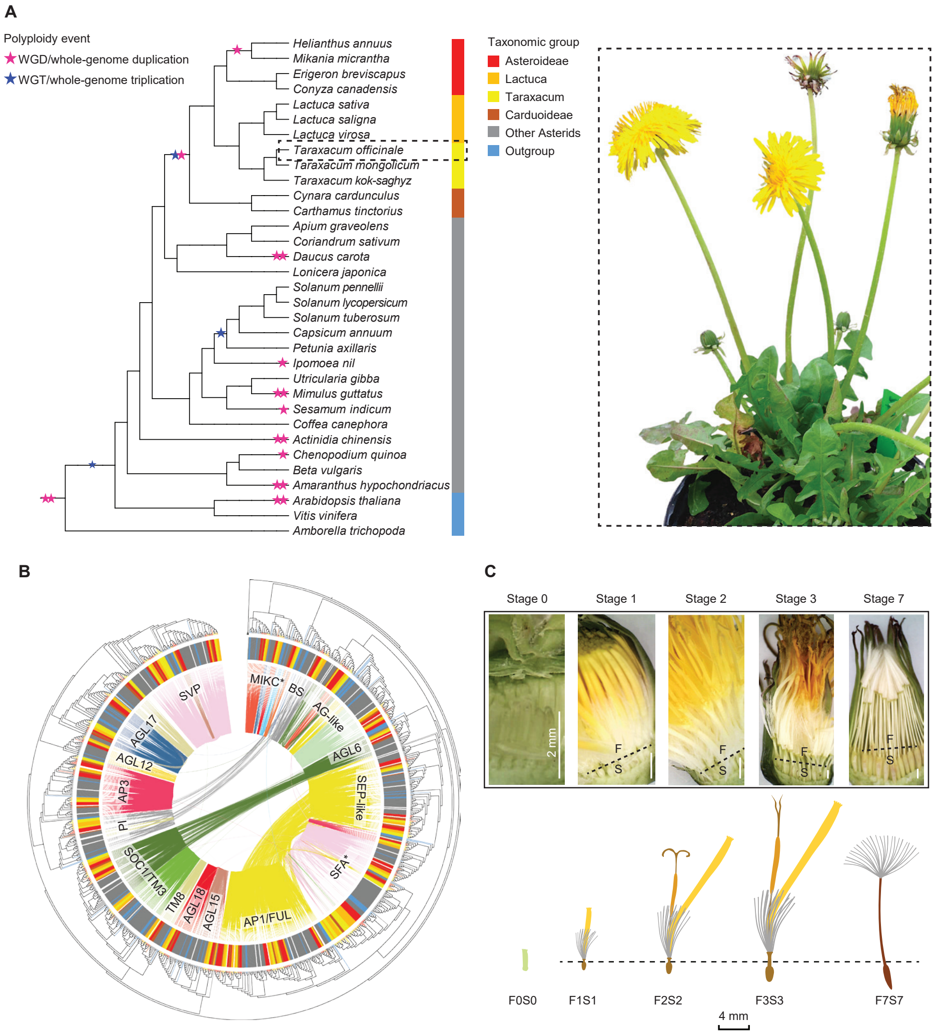


Figure 1. Overview of our study. **A,** Summarized phylogeny of the Angiosperms, with a focus on Asteraceae and the position of Taraxacum therein, and with the ancestral whole genome duplications and triplications indicated (left) and a dandelion plant (right). **B,** Phylogenetic tree of Type II *MADs-box* genes based on MADS and K-box domain protein sequences. **C.** Dandelion floral tissues and stages used in the gene expression analysis: F = Upper floral part and S = lower floral part, separated through the beak (dotted line) except for the youngest stage (F0S0); Stage 1 = bud just before opening; 2 = open flower; 3 = 3 days after pollination (3 DAP); 7 = 7 DAP. © Kitty Vijverberg (**A, C**) and Wei Xiong (**C**).

Table 1. Main characteristics of *T. officinale* genome.

Genome assembly and annotation	**Statistics**
Assembly size (Mb)	936
Expected genome size (Mb)	831
Number of scaffolds	4,059
N50 Super-scaffolds (Kb)	757
Heterozygosity (%)	1.5
BUSCO completeness (%)	95.4
Repeats (%)	63
Predicted high confident genes	64,089
Functional annotated transcripts	56,560
Sequence identical genes (%)	2.8
Protein >99% similar genes (%)	8.0

2.2 Genome Annotation

The assembled genome was repeat-masked using RepeatModeler with Long Terminal Repeat (LTR) detection and using RepeatMasker. In total 63% of all bases were masked, which is in line with a previous study (Ferreira de Carvalho et al., 2016a) and corresponding to the repeat content in *T. mongolicum* and *T. kok-saghyz* (Lin et al., 2021). The repeat content was to a large extent driven by LTRs, namely *Copia* (~214 Mb, 22.9% of the genome) and *Gypsy* (~135 Mb, 14.5%) retrotransposons (Supplementary Table S2).

We annotated the genome using an RNAseq library based on four different tissue types of the sequenced dandelion genotype: leaf, bud, open flower and roots, using BRAKER2. A total of 64,089 high confident genes (i.e., size >= 150 amino acids [aa] or >=50 aa with homology annotation) with 66,956 transcripts was found (Supplementary Table S3; Supplementary Data S1). The mean gene length was 2,110 bp with on average 4.7 exons of mean total CDS length of 971 bp (Supplementary Table S3). For 88.2 % of the genes (56,560) the transcripts have a description of which 68.3 % (43,771) are associated with at least one Gene Ontology (GO) term (Supplementary Table S3).

A total of 1,787 high-quality genes (2.8 %) were found that had at least one identical sequence copy in the annotation and 5,147 genes (8.0 %) showed more than 99 % amino acid identity with another annotated gene (Supplementary Table S3; indicated in Supplementary Data S1), which are either sequence duplicates, true duplicates, closely related family members or alleles. The most abundant genes showed 15 and 11 copies, respectively, representing *Histone H*4 and *GOS*9-like isoforms (Supplementary Table S3; Supplementary Data S1). BUSCO analysis of the translated transcripts showed 90 % completeness with 19.4 % duplicated BUSCOs.

An unfiltered gene set that includes the high confidence gene models as well as smaller transcripts (50-100 aa) and genes without homology annotation of 81,291 genes in total with 85,093 transcripts was used in the gene expression analyses and synteny mapping results (see below).

2.3 Genome comparison between *Taraxacum* spp. assemblies

The *T. officinale* genome assembly was compared to those of the recently published whole genome sequences of two other sexual diploid Taraxacum species, *T. mongolica* (*Tmo*) and *T. kok-saghyz* (*Tks*) (Lin et al., 2021), showing a relatively fragmented assembly (4,059 scaffolds versus 65 in *Tmo* and 160 in *Tks*; Supplementary Table S4). The annotation of gene space was, however, far more complete in *T. officinale* based on the BUSCO results (90 % completeness versus 69 % in *Tmo* and 74 % in *Tks*). The GC content of 37 % was similar to the other two species, whereas the heterozygosity varied from 1 % (*Tks*) to 1.5 % (*Tof*). The assemblies are collinear without major structural rearrangements if compared by alignments and dotplots (Supplementary Figure S4).

2.4 Expression analysis of floral tissues and stages

A total of 25 samples of the sexual diploid dandelion plant FCh72, including different floral tissues and stages (Figure 1c), were analyzed for gene expression to obtain a global overview of the genetic basis underlying floral development. The samples included triplicates of very young whole buds (F0S0; organs initiating), older buds just before opening (F1, S1) and open flowers (F2, S2), the latter two stages with the florets separated into an upper (F) and lower (S) part by cutting through the beak (Figure 1c; see for exact stages and method Vijverberg et al. 2021). In addition, duplicates of the two floral parts at three days after pollination (3 DAP; F3, S3) and 7 DAP (F7, pappus only; S7, ripening seeds) and leaves (LF) were analyzed. RNA sequencing generated on average 33.8 million read pairs per sample (Supplementary Table S5) of which >99.9 % was maintained after trimming. The reads were randomly sampled to 60 million single reads per sample and then mapped to the annotated *T. officinale* genome including the small genes in CLC-GW, and saved as Total Exon Reads per gene (TER), Unique Gene Reads (UGR), and Unique Exon Reads per transcript (UER) (Supplementary Data S2a, Raw data). Data were manually transferred to Transcripts per Million (Supplementary Data S2b, TPM) and averaged per replicate, in Transcript per 10 Million (Supplementary Data S2c, TP10M).

The data quality was checked with a Principal Coordinate Analysis in CLC-GW (TERs in TPM; Supplementary Figure S5). This showed clear clustering of the replicates per stages and tissues, with particularly tight clustering of replicates in the youngest stage (F0S0) and younger seed stages (S1 and S2). In the upper floral parts (F) some more variation was detected within and between replicates, reflecting fast changes in gene expression in these rapidly developing tissues and close successive stages, e.g., the upper floral parts of mature buds (F1) and just opened flowers (F2) follow each other quickly in time. The leaf duplicates also nicely clustered together and diverged from the floral tissues.

To further analyze the data quality and visualize expression patterns, a heatmap of all expressed genes was constructed in CLC-GW (TERs in TPM; Supplementary Figure S6), confirming the reproducibility of the replicates. Also, results show the clustering of similar tissues in subsequent stages, particularly S1, S2 (F0S0) and F1, F2. The heatmap also indicated 'expression blocks' of which the most obvious genes were manually selected, boxed and numbered 1-12 (Supplementary Data S2, columns S and T). Some blocks were confined to one tissue type and stage only, for instance, blocks 1 for F0S0 and 3 for F1, and others were shared by the same tissue type of subsequent stages, e.g., blocks 2 for F0S0, S1, S2; 4 (and 5) for F1, F2 (F3); 10 for F3, F7. Blocks with the highest numbers of genes were found among the youngest stages, particularly blocks 2 (F0S0, S1, S2) and 3 (F1) and to a lesser extent 1 (F0S0) and 4 (F1, F2) (Supplementary Table S6a and graph therein), indicating high transcript activity in young floral developmental stages. A relatively high number of genes was also found in block 10 (F3, F7), indicating another stage of diverse gene activity.

To analyze gene expression in dandelion tissues, the expression of all 25 samples, both the Total Exon Reads (TER) and Unique Gene Reads (UGR), were summed for each gene and classified in seven groups from 'true zero' to 'extremely high' expression (>10,000 TPM) (Supplementary Data S2b, columns BU-BX, and summary thereof in Supplementary Table S6b and graphs therein). A total of 49,102 genes were expressed (60.4 %; sum > 1 TPM) of which a minority showed very high (7.4 %; sum > 1000 TPM) to extremely high (0.5 %; Sum > 10,000 TPM) expression. Also, the TER versus UGR were compared, showing the majority of genes (82.9 %) similarly expressed while a small part (1,5 %) showed significantly higher TERs and a larger part (15.6 %) significantly higher UGRs. The latter could be explained by the mapping of reads to introns and ≧ 400 nt untranscribed regions (UTRs) in addition to the exons. The most highly expressed genes (Supplementary Table S6c) included four genes with a summed expression of > 100,000 TPM of which three were related to anthers: pollen allergen *Art v*1-like (2x) and anther-specific *SF18*-like, expressed in F1 and F2 only, and the fourth a hypothetical protein, expressed in floral tissues, but not leaves. The next six highest expressed genes contained four that were overall highly expressed, including elongation factor *EF*1α, histone *H*3, *Acyl-CoA-binding* protein and polyubiquitin, one that was particularly expressed in the upper floral tissues (F1-F7) and seeds after pollinations (S3-S7), copper transport protein *ATX*1, and one that was expressed in all but the young (F0S0, F1) floral tissues, *Dormancy-associated* protein 1-like (additional information in Supplementary Data S2b and S2c).

To obtain insights into the expression of genes related to floral development, including the *MADS-box* Type I and II genes and *TCP* genes (and *APETALA*-2 [*AP2*]), the average expression per tissue type and stage (Supplementary Data S2c, within Columns F, G and H the relevant genes indicated, see next paragraph about the selection of those genes) was extracted and summarized (Supplementary Table S7). Several genes (13 of the 78 *MADS-box* genes and 5 of the 33 *TCP* genes; 15-17 %) were represented by two alleles in the genome assembly rather than one based on sequence similarity, partiality of some gene(s), flanking sequences and similar expression. These were taken together, and their sum of expression used in the final analysis (indicated with a double gene name and asterisk in Supplementary Table S7). Gene expression is visualized in a heatmap per gene class (LOG2 [averaged TERs per replicate + 1]; Supplementary Figures S7a-d).

The results of the *MADS-box* Type II genes showed clustering of several tissue groups: young upper floral tissues (F0S0, F1, F2), seed tissues (S1-7) and older upper floral tissues (F3, F7), and a clear differential expression in leaves (LF) (Supplementary Figure S7A). Virtually all ABC(D)E genes, indicated with an A, B, C, D or E-prefix in the gene name, were among the most highly expressed *MADS-box* genes. Some expressions were clearly defined according to expectations, for example, the class B-genes *PISTILLATA (PI)* are highly expressed in young upper floral tissues only. Most other *MADS-box* Type II genes, indicated with an M-prefix, showed (very) low expression in the floral tissues. These results confirm the important role of *MADS-box* Type II genes, particularly the ABC(D)E genes, in floral development. Based on the expression of *MADS-box* Type I genes, the seed tissues cluster together as do the upper floral tissues (Supplementary Figure S7B), supporting their importance in ovule and seed development. A few genes were specifically expressed in young buds: *AGL*47 and *AGL*62, which also confirms expectations. The heatmap of *TCP* Class I and Class II gene expression showed a similar clustering of tissues as the *MADS-box* Type II genes (Supplementary Figure S7C versus S7A), also supporting their role in floral development. Particularly the *CIN* genes were highly expressed in the floral tissues and so were some of the *PCF* genes, while most *CYC* genes show (very) low expression. Examples of tissue specificities are the high expression of a *TCP*5-*like gene* in young buds and a *TCP*8-*like* gene in tissues after pollination (F3, S3, F7, S7). Finally, the expression of the *AP2*-like homologs, an A-class non-*MADS-box* transcription factor gene, is shown (Supplementary Figure S7D) of which some showed expression in the young buds according to their role in early floral organ ontogenesis.

2.5 Synteny Network Analysis: Identification of Asteraceae-specific MADS-box and TCP gene synteny clusters

To compare the genomic context of genes within species of the Asteraceae and between Asteraceae and non-Asteraceae, we conducted a synteny network analysis of 33 angiosperm species for which high-quality Whole Genome Sequences (WGS) are available (12 Asteraceae, 18 other Asterids, 2 Rosids and 1 early diverging Angiosperm; Supplementary Table S8; Figure 1A). Within the Asteraceae, six species were from the Cichorioideae, including our *de novo* sequenced *T. officinale* (*Tof)*, two additional Taraxacum species (*Tmo* and *Tks*) (Lin et al., 2022) and three lettuce species. In addition, four species of the Asteroideae and two of the Carduoideae were analyzed. The synteny network database was built using the SynNet pipeline (Zhao and Schranz, 2017; Gamboa-Tuz et al., 2022) and contained 718,070 nodes (genes found in syntenic blocks) and 7,603,091 edges (connections between syntenic genes) (data on which subsequent analyses were based). We further focused on the sub-networks of the *MADS-box* and *TCP* gene families.

Using HMMER analysis of the 33 proteome sequences, the *MADS-box* genes were identified by searching for the *MADS-box* (SRF-TF: PF00319.20) and *K-box* (PF01486.20) domains and the *TCP* genes by searching for the *TCP*-specific *bHLH* domain (PF03634.15). We further classified the identified candidates by their sequence-similarity and phylogenetic relationship to well-known reference genes, particularly from Arabidopsis, Petunia and Gerbera (*MADS-box* genes), Arabidopsis and rice (*TCP* genes) (Supplementary Table S9). In total, 2,525 *MADS-box* and 1,019 *TCP* genes were identified (Supplementary Data S3). After classification, the normalized gene count (i.e., Z-score) for each clade was calculated. Results identified several gene expansions in different plant families (Figure 2; Supplementary Table S10), particularly, Type I and MIKC* in the Solanaceae and MIKC^c and *CYC* in the Asteraceae. Within Taraxacum, we found more *MADS-box* genes in *T. officinale* (78) than in *T. mongolicum* (54) and *T. kok-saghyz* (57), and a similar number of *TCP* genes (31-34), with the former possibly as a result of their genome completeness.

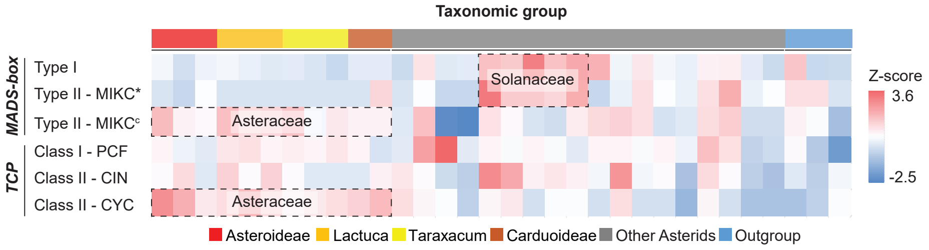


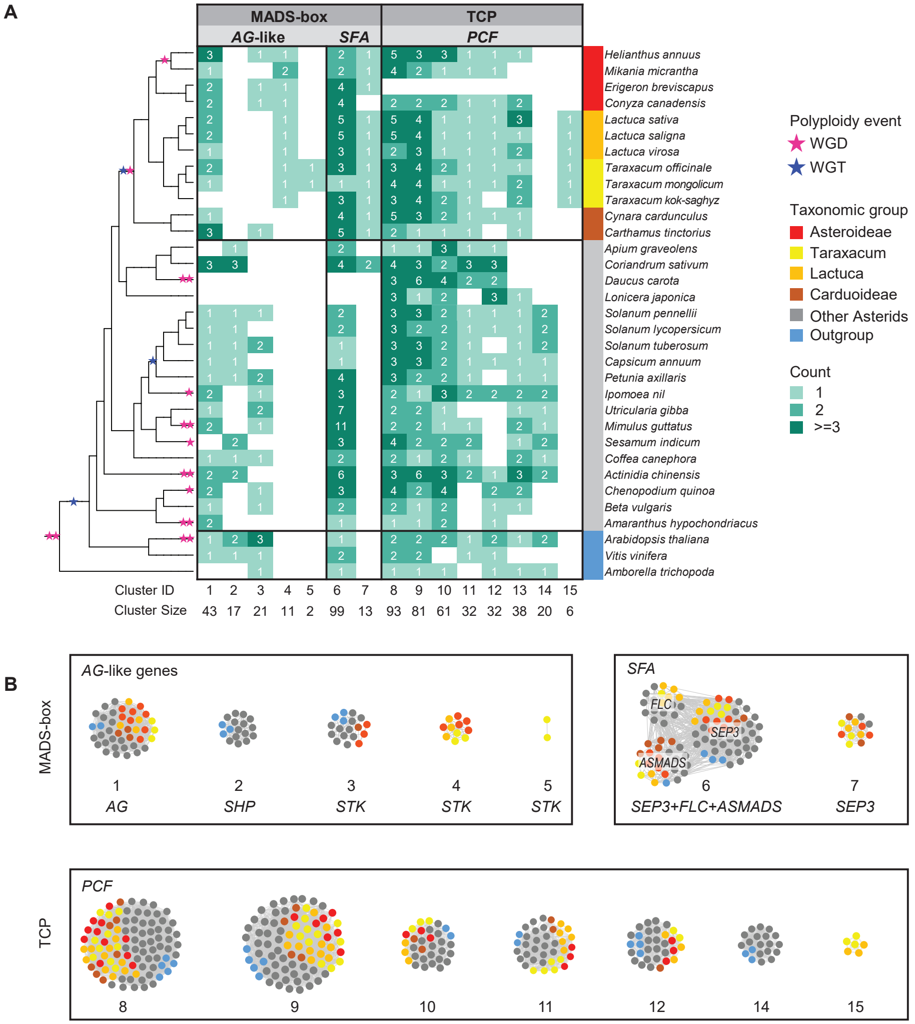
Figure 2. Z-score heat map of MADs-box and TCP genes. Target genes were identified and classified into sub-clades (row) for each species (column) from three taxonomic groups. The gene count was scaled (cell) by gene clades using z-score. Colors illustrate the deviation from average, with blue for small size and red for the big size. *Erigeron breviscapus* was excluded from this visualization due to its incompleteness of targe genes.

The complete lists of *MADS-box* and *TCP* genes were used to extract their synteny sub-networks from the whole network database. The resulting *MADS-box* sub-network contained 1,677 nodes and 16,697 syntenic edges and the *TCP* sub-network contained 835 nodes and 14,716 syntenic edges (Supplementary Data S4a, b). To associate the syntelogs (the syntenic homologous genes) with each other, we conducted phylogenetic profiling of all obtained synteny clusters of MADS-box and TCP proteins and visualized the primary clusters in a heatmap for each family (Supplementary Data S4c, d). For this, the number of syntelogs in each cluster was counted for each species and the clusters were ordered by hierarchical clustering based on the index of dissimilarity derived from the syntelog counts. Consequently the clusters specific to the Asteraceae, Cichorioideae and Taraxacum were determined. In Figure 3, we highlight 15 synteny clusters that illustrate our most relevant findings: the Asteraceae or Taraxacum specific *MADS-box* clusters *AG*-like (CL4-5) and *SEP3/FLC/AS-MADS* (SFA; CL7), and *TCP-PCF* cluster 15 (CL15), and absence of *AG*-like cluster 2 (CL2) and *TCP-PCF* cluster 14 (CL14). The selected clusters are also displayed in a network format, pruning the non-primary syntelogs (Figure 3B).

The *AG*-like genes include the C-class gene *AGAMOUS* (*AG*; CL1, Figure 3) and C/D-class genes *SHATTERPROOF*-like (*SHP*-like; CL2) and *SEEDSTICK*-like (*STK*-like; CL3-5). *AG* is critical for anther and carpel development; *SHP* regulates aspects of fruit development in core eudicot, such as fruit dehiscence in dry fruits as in Arabidopsis and fruit expansion and ripening in fleshy fruits like in tomato; and *STK* is involved in ovule development. For *AG*, most orthologous genes resided in the conserved synteny cluster 1 (CL1; Figure 3), including two genes of *T. officinale*. Syntelog(s) of *SHP* (CL2) were absent in the Asteraceae. Since Asteraceae fruits are single-seeded indehiscent dry fruits (cypsela), this is consistent with a loss or absence of a gain of *SHP* homologs. More than 60 % of the *STK* orthologs were in one single synteny cluster (CL3), mainly from non-Asteraceae species. *STK* orthologs from Asteraceae exclusively formed a second synteny cluster (CL4). Moreover, there was an extra pair of syntenic *STK* genes unique in the Taraxacum species (CL5) and one more present in *T. officinale* only (Supplementary Table S7; since unique, this is not detected as a cluster).

The *SEP3*-like genes, the E-class genes of floral development, exemplify another Asteraceae-specific relationship. *SEP* genes underlie the development of all floral organs. A conserved cluster of *SEP*3-lik*e* genes was shared by all genomes analyzed, including most Asterids, the two Rosids and the first-diverging angiosperm *A. trichopoda* (CL6; Figure 3). In addition, the majority of another *SEP3*-lik*e* cluster (CL7) was predominantly Asteraceae specific (plus *Coriandrum sativum* [Apiaceae]), and likely a transposed duplicated copy preserved in the Asteraceae.

A third example of Asteraceae-specific synteny was found in the *TCP* class II *PCF* genes, which are plant-specific transcription factors that play a role in cell differentiation and plant growth. For the *PCF* genes, we identified a Cichorioideae specific cluster (CL15; Figure 3), while a second cluster was specific for non-Asteraceae (CL14), hinting of a transposition in the ancestor of the Cichorioideae. The other six *PCF* clusters were relatively conserved in all species analyzed (CL8-13). Similarly, the *TCP* subclasses, *CIN* and *CYC* (Supplementary Dataset 4d) were conserved between the Asteraceae and non-Asteraceae.



	MADS-box							TCP							
	AG-like					*SFA*		*PCF*							
Cluster ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Helianthus annuus	3		1	1		2	1	5	3	3	1	1	1		
Mikania micrantha	1			2		2	1	4	2	1	1	1			
Erigeron breviscapus	2		1	1		4	1								
Conyza canadensis	2		1	1		4		2	2	2	1	1	2		
Lactuca sativa	2			1		5	1	5	4	1	1	1	3		1
Lactuca saligna	2			1		5	1	5	4	1	1	1	1		1
Lactuca virosa	1			1		3	1	2	3	1	1	1	2		1
Taraxacum officinale	2			1	1	3	1	3	4	2	1	1	1		1
Taraxacum mongolicum	1			1	1	1	1	4	4	1	1	1	2		1
Taraxacum kok-saghyz				1		3	1	3	4	2	1		2		1
Cynara cardunculus	1					4	1	5	3	2	1	1	1		
Carthamus tinctorius	3		1			5	1	2	1	1	1		1		
Apium graveolens		1				2		1	1	3	1	1			
Coriandrum sativum	3	3				4	2	4	3	2	3	3			
Daucus carota								3	6	4	2	2			
Lonicera japonica								3	1	2		3	1		
Solanum pennellii	1	1	1			2		3	3	2	1	1	1	2	
Solanum lycopersicum	1	1				2		3	2	2	1	1	1	2	
Solanum tuberosum	1	1	2			1		3	3	2	1		1	2	
Capsicum annuum	1	1				1		3	3	2	1	1	1	1	
Petunia axillaris	1	1	2			4		3	2	2	1		1	1	
Ipomoea nil	2		1			3		2	1	3	2	2	2	2	
Utricularia gibba	1		2			7		2	2	1			1	1	
Mimulus guttatus	2		1			11		2	2	1	1		2	1	
Sesamum indicum		2				3		4	2	2	2		1	2	
Coffea canephora	1	1	1			2		2	1	2	1		2	1	
Actinidia chinensis	2	2				6		3	6	3	2	1	3	2	
Chenopodium quinoa	2		1			3		4	2	4		2	2		
Beta vulgaris	1		1			2		2	1	2		1	1		
Amaranthus hypochondriacus	2					1		1	1	2		1			
Arabidopsis thaliana	1	2	3			1		2	2	2	1	2	1	2	
Vitis vinifera	1	1	1			2		2	2		1	1			
Amborella trichopoda			1			1		1	1	1		1	1	1	
Cluster Size	43	17	21	11	2	99	13	93	81	61	32	32	38	20	6

Figure 3. Synteny Network Clusters reveal the Asteraceae-specific context for a set of important genes during flower development. **A,** Phylogenetic profiling of selected examples for 7 *MADS-box* (AG-like, and SFA clades) clusters and 8 *TCP* (PCF clade) clusters. Gradient red cells show the number of syntelog (syntenic homolog) for each cluster in the different species. For targeted genes, phylogenetic profiling identified lineage-specific

clusters, such as Cluster 4, 7 and 15 for Asteraceae family, or Cluster 5 for *Taraxacum* genus. Species names on the right side of profiling heatmap were further divided into 6 groups based on taxonomy and indicated by different colors: *Asteroideae* sub-family (red); *Taraxacum* genus (yellow); *Lactuca* genus (orange); *Carduoideae* (brown); other species from Asterids clade (grey); three basal species as outgroup (blue). Pink and blue stars placed on the tree demonstrate the known whole-genome duplication (WGD) and whole-genome triplication (WGT) events. At the bottom, there are cluster ID and size. **B,** Visualized network of 12 selected Clusters from **A** The nodes color represents the same taxonomic group. This result is supported by Supplementary Table S7 and Supplementary Data 4. *, Cluster 6 is comprised of *SEP3*, *FLC* and *AS-MADS* types of *MADS-box* in profiling heatmap (**A**).

2.6 Phylogenomic analysis of MADS-box and TCP genes

To depict the evolutionary relationships between the different *MADS-box* and *TCP* genes, we mapped the syntenic connections (genomic context) onto the gene trees (Figure 1B and Figure 4). The phylogeny reconstruction was first based on the amino acid sequence alignments of the MADS domains (*MADS-box* genes) and *bHLH* domain (*TCP*-genes), splitting the *MADS-box* genes into Type I and Type II (including MIKC* and MIKC^c) genes and the *TCP* genes into *PCF*, *CIN* and *CYC/TB1* genes (Supplementary Figure S8). To improve the resolution of the *MADS-box MIKC^c* genes, an independent phylogenetic tree was built using 1,154 Type II *MIKC^c* genes where the K-box domain(s) was considered in the alignment beside the *MADS-box* domain (Supplementary Dataset 3a: column C). Both the phylogenies of the *MADS-box MIKC^c* genes and *TCP-PCF* genes (Figure 1B and Figure 4C) clearly classified the various gene clades. The syntenic relationships were mapped onto these trees (colored connection lines within the circles) to compare the gene evolution based on genomic context (synteny) with those of sequence divergence (gene tree). In addition, the Asteraceae sub-families were highlighted in the phylogenetic trees (colored sections of the circle). Both the *MIKC^c* and *PCF* results showed a high level of similarity between the syntenic and gene sequence relationships, with some interesting exceptions that are described in the next paragraphs.

In Figure 1A, a difference between the syntenic and genetic relationships was particularly seen for *BS* versus *PI* genes (grey lines), *AGL6* versus *SOC1/TM3* genes (dark green lines) and *SEP*-like versus *AP1/FUL* genes (yellow lines). These genes show a clear close relationship based on their genomic context (are syntenic) but occur in different clades in the phylogenetic tree based on their sequences. It suggests that these genes have been diverged, possibly as a result of selection or by a duplication followed by a loss of one of the two copies. Figure 1A also visualizes the Asteraceae specific MADS-box lineages, one within the *AG*-like clade (*STK*-like genes, dark red lines) and one within the SFA clade (*SEP3*-like genes, dark pink lines). These two syntenic clusters are shown separately in Figures 4A and 4B and described below.

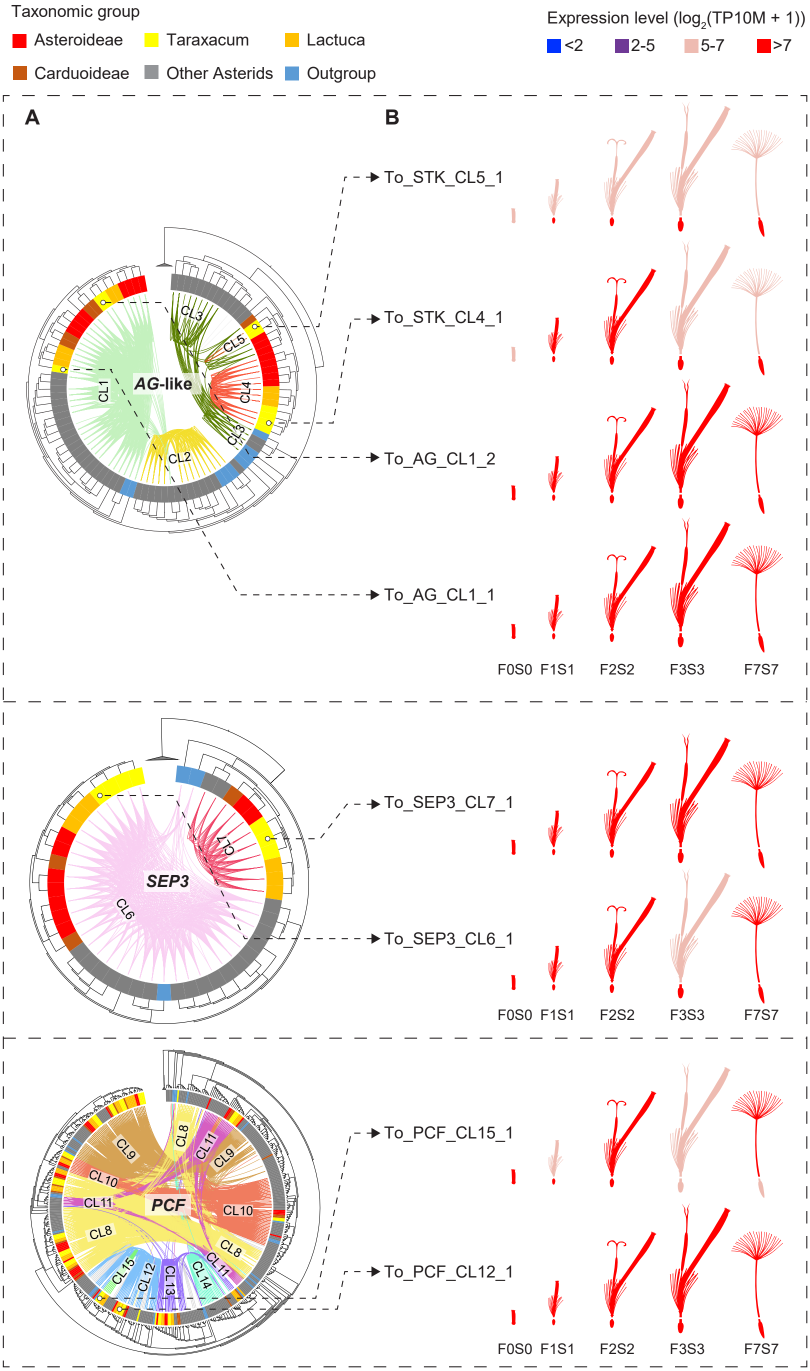


Figure 4. Flower-related genes from the same ortholog group with varying expression levels in different genomic contexts. **A.** Identified type II (MIKC^c) *MADS-box* and *TCP* genes were used to create maximum-likelihood gene tree including the syntenic relationships between the genes. Three ortholog clades of selected genes with Asteraceae-specific clusters (Figure 3) were extracted from the complete gene trees, including *AG*-like of

MADS-box (top) *SEP3* of *MADS-box* (middle), and *PCF* of *TCP* (bottom). The internal color strip indicates the taxonomic group for every gene from different species. Color lines connecting between genes indicate their syntenic relationship labelled by their cluster (CL) ID. **B.** Three gene pairs from *T. officinale* (To) represent the comparison of expression level for orthologs from **A** in different synteny clusters. The cartoon heatmap shows the transformed expression for five flower-developmental stages (0, 1, 2, 3, and 7) and two types of tissues including top flower (F) and bottom seed (S). Transformed expressions were divided into four levels: blue (<2), purple for medium (2 – 5), pink (5-7) and red for high (> 7).

The *AG*-like (*AG*, *SHP* and *STK*) gene tree, supported by syntenic connections (Figure 4A), showed the three gene clades with five synteny clusters: the *AG* clade (CL1), shared by all species and putatively ancestral, the *SHP* clade (CL2), present in non-Asteraceae species only, and the *STK* clade (CL3-5), showing evidence for gene duplications and divergence. The separation of the three *AG*-like gene clusters is in line with the previous C/D-class genes classification in angiosperm (Kramer et al., 2004). The tree validates the overall high conservation of *AG* genes and the absence of *SHP*-like genes in the Asteraceae, as mentioned above in relation to the synteny analysis (Figure 3). In addition, the *AG*-like gene tree shows that the genes in the *STK*-like clusters (CL3, dark green lines; CL4 and CL5, red lines) underwent different modes of evolution: the genes in CL3 are syntenically related, but distributed over different clades based on their sequences, while within the Asteraceae the genes in CL3 are genetically related to those in CL4, but syntenically diverged. The finding of extra *STK*-like copy in Taraxacum (CL5), and unique copy found in *T. officinale*, suggests a unique evolution of *STK*-like genes in dandelions in addition. Different expression patterns in the *AG*-like genes in Taraxacum (Figure 4A, floret cartoon), with overall high expression of CL1 genes (*AG,* C-class, dark red), and various, less high expression of the CL4 and CL5 genes (*STK*-like genes, C/D-class, light red), support their divergence.

For the *SEP3* orthologous group (Figure 4B), most gene copies resided in CL6 (light pink lines), which form one clade in the gene tree, supporting the high conservation of this expanded gene group in genomic as well as sequence context. A second group of syntelogs was found in the Asteraceae (CL7; red lines) and is also supported by the gene tree. Possibly these genes result from a transposition after a duplication in the ancestor of the Asteraceae. The expression of the *SEP*3-like genes in Taraxacum (Figure 4B, floret cartoon), shows some reduction in mature floral tissues in the conserved, putatively ancestral CL6 gene, and overall high expression in the Asteraceae-associated CL7 gene. This supports the importance of E-class genes in a wide range of floral developmental aspects and indicates that the two gene copies have not (yet) much diverged and may have maintained a similar function.

The *PCF* genes formed a large clade within the *TCP* genes and were divided into three subclades based on their gene tree (Figure 4C). Cluster 8 (yellow lines) and CL11 (pink lines) both had syntenically connected genes from all three subclades, indicating their paralogous relationships. The Asteraceae-specific CL15 genes (fluorescent green lines) are monophyletic according to the gene tree, and likely a result of a duplication within the second *PCF* clade (CL12 and CL13). Divergence in their expression pattern (Figure 4C, floret cartoon) suggests that these genes are diverging. CL10 shows an Asteraceae-specific lineage based in the third *PCF* clade that was syntenically connected to genes in the first clade. This fits a gene duplication model followed by a translocation of one of the two genes and subsequent gene loss of either the first or second copy in the different lineages. The clear synteny relationship of CL10 genes in the gene tree also illustrates an Asteraceae-associated reduction of genes in this group, which is less obvious, but present in the heatmap, illustrating another strength of the gene tree – synteny comparison.

To summarize, the extracted gene trees of *AG*-like (*AG, SHP* and *STK*: CL1-5), *SEP3* (CL6-7) and *PCF* genes (CL8-15) confirmed the orthologous relationship of genes within Asteraceae-specific synteny clusters (CL4, 5, 7 and 15) and of genes widely conserved within the angiosperms (CL1, 6, 8-13) or being non-Asteraceae specific (e.g., CL2 and 14). By combining the gene phylogeny and synteny, we validated the occurrence of duplications and/or transpositions of *AG*-like, *SEP3* and *PCF* genes in ancestral species of the Asteraceae or subsets thereof and added an extra level of evolutionary history to the traditional gene tree phylogenies.

2.7 Inference of origin and function for Asteraceae-specific MADS-box orthologs

In a recent paper about MADS-box genes in Chrysanthemum, a unique, monophyletic clade was found, including eleven Chrysanthemum genes (*CnMADS*54-64) and one from lettuce (*LsMADS*16), together named *Asteraceae-Specific MADS-box* (*AS-MADS*) genes (Won et al., 2021). To characterize this potentially novel sub-group, we included *LsMADS*16 as a reference in our MADS-box search and annotation. Most *AS-MADS* genes found belonged to the syntenic cluster 6 (CL6, Figure 3), which is one of the largest clusters in our analysis with 99 nodes. CL6 includes *SEP*3 and *FLC* in addition to *AS-MADS* genes (SFA) and its network showed nodes of *SEP3*-like genes widely, but slackly connected to the sub-clusters of *FLC*-like and *AS-MADS*-like genes, both inter- and intra-specifically (Figure 3B, FSA network). We analyzed the relationships and expression of the SFA genes in more detail (Figure 5).

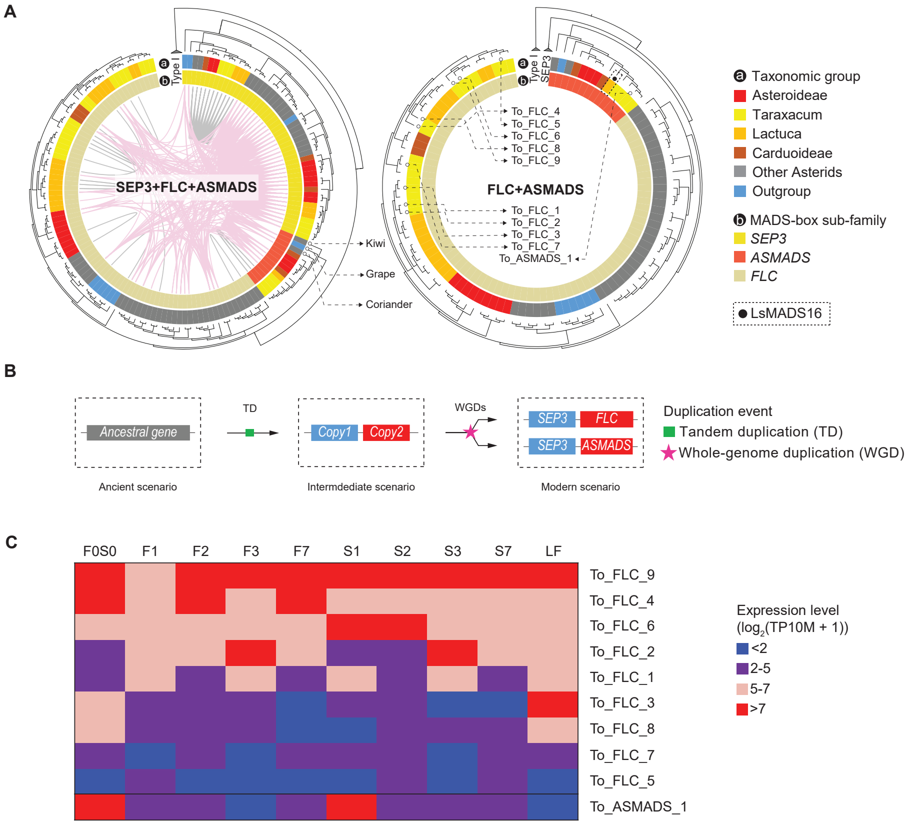


Figure 5. Inferred evolution of *Asteraceae specific MADS-box* (*AS-MADS*) gene. **A,** Left gene tree of *SFA* (*SEP3* + *FLC* + *AS-MADS*) rooted by Type I *MADS-box* reference genes, including the syntenic relationship of cluster 6 (pink) and other clusters (grey) by internal lines. The syntenic relationship of *ASMADS* suggests a duplication event of *AS-MADS* related to its syntelog *SEP3* and *FLC*. Right gene tree with the SEP3 clade collapsed for a better view of the paralogous relationship between *AS-MADS* and *FLC*. Three genes from non-Asteraceae species (Coriander, Grape, and Kiwi) indicate an ancient origin of *AS-MADS* gene for all flowering plants. The black dot squared by dash-line points the *AS-MADS* (LsMADS16) found in lettuce (Won et al., 2021). **B,** Proposed evolutionary history of *AS-MADS* showing ancestral gene (Type II MIKC^c MADS-box) firstly get tandem duplicated and then go through polyploidy events. Finally, *FLC* and *AS-MADS* diverge from the same common ancestor Copy2. **C.** The heatmap shows the different expression pattern between 9 *FLC* and one *AS-MADS* of *T. officinale* along flower development, indicating the neo-function of *AS-MADS*. The transformed expression level were divided into four levels: blue (< 2), purple for medium (2 – 5), pink (5- 7) and red for high (> 7).

The phylogenetic tree based on gene sequences versus the syntenic relationships suggested tandem gene arrangement events (Figure 5A). This was supported by the *SEP3-FLC* tandem gene arrangement found in the non-Asteraceae species *Solanum tuberosum*, *Coffee canephora* and *Beta vulgaris* (Supplementary Table S11; Supplementary Data S5). Furthermore, we found a preserved example of a *SEP3-AS-MADS* tandem in *Chenopodium Quinoa,* using the complete MADS-box list (Supplementary Table S11).

We further characterized that the *AS-MADS* is in-paralogous to *FLC* while out-paralogous to *SEP3* referring to the common ancestor prior to the duplication of the *SEP3-FLC* tandem (Figure 5A). Interestingly, out of the 13 genes in the *AS-MADS* ortholog group (indicated in orange in the inner circle), we found three that were not from Asteraceae species, namely *Coriandrum sativum* (coriander), *Actinidia chinensis* (kiwi) and *Vitis vinifera* (grape), that lacked the gene from the basal angiosperm *Amborella trichopoda*. This suggests that *AS-MADS* is from a much more ancient root before the emergence of the Asteraceae family. Combining the synteny and phylogeny results, we infer that the *AS-MADS* and *FLC* genes possibly derived from a series of whole-genome duplications (WGDs) after the tandem duplication (TD) between their common ancestor and the ancestor of *SEP3* gene (Figure 5B).

In lettuce and Chrysanthemum, the copies of *FLC* and *AS-MADS* both show various expression patterns through different floral development stages and tissues (Ning et al., 2019; Won et al., 2021). Similarly, for the representative *MADS-box* genes in *T. officinale*, we present the heatmap of expression patterns for *FLC*-like and *AS-MADS*-like genes encoding the complete MIKC^c protein in Figure 5C (i.e., contain MADS-box domain and K-box domain). In total, we found 9 *FLC* (To_FLC_1-9) and one *AS-MADS* gene (To_ASMADS_1) in *T. officinale* (Figure 5B). All of them were expressed in at least one stage and tissue (Figure 5C). Genes with rather high expression levels (> 5 with pink or red color) can be divided into two major groups: one had no specific stage and tissue, such as To_FLC_4 and To_FLC_9 expressed in all stages and both tissues; one expressed at a specific stage in specific tissue, such as To_FLC_3, To_FLC_8 and To_ASMADS_1. And all the remaining genes were in the middle of these two mentioned types. Among them, the ToASMADS_1 shows a distinctive pattern compared *FLC*s, where it specifically expressed in the seed part at the early floral stages (F0S0 and S1). To conclude, our phylogenomic and transcriptomic suggest a non-Asteraceae specific origin of *AS-MADS* genes and advocate its different function for dandelion flower development compared to *FLC*.

3. Discussion

The transcription factor gene families *MADS-box* (*MIKC^c* type) and *TCP* (*CYC* type) are critical regulators of angiosperm floral organ identity (Becker and Theißen, 2003) and flower symmetry (Luo et al., 1996), respectively. They also are critical in the evolution and control of the iconic inflorescence and florets of the Asteraceae, for example, Zhao et al. (2020) reported a *TCP-MADS-box* transcription factor network for ray floret development in *Gerbera hybrida*. Here, we provide the first complete overview, and first inventory in *Taraxacum*, of *MADS-box* and *TCP* genes in the Asteraceae, by comparing their results with those in other Asterids and a few representative model angiosperms.

Our search was robust and the identified gene numbers in the selected species were comparable to those in previous studies, for example: 82 *MADS-box* genes, including 23 Type I and 59 Type II genes, were reported in lettuce (Ning et al., 2019), while we found 78, including 20 Type I and 58 Type II *MADS-box* genes (Supplementary Table S10). A comparative analysis of the *TCP* genes in the Apiaceae family identified 29 genes in *Apium graveolens*, 43 in *Coriandrum sativum* and 50 in *Daucus carota* (Pei et al., 2021), while we found 32, 45, and 50 members, correspondingly (Supplementary Table S10). In line with this, we found 27-41 *TCP* genes in the Asteraceae (with one exception of five genes in *Erigeron breviscapus*) and 31-34 in the three Taraxacum species.

3.1 Unique patterns of gene family expansion and loss of MIKC^c MADS-Box and CYC/TB1 TCP in Asteraceae

After classification, we found that the Asteraceae, represented by 12 species, contained a lower number of copies for Type I (on average 18.5) and Type II *MIKC** (on average 2.8) *MADS-box* genes as compared to the other selected eudicots (on average 41.8 Type I and 6.1 *MIKC** genes). In contrast to the Asteraceae, the Solanaceae, represented by five species, have around four-times as many Type I (on average 74.8) and Type II MIKC* (on average 11.0) genes, which indicates a larger gene retention after its recent Whole Genome Triplication (WGT) (91 – 52 Mya; Sato et al., 2012). Similarly, the Asteraceae has undergone a *MADS-box* Type II gene family expansion and retention after a WGT (from two successive rounds of paleopolyploidy; (Barker et al., 2008, 2016), maintaining more *MIKC*^{c} (on average 45.8) than the other eudicots (on average 36.4), such as Solanaceae (on average 37.2; Supplementary Table S10). Thus, both the Solanaceae and Asteraceae show lineage-specific gene expansions and high levels of gene retention of Type II *MIKC*^{c} *MADS-box* genes following a WGT (Figure 2). These duplicated copies might evolve into new functions, for *example, in Gerbera* eight *SEP*-like *GRCD* genes were found that individually showed conserved, sub-functional, and neo-functional roles in floral organs development in contrast to the four redundant copies present in Arabidopsis (reviewed by Elomaa et al. 2018).

4

A similar scenario of gene expansion was found for the *TPC* Class II *CYC* gene family (Figure 2), where Asteraceae (except for *Erigeron breviscapus*) contains nearly twice as many *CYC* genes (10.7) than other non-Asteraceae species (on average 5.9; Supplementary Table S10). This extensive *CYC* duplication has been reported for many specific Asteraceae subfamilies (Kim et al., 2008; Chapman et al., 2008; Tahtiharju et al., 2012; Huang et al., 2016), including Senecioneae (Senecio), Mutisieae (Gerbera), Asteroideae/Heliantheae (sunflower) and Asteroideae/Anthemideae (Chrysanthemum). Our study further supports the duplication of CYC genes in the Cichoroideae (dandelion and lettuce) and Carduoideae (cardoon and safflower) subfamilies, confirming a whole-family duplication event of the *CYC* clade.

Moreover, the regulatory function of duplicated *CYC*s for Asteraceae ray flower development reported in gerbera and sunflower can be assigned to different copies in the *CYC*2 clade (Chapman et al., 2008; Tahtiharju et al., 2012). In our phylogeny (SF8), we also classified the *CYC* based on the study from Tahtiharju et al. (2012) using reference genes of *Arabidopsis*. The classified *CYC2* clade contains two genes from *T. officinale* (Toff_WURv1_g36520 and Toff_WURv1_g24074), and their expression suggest that Toff_WURv1_g36520 regulates the formation between ligulate and disc flower in dandelion, while Toff_WURv1_g24074 was likely not involved in the flower development (Supplementary Table 12). Genes of other Asteraceae species in the *CYC2* clade should also be experimentally checked to determine their function on floral identity (Supplementary Figure S8).

3.2 Lineage-specific gene-loss and genomic context of *MADS-box* and *TCP* genes

Our phylogeny shows that the *AGAMOUS*-like (*AG*-like) clade of *MADS-box* also contains *SHATTERPROOF* (*SHP*) and *SEEDSTICK* (*STK*) (Figure 4A: top, represented by cluster 2, and cluster 3-5 respectively), agreeing with previous phylogenies and classifications (Theissen et al., 2000). In the model plant *A. thaliana*, *SHP* and *STK* are both involved in the ovule/fruit development: *SHP* can activate an AG-independent carpel development and subsequently control fruit dehiscence for seed dispersal (Pinyopich et al., 2003; Liljegren et al., 2000), while *STK* regulates the development of funiculus connecting the seed to ovary wall (Pinyopich et al., 2003). Interestingly, our phylogenomic result revealed that *SHP* type (represented by synteny cluster 2) is lacking in all selected Asteraceae species and *STK*s of Asteraceae (represented by synteny cluster 4) are primarily located in a different genomic context (synteny cluster 4) than other angiosperms (cluster 3; Figure 3 and 4). In a previous phylogeny of eudicot *MADS-box* genes, the *PLENA* (*PLE*) lineage of clustered *SHP*s has no protein from Asteraceae, suggesting that the *SHP* is missing in this family (Dreni and Kater, 2014). In this study, we further determine the orthologous clade of *SHP* via synteny, thus confirming the loss of *SHP* in Asteraceae. Compared to Brassicaceae (Arabidopsis), the observed absence of *SHP* and specific *STK* copies potentially could be linked to the Asteraceae-unique seed dispersal by wind or animal through pappus (Jana and Mukherjee, 2012): on the one hand, the function of *SHP* for fruit dehiscence and single is probably not required anymore (Liljegren et al., 2000), while STK transposition might influence fruit difference between single-ovule (Asteraceae) and multi-ovules (Brassicaceae), which needs further validation.

SEP-like (Class E) genes are essential regulators that orchestrate the formation of different floral organs (Theißen et al., 2016). The *SEP*-like genes can be divided into the *SEP1/2/4* clade and the *SEP3* clade, where the *SEP3* has been shown to co-regulate the activation of class B and C *MADS-box* genes in the model plant Arabidopsis and gerbera (Kotilainen et al., 2000; Liu et al., 2009). We found that the Asteraceae (represented by

12 species of 3 subfamilies) has a gene duplication of *SEP3* lineage genes, consistent with an earlier report of an additional copy in gerbera (Mutisioideae; Zhang et al., 2017). Moreover, we further revealed one Asteraceae-dominant clade of *SEP3* duplication having a lineage-specific synteny (cluster 7) compared to a conserved synteny shared by other angiosperms (cluster 6; Figure 4 and 5). *Taraxacum officinale* has two *SEP3* genes with complete type II structures, one in each cluster (To_SEP3_CL6_1 and To_SEP3_CL7_1 in Figure 4). Considering the diverse functions of *SEP* genes in gerbera (reviewed by Elomaa et al., 2018), it will also be interesting to examine patterns of neo- and subfunctionalization of different *SEP3* copies in *T. officinale*, as well as the potential effect caused by a positional change (i.e., different syntenies).

Although *CYC* genes are the typical *TCP* of special interest due to their control of Asteraceae floret symmetry we also included the Class I *PCF* genes in our study. We found a syntenic depth of three for *PCF* genes, which likely are derived and retained from the ancient γ WGT shared by eudicots (Figure 4A: bottom). Compared to the function of *CYC* in flower development, PCF genes (PCF1 and PCF2) were first defined in rice and found to regulate the expression of meristematic tissue primarily via heterodimers (Kosugi and Ohashi, 1997). In this study, we find an Asteraceae-specific synteny of the Class I *PCF* genes (Figure 3 and 4: cluster 15). It is essential first to check its expression during flower development using our RNA-seq data of *T. officinale* (see below ‘expression pattern’). A further test of its cis-regulatory elements can be done to examine whether it has a regulatory novelty under the Asteraceae-specific genomic context.

3.3 Origin and revised classification of Asteraceae specific-MADS (*AS-MADS*) gene

An Asteraceae specific-MADS type (*AS-MADS*) was recently described in chrysanthemum (Won et al., 2021). They identified a monophyletic clade comprising multiple *AS-MADS* copies from chrysanthemum and one copy from lettuce. The single lettuce copy (*LsMADS16*) was earlier found to be in the *FLC*-like clade (Ning et al., 2019). In this study, we identified a monophyletic clade of *AS-MADS* anchored by *LsMADS16* (Figure 5A). Surprisingly, the *AS-MADS* clade also contains proteins from coriander (Apiaceae as Asteraceae outgroup), kiwi (basal Asterid), and grape (basal rosid) in addition to Asteraceae. Our phylogenomic analysis also demonstrates that *AS-MADS* and *FLC* share the same last common ancestor, and both are syntenic to *SEP3* (Figure 5A). The previous study indicates that *FLC* is derived from a *TM8* homolog (Gramzow et al., 2014), which shares the same ancestor with *SEP* before a tandem duplication event in seed plants (Ruelens et al., 2013). The *SEP3-TM8* tandem is believed to be more ancestral than the *SEP3-FLC* tandem (Zhao et al., 2017). Based on phylogenomic analysis, we propose that the *AS-MADS* is also derived from the *TM8* homolog, like *FLC* (Figure 5B). Moreover, the synteny of *AS-MADS* is maintained in one single cluster comprised of Asteraceae species and other eudicots, which indicates that this synteny has been retained for *AS-MADS* at least since the last common ancestor of eudicots. Furthermore, *FLC* was found in a more ancestral species, *Amborella trichopoda* (basal angiosperm), hence its paralog, *AS-MADS*, could also diverge from the ancient ζ WGD (i.e., shared by angiosperms) after the *SEP3-TM8* tandem in the ancient flower plant ancestor. In summary, our results anticipated that the *AS-MADS* is a paralog of *FLC* and has a more ancient origin but prevalent reservation in Asteraceae species compared to other eudicots.

3.4 Expression pattern of lineage-specific suggests specialized function/novel regulation during flower development

By combining phylogeny and synteny information, we have validated and expanded gene family classifications and the identification of orthologous relationships between genes in conserved and lineage-specific genomic contexts (Figure 4A; Figure 5A). Based on phylogenomic analyses, we further examined the expression of all mentioned genes in Asteraceae, including *STK*-like, *SEP3*-like, *PCF*-like and *AS-MADS* using our *T. officinale* genome as reference. We found a diverse pattern of expression (Figure 4B; Figure 5C). The *STK* in *T. officinale* specific synteny (To_STK_CL5_1) has a partial expression pattern of another copy in Asteraceae-specific synteny (To_STK_CL4_1), implying a potential subfunctionalization event. For *SEP3*, the expression pattern of *T. officinale* copy in the Asteraceae-specific context is highly similar to the second copy in the conserved synteny with other angiosperms, which likely maintains the conserved function. Unlike *SEP3*, the two closely related *T. officinale PCF*s expressed differently during flower development, which indicates a potential regulatory novelty after gene transposition. In *Phalaenopsis* species, the PCF genes were found to co-express with other transcription factors like *MADS-box* (e.g., AP3, PI, and SEP3) and MYB (e.g., TCP) in bud, callus and gynostemium (Pramanik et al., 2020). A similar balancing role might be true to either one of the *PCF* copies. For *AS-MADS*, the one *T. officinale* copy has a different expression pattern than the other 9 *FLC* copies, which is highly expressed in the seed part (i.e., ovary) at early stage F0S0 and S1. This result functionally suggests the specialization of *AS-MADS* genes as a separate subclade of the *MIKC*^{c} type.

We sequenced the genome and transcriptome of the common dandelion. While the genome assembly is still fragmented and not on a par with more recently published genomes, it has good completeness both of the assembly and the annotation. The work presented in this study shows the usefulness of the *de novo* genome for understanding Asterid evolutionary history. Combining this *de novo* genome with genomic data of other Asterids, we systematically studied the genes highly related to Asteraceae floral development and of *MADS-box* and *TCP* in particular. Future high-quality genome assemblies of other Asteraceae species and subfamilies can facilitate and validate our conclusions about *MADS-box* and *TCP* contribution to Asteraceae floral evolution. We also validated gene expression in lineage-specific synteny or phylogeny (*AS-MADS*) using referenced-based mapping on our *T. officinale* genome. In addition to floret, *T. officinale* material from the inflorescence meristem stage could also be sequenced to explore the function of *MADS-box* and *TCP* highlighted in this paper.

4. Materials and Methods

4.1 Plant material

The common dandelion accession sequenced is a member of a diploid dandelion population in France near the village of Châtillon, Jura (FCh72; population F3 in Verhoeven and Biere 2013). It was grown from a field-collected seed and maintained in the greenhouse via cuttings, under 16/8 h light/dark conditions, frost free and a maximum temperature of 20 °C. FCh72 is a sexual plant with *2n = 2x = 16* chromosomes and an estimated genome size of 831 Mb (W1; Dolezel et al., 2005).

4.2 DNA preparation

One of the cuttings of plant FCh72 was placed in the dark (etiolated) for three days, after which young leaves were harvested, the largest veins removed and the remainder frozen in liquid N_2 and stored at -80°C. DNA extraction was performed according to the Cetyl-Trimethyl-Ammoniumbromide (CTAB) method by Chang et al., (1993) with minor modification, while care was taken in all steps to keep the High Molecular Weight (HMW) DNA. In brief: a total of 2-3 gr leaf material was grounded in liquid N_2, the DNA was extracted in pre-warmed CTAB buffer at 65°C for 1 hr, the DNA was purified via two subsequent Chloroform extractions and then precipitated using 0.7 volumes Isopropanol (4°C overnight). Pellets were resuspended in 450 µl RNase- and DNase-free MilliQ water (MQ) and the RNA removed by an RNase treatment with 50 Units RNaseOne™ Ribonuclease (Promega, Madison USA). An equal volume of Sodium Chloride-Tris-EDTA (SSTE) 2x buffer was added, a third Chloroform extraction was performed and the DNA precipitation in Ethanol. DNA pellets were dissolved in MQ and the concentration and quality examined on a NanoDrop 2000 (Thermo scientific) and Qubit 2.0 (Invitrogen, Life Technologies, Carlsbad CA), the latter using the dsDNA HS assay (Invitrogen, Life Technologies). A total of 40 µg HMW DNA was prepared for PacBio and Illumina library preparations for sequencing.

4.3 RNA preparation

To facilitate gene annotation, a mix of RNA from *T. officinale* flower, bud, leaf and root tissues was prepared. Tissues were collected from cuttings of the above mentioned plant FCh72 over different days, depending on tissue availability. The largest veins were removed from the leaves and the roots were quickly rinsed with MQ, after which the tissues were frozen in liquid N_2 and stored at-80˚C. Total RNAs were extracted from the each of the tissue types separately following the TRIzol reagent (Invitrogen) method with the adjustments by Ferreira de Carvalho et al., (2016b). RNAs were treated with DNAse (Turbo DNA free kit; Ambion) according to the manufacturers protocol. The RNA integrity and concentration were checked on a Nanodrop 2000 and by examining the 25S:18S quality and ratio on a 1% agarose gel. Samples were then pooled to equimolar concentrations and a total of 1.5 µg RNA prepared for Illumina library preparation.

For floral expression analysis, RNAs from *T. officinale* buds and flower heads at different developmental stages were harvested from cuttings of plant FCh72, with the younger stages (stages 0, 1 and 2) in triplicate and the older stages (stages 3 and 7) as duplicates. Harvesting was performed over different days, depending on tissue availability. The samples included very young, whole buds (F0S0; organs initiating, stem ~0 cm), and buds and flower heads of older stages separated through the beaks in an upper (F) and lower (S) floral part: mature buds (F1, S1; organs determined and elongated, stem ~10 cm); open flowers (F2, S2), old flowers (F3, S3; 3 days after pollination [DAP]); mature pappus (F7) and ripening seeds (S7; 7 DAP), and leafs (LF) (see for eact stage definitions and sample preparation Vijverberg et al., 2021). A total of 10-40 mg tissue was collected for each sample, quickly prepared at room temperature and then frozen in N_2. Total RNAs were isolated using Trizol reagent as described above and dissolved in DEPC-MQ to a final concentration of 200 ng / µl.

4.4 DNA and RNA sequencing

The *Taraxacum* genome was sequenced in three rounds, using a PacBio RSII sequencing system (W2), 10X Genomics with Illumina HiSeq2500 125 paired end sequencing (W3), and BioNano Genomics technology (W4), respectively. All library preparations and sequencing were performed by the sequence facility of Wageningen University & Research (W5). PacBio uses Single Molecule Real-Time (SMRT) sequencing technology, providing long reads averaging 10-15 kb. The 10X Genomics method is droplet-based, enabling barcode-specific sequencing of small amounts of DNAs / single DNA strands, facilitating the haplotype detection and sequence assembly. The illumina reads were also used to polish the sequences. Optical mapping by BioNano further improved the contig assembly.

The RNA library preparations and sequencing for gene annotation was performed at the same sequence facility at Wageningen University & Research, using Illumina HiSeq2500 125 nt paired end sequencing. The RNA library preparation and sequencing of samples of the floral expression analysis was performed at BaseClear BV (Leiden, The Netherlands), using NovaSeq 150nt paired end sequencing.

4.5 Genome assembly

We obtained PacBio reads with the mean subread length of 12,259 bp and a total length of 62,496,657,252 bp, corresponding to ~75x coverage of the *Taraxacum* genome. In addition, we obtained ~161x of Illumina 10X 150 nt paired end reads. The PacBio subreads were assembled using Canu (version 1.3, corMaxEvidenceErate 0.15) (W6). The resulting contig assembly was checked for contaminants using blobtools (v1.0) (Laetsch and Blaxter, 2017) and assessed for completeness with BUSCO (v5.2.2 using eudicot_odb10) (Manni et al., 2021). Assembly statistics were gathered using Quast (v5.02) (Gurevich et al., 2013). To collapse separately assembled haplocontigs the purge_dups manual protocol (W7) was followed. In brief, any contigs with assembly ambiguities were split using tigmint (v1.0.0) (Jackman et al., 2018), reads mapped back to the split assembly using minimap2 (Li, 2018), and putative haplocontigs collapsed by purge_dups using coverage information. Internal joins in scaffolds by purge_dups were then split on all 22N recognition sequences. This assembly was then polished with two rounds of RACON (v1.4.11) (Vaser et al., 2017) using the original PacBio data. Next, the polished assembly was scaffolded with the Illumina 10X data using ARCS (v1.1.0) (Yeo et al., 2018). The assembly was further scaffolded with BioNano Irys data using hybrid-scaffolding. In a final step, the assembly was polished with the 10X Illumina data using Pilon (v1.22) (Walker et al., 2014).

4.6 Repeat Masking

Repetitive sequences and transposable elements (TE) in the *T. offcinale* genome were identified using a combination of de novo and homology-based approaches at the DNA level. De novo: RepeatModeler (v.2.0.1 with the LTRstruct option) was used to create a de novo repeat dataset (Flynn et al., 2020). The results from RepeatModeler were combined with the RepeatMasker combined data subset relevant for *T.officinale* (i.e., *viridiplantae,*) and used as input for RepeatMasker (open-4.0) The results from RepeatMasker were used to softmask the genome assembly prior to annotation (Smit et al., 2019).

4.7 Gene Prediction and Functional Annotation

We employed the BRAKER2 (Brůna et al., 2021) pipeline for *ab initio* gene prediction. First, stranded RNAseq data from four tissues were quality and adapter trimmed using cutadapt (v1.11) (Martin, 2011). The trimmed reads were aligned against the assembly (sans mitochondrial scaffold) using STAR (v2.6.1c) (Dobin et al., 2013). The aligned reads were separated into forward and reverse read for BRAKER2 stranded mode. The reads were used as input for BRAKER2, together with the softmasked reference. The BRAKER2 RNA evidence-based pipeline uses GeneMark-ET (Lomsadze et al., 2014) to generate initial gene structures using transcript support from RNA-Seq alignment. Next, AUGUSTUS (Stanke et al., 2008) uses the filtered predicted genes for parameters training and then integrates RNA-Seq information as extrinsic evidence into final gene predictions. For functional annotation and filtering, the transcript sequences predicted by BRAKER2 were extracted using gffread (Pertea and Pertea, 2020) and converted to protein sequences using EMBOSS transeq (v6.6.0,) (Rice et al., 2000). To identify homologous sequences, we used DIAMOND (Buchfink et al., 2021) blastp (v2.0.7, "-b 10 –c1 –outfmt 5--sensitive") against nr (downloaded 06-03-2021). In addition, we analysed the transcripts with InterProScan (v5.50-84.0) (Jones et al., 2014). Protein sequences, blast output and InterProScan results were then imported into Blast2Go (Conesa et al., 2005) Basic (v5.2.5) and annotated with gene names and GO terms following the standard annotation pipeline. The resulting annotation was exported as gff3 file and subsequently formatted, filtered and annotated using custom scripts. Mainly, transcripts shorter than 150aa without homologous evidence were removed, duplicated transcripts marked in the Note field with "Sequence identical to:" and transcripts with more than 99% aa identity were labelled with " Protein > 99 perc identical to: " followed be the matching gene identifiers. Genes were relabeled in order of appearance on the assembly.

4.8 Genome comparison

We aligned our assembly with those of Tmon and Tkok using minimap2 (v2.24-r1122: -x asm5 -K 4g --cap-kalloc=2000m -t 16) and visualized the outcome in a dot-plot using dotplotly (-s -t -m 5000 -q 50000 -k 40 -x). We ran BUSCO (v5.2.2 with using eudicots_odb10), on three transcriptomes to compare genome quality (also see below section).

4.9 Genome database

Plant whole genome sequences of 33 species were selected for synteny network and phylogenomic analysis, including species of the two large, derived crown groups of the Asteraceae: Cichorioideae (covering *Taraxacum)* and Asteroideae, two species of a basal subfamily: Carduoideae, four none-Asteraceae Asterid II members, 11 species from the Asterid I clade, four early diverging Asterids, two species of the Rosids and the basal *Amborella trichopoda.* Among them, 14 species were retrieved from Zhao et al. (2019; indicated with * in Supplementary Table S8) and more recent researches (19 species mainly Asteraceae). The protein sequences of primary transcripts and corresponding gene position were extracted from selected genomes for downstream phylogenic and syntenic analysis. BUSCO (v5.2.2) was used to assess the completeness of proteomes using eudicots_odb10 dataset.

4.10 Identification and classification of MAD-box and TCP genes

For *MADS-box* genes, HMMER (v3.3.2) was used to search for the MADS- (PF00319.20) and K-box (PF01486.20) domains in all amino acid (aa) sequence from 33 species, with a default cut-off using the profiles of Hidden Markov Models (HMMs) collected from pfam (Mistry et al., 2013). To classify the identified *MADS-box* candidates, a reference database of 162 *MADS-box* genes was prepared, including 107 *Arabidopsis thaliana*, 32 *Petunia hybrida*, 21 *Gerbera hybrida*, one *Solanum lycopersicum* (tomato) gene (TM8), and one Asteraceae specific *MADS-box* gene (*AS-MADS*) from *L. sativa* (lettuce) (Supplementary Table S9). To quickly classify the sub-families of the identified *MADS-box* genes, BLAST (v2.12.0) was applied to search for the best match of each candidate using aa sequence encoded by the reference genes as database with default cut-off.

For *TCP* genes, 53 classified genes were collected, including 24 from *A. thaliana*, 26 from *Oryza sativa* (rice), two from *Antirrhinum majus* (Garden snapdragon), and one from *Zea mays* (maize). The source of the sequence data from this section can be found in the Supplementary Table S9. HMMER was used to search for the TCP domain (PF03634.15) in 33 proteomes with default setting. To further classify the *TCP* homologs, BLAST (v2.12.0) was applied to search the best match for each candidate using the reference genes as database.

4.11 Synteny Network analysis

Complete synteny networks of proteomes for the 33 plant species was created by the SynNet-Pipleine from Zhao and Schranz (Zhao and Schranz, 2017), which is available at https://github.com/zhaotao1987/SynNet-Pipeline. In this pipeline, Diamond (v2) was applied to conduct the whole-genome protein comparison (Buchfink et al., 2015). Then MCScanX was used to detect the syntenic blocks (minimum homologs = 6 genes, max gaps =25 genes) and the output was merged into the synteny network database (Wang et al., 2012). The syntenic connections of identified *MADS-box* and TCP genes were extracted from the synteny network separately excluding the duplicated alleles in *T. officinale* genome (Supplementary Table 10). Then extracted sub-networks were further clustered (i.e., cut into small networks) by the Infomap algorithm in R (Rosvall and Bergstrom, 2008). Clustered synteny networks were visualized in CYTOSCAPE (v3.7.1) (Shannon et al., 2003). Next, phylogenomic profiles were built by quantifying syntenic genes per syntelog (syntenic homolog) cluster in all 33 species. Subsequently, hierarchical clustering (ward.D) was done to re-order the synteny clusters using jaccard index. To study the genomic context of interesting genes, clusters were annotated by their primary syntelog(s) (> 10 % composition). Clusters were determined as Asteraceae-specific if more than 80 % of the syntelogs from Asteraceae species. The 80% cutoff instead of 100% was selected to maintain the evidence of close species that shared same WGD or WTD events with Asteraceae.

4.12 Phylogeny construction of identified genes

For both, the *MADS-box* and TCP *genes*, the protein sequences of all identified homologs were aligned based on their domains' HMM (PF00319.20 and PF03634.15) using HmmerAlign (Kristensen et al., 2011). Next, PAL2NAL (v14) was used to convert the protein alignments back to codon alignment, and the codon alignments of *MADS-box* and *TCP* were trimmed by TrimAl (v1.4.1) using -automated1 and -gappyout mode respectively (Suyama et al., 2006; Capella-Gutiérrez et al., 2009). RAxML phylognetic trees were constructed for both gene families by IQ-TREE (v1.6.2) with 1000 ultrafast bootstrap (UFBoot) replicates to assess final tree topology (Nguyen et al., 2015). For *MADS-box* genes, the best-fit model GTR+F+ASC+R10 was used by IQ-TREE (-pers 0.1, -nm 500) for 10 independent runs. For *TCP* genes, the best-fit model GTR+F+R6 was used by IQ-TREE to infer the phylogeny (default for other options). The consensus tree was annotated and visualized by iTOL (v6) (Letunic and Bork, 2021).

To better identify and classify the Type II type of *MADS-box*, genes, classified as Type II by previous phylogeny and with a K-box domain identified by HMMER and curated by SMART, were selected for a second-round construction of phylogeny, together with collected reference genes. The complete amino acid sequences were first aligned by MAFFT (v7.490) by FFT-NS-2 strategy. Then, the protein alignment was converted back to codon alignment using PAL2NAL. Further, the residues shared by less than 5% (-gt 0.05) in alignment were trimmed by TrimAl (v1.4.1). In addition, the trimmed alignment was manually curated in Mesquite (v3.61). Finally, IQ-TREE (v1.6.12) was used to infer the the maximum-likehood trees using GTR+F+ASC+R10 model with 1000 ultra-fast bootstrap (UFBoot) and SH-aLRT test replicates.

4.13 Expression Analysis

Analysis of gene expression and visualization thereoff was performed by using CLC-GW, Excel and R. Maximum distance between paired reads was set to 2000 nt, raw sequence reads trimmed on quality (0.05), ambiguity (2 nt), adapters, and length (>30 nt), and both paired and broken pairs saved for mapping. Samples with high read numbers were sampled back to 60 M single reads by using 'Random sampling' tool. Read mapping was performed to the annotated *Taraxacum* genome, including all genes of length 150 nt and longer (n = 81.291), using the 'RNAseq analysis' tool and following settings: Mismatch cost = 2; Insertion cost = 3; Deletion cost = 3; Length fraction = 0.5; and Similarity fraction = 0.9. Two expression values were collected: Total Exon Counts (TEC) and Unique Gene Counts (UGC). For the latter, all genes were extended with an extra 400 nt up- and downstream of the genes to ensure including the reads that map partly or entirely in the 5'- and 3'-UTRs in the counts. For the final analysis and heatmaps, Total Exon Reads (TER) were used, while genes with significant different Unique Gene Reads (UGR) were labeled, allowing considering for genes of interest. Total Exon Reads were normalized to Transcripts Per Million (TPM) in CLC-GW, and the data checked for quality with a Principal Component Analysis (PCA) and a Heatmap, using the 'PCA for RNAseq' and 'Create heatmap for RNAseq' tools, respectively, with the latter based on Euclidean distances and Complete cluster linkage. In the heatmap, stage(s) and tissue(s) related 'expression blocks' were defined manually, and the corresponding 'block' numbers added to the genes involved. The data was exported to Excel, in which the read values were averaged over the tissue types and stages. For this, first raw values were transformed to reads per ten million (RP10M) and averaged (AvTEC), then the averaged values were transformed to TP10M. Subsequently, the Minimum (MIN), Maximum (MAX), Range (MAX-MIN) and Ratio (MAX/MIN) calculated over the nine floral-related tissue types, excluding the leaf sample, for each gene. In the cases in which MIN = 0, the Ratio was based on the lowest non-0 value and the number of samples without expression indicated. An interactive heatmap was calculated based on the averaged values after excluding all non-expressed genes (MAX < 5) and non-differential expressed genes (Ratio < 3), after transformation to Log2(value +1), using heatmaply in R. More focused heatmaps were calculated for the subsets of genes interest, the *MADS-box* gens and *TCP* genes.

5. Acknowledgments

We thank Elio Schijlen, Linda Bakker, Jan van Haarst (Bioscience, Wageningen University & Research [WUR]), Rens Holmer (Bioinformatics, WUR), and Frank Becker (Genetics, WUR) for bioinformatic and technical assistance. We thank Cloé Villard (Biosystematics, WUR), Wenhao Li (Biometris, WUR) and Xinyu Hu for the advice on graph design. This work was supported by grants from the Dutch Research Council Applied and Engineering Sciences grant# 13700 (ParTool) to T.Z., C.O., M.B., M.E.S., and K.Vi., and Chinese Scholarship Council (CSC) to W.X..

6. References

Alvarez-Buylla, E.R., Pelaz, S., Liljegren, S.J., Gold, S.E., Burgeff, C., Ditta, G.S., De Pouplana, L.R., Martinez-Castilla, L., and Yanofsky, M.F. (2000). An ancestral *MADS-box* gene duplication occurred before the divergence of plants and animals. Proc. Natl. Acad. Sci. U. S. A. 97: 5328–5333.

Anderberg, A.A. et al. (2007). Compositae. In Flowering Plants · Eudicots, J.W. Kadereit and C. Jeffrey, eds (Springer Berlin Heidelberg: Berlin, Heidelberg), pp. 61–588.

Barker, M.S., Kane, N.C., Matvienko, M., Kozik, A., Michelmore, R.W., Knapp, S.J., and Rieseberg, L.H. (2008). Multiple paleopolyploidizations during the evolution of the compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol. Biol. Evol. 25: 2445–2455.

Barker, M.S., Li, Z., Kidder, T.I., Reardon, C.R., Lai, Z., Oliveira, L.O., Scascitelli, M., and Rieseberg, L.H. (2016). Most compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the calyceraceae. Am. J. Bot. 103: 1203–1211.

Becker, A. and Theißen, G. (2003). The major clades of *MADS-box* genes and their role in the development and evolution of flowering plants. Mol. Phylogenet. Evol. 29: 464–489.

De Bodt, S., Maere, S., and Van De Peer, Y. (2005). Genome duplication and the origin of angiosperms. Trends Ecol. Evol. 20: 591–597.

Brock, M.T., Weinig, C., and Galen, C. (2005). A comparison of phenotypic plasticity in the native dandelion *Taraxacum ceratophorum* and its invasive congener *T. officinale*. New Phytol. 166: 173–183.

Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M., and Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics Bioinforma. 3: 1–11.

Buchfink, B., Reuter, K., and Drost, H.G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods 18: 366–368.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12: 59–60.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.

Carlquist, S. (1976). Tribal Interrelationships and Phylogeny of the Asteraceae.

Chang, S., Puryear, J., and Cairney, J. (1993). A simple and efficient method for isolating RNA from pine trees. Plant Mol. Biol. Report. 11: 113–116.

Chapman, M.A., Leebens-Mack, J.H., and Burke, J.M. (2008). Positive selection and expression divergence following gene duplication in the sunflower *CYCLOIDEA* gene family. Mol. Biol. Evol. 25: 1260–1273.

Chase, M.W. et al. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot. J. Linn. Soc. 181: 1–20.

Chen, J., Shen, C.-Z., Guo, Y.-P., and Rao, G.-Y. (2018). Patterning the Asteraceae capitulum: duplications and differential expression of the flower symmetry *CYC2*-Like Genes. Front. Plant Sci. 9: 551.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.

Dewey, C.N. (2011). Positional orthology: putting genomic evolutionary relationships into context. Brief. Bioinform. 12: 401–412.

Dezar, C.A. (2003). Identification of three *MADS-box* genes expressed in sunflower capitulum. J. Exp. Bot. 54: 1637–1639.

Van Dijk, P.J., Op den Camp, R., and Schauer, S.E. (2020). Genetic dissection of apomixis in dandelions identifies a dominant parthenogenesis locus and highlights the complexity of autonomous endosperm formation. Genes (Basel). 11: 961.

Van Dijk, P.J., Tas, I.C.Q., Falque, M., and Bakx-Schotman, T. (1999). Crosses between sexual and apomictic dandelions (*Taraxacum*). II. The breakdown of apomixis. Heredity (Edinb). 83: 715–721.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Doebley, J., Stec, A., and Hubbard, L. (1997). The evolution of apical dominance in maize. Nature 386: 485–488.

Doležel, J. and Bartoš, J. (2005). Plant DNA flow cytometry and estimation of nuclear genome size. In Annals of Botany (Oxford Academic), pp. 99–110.

Dreni, L. and Kater, M.M. (2014). *MADS* reloaded: evolution of the *AGAMOUS* subfamily genes. New Phytol. 201: 717–732.

Elomaa, P., Zhao, Y., and Zhang, T. (2018). Flower heads in Asteraceae—recruitment of conserved developmental regulators to control the flower-like inflorescence architecture. Hortic. Res. 5: 36.

Falque, M., Keurentjes, J., Bakx-Schotman, J.M.T., and Van Dijk, P.J. (1998). Development and characterization of microsatellite markers in the sexual-apomictic complex *Taraxacum officinale* (dandelion). Theor. Appl. Genet. 97: 283–292.

Ferreira de Carvalho, J., de Jager, V., van Gurp, T.P., Wagemaker, N.C.A.M., and Verhoeven, K.J.F. (2016a). Recent and dynamic transposable elements contribute to genomic divergence under asexuality. BMC Genomics 17: 884.

Ferreira de Carvalho, J., Oplaat, C., Pappas, N., Derks, M., de Ridder, D., and Verhoeven, K.J.F. (2016b). Heritable gene expression differences between apomictic clone members in *Taraxacum officinale*: Insights into early stages of evolutionary divergence in asexual plants. BMC Genomics 17: 203.

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

Folk, R.A., Siniscalchi, C.M., and Soltis, D.E. (2020). Angiosperms at the edge: extremity, diversity, and phylogeny. Plant. Cell Environ. 43: 2871–2893.

Gamboa-Tuz, S.D., Pereira-Santana, A., Zhao, T., and Schranz, M.E. (2022). Applying synteny networks (SynNet) to study genomic arrangements of protein-coding genes in plants. In Methods in Molecular Biology (Humana Press Inc.), pp. 199–215.

Gramzow, L., Weilandt, L., and Theißen, G. (2014). *MADS* goes genomic in conifers: Towards determining the ancestral set of *MADS-box* genes in seed plants. Ann. Bot. 114: 1407–1429.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072–1075.

Henschel, K., Kofuji, R., Hasebe, M., Saedler, H., Münster, T., and Theißen, G. (2002). Two ancient classes of MIKC-type *MADS-box* genes are present in the moss *Physcomitrella patens*. Mol. Biol. Evol. 19: 801–814.

Huang, C.H., Zhang, C., Liu, M., Hu, Y., Gao, T., Qi, J., and Ma, H. (2016). Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. Mol. Biol. Evol. 33: 2820–2835.

Ilic, K., SanMiguel, P.J., and Bennetzen, J.L. (2003). A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. Proc. Natl. Acad. Sci. U. S. A. 100: 12265–12270.

Jackman, S.D., Coombe, L., Chu, J., Warren, R.L., Vandervalk, B.P., Yeo, S., Xue, Z., Mohamadi, H., Bohlmann, J., Jones, S.J.M., and Birol, I. (2018). Tigmint: correcting assembly errors using linked reads from large molecules. BMC Bioinformatics 19: 393.

Jana, B.K. and Mukherjee, S.K. (2012). Pappus Structure in the Family Compositae- A Short communication. Int. J. Sci. Res. 3: 29–30.

Jones, P. et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics 30: 1236–1240.

Kim, M., Cui, M.L., Cubas, P., Gillies, A., Lee, K., Chapman, M.A., Abbott, R.J., and Coen, E. (2008). Regulatory genes control a key morphological and ecological trait transferred between species. Science 322: 1116–1119.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27: 722–736.

Kosugi, S. and Ohashi, Y. (1997). PCF1 and PCF2 specifically bind to cis elements in the rice proliferating cell nuclear antigen gene. Plant Cell 9: 1607–1619.

Kotilainen, M., Elomaa, P., Uimari, A., Albert, V.A., Yu, D., and Teeri, T.H. (2000). *GRCD1*, an *AGL2*-like *MADS-box* gene, participates in the C function during stamen development in *Gerbera hybrida*. Plant Cell 12: 1893–1902.

Kramer, E.M., Jaramillo, M.A., and Di Stilio, V.S. (2004). Patterns of gene duplication and functional evolution during the diversification of the *AGAMOUS* Subfamily of *MADS-box* genes in Angiosperms. Genetics 166: 1011–1023.

Kristensen, D.M., Wolf, Y.I., Mushegian, A.R., and Koonin, E. V. (2011). Computational methods for Gene Orthology inference. Brief. Bioinform. 12: 379–391.

Laetsch, D.R. and Blaxter, M.L. (2017). BlobTools: Interrogation of genome assemblies. F1000Research 6: 1287.

Langham, R.J., Walsh, J., Dunn, M., Ko, C., Goff, S.A., and Freeling, M. (2004). Genomic duplication, fractionation and the origin of regulatory novelty. Genetics 166: 935–945.

Letunic, I. and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 49: W293–W296.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100.

Li, S. (2015). The *Arabidopsis thaliana* TCP transcription factors: a broadening horizon beyond development. Plant Signal. Behav. 10.

Liljegren, S.J., Ditta, G.S., Eshed, Y., Savidge, B., Bowmant, J.L., and Yanofsky, M.F. (2000). *SHATTERPROOF MADS-box* genes control dispersal in *Arabidopsis*. Nature 404: 766–770.

Lin, T. et al. (2022). Extensive sequence divergence between the reference genomes of *Taraxacum kok-saghyz* and *Taraxacum mongolicum*. Sci. China Life Sci. 65: 515–528.

Liu, C., Xi, W., Shen, L., Tan, C., and Yu, H. (2009). Regulation of floral patterning by flowering time genes. Dev. Cell 16: 711–722.

Lockton, S. and Gaut, B.S. (2005). Plant conserved non-coding sequences and paralogue evolution. Trends Genet. 21: 60–65.

Lomsadze, A., Burns, P.D., and Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 42: e119–e119.

Luo, D., Carpenter, R., Vincent, C., Copsey, L., and Coen, E. (1996). Origin of floral asymmetry in *Antirrhinum*. Nature 383: 794–799.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., and Ravikesavan, R. (2013). Gene duplication as a major force in evolution. J. Genet. 92: 155–161.

Mandel, J.R., Dikow, R.B., Siniscalchi, C.M., Thapa, R., Watson, L.E., and Funk, V.A. (2019). A fully resolved backbone phylogeny reveals numerous dispersals and explosive diversifications throughout the history of Asteraceae. Proc. Natl. Acad. Sci. U. S. A. 116: 14083–14088.

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol. Biol. Evol. 38: 4647–4654.

Martín-Trillo, M. and Cubas, P. (2010). TCP genes: a family snapshot ten years later. Trends Plant Sci. 15: 31–39.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal 17: 10.

Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 41: e121–e121.

Nath, U., Crawford, B.C.W., Carpenter, R., and Coen, E. (2003). Genetic control of surface curvature. Science (80-.). 299: 1404–1407.

Navaud, O., Dabos, P., Carnus, E., Tremousaygue, D., and Hervé, C. (2007). TCP transcription factors predate the emergence of land plants. J. Mol. Evol. 65: 23–33.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32: 268–274.

Ning, K., Han, Y., Chen, Z., Luo, C., Wang, S., Zhang, W., Li, L., Zhang, X., Fan, S., and Wang, Q. (2019). Genome-wide analysis of *MADS-box* family genes during flower development in lettuce. Plant. Cell Environ. 42: 1868–1881.

Ohno, S. (1970). Evolution by gene duplication (Springer-Verlag , New York).

Palazzesi, L., Pellicer, J., Barreda, V.D., Loeuille, B., Mandel, J.R., Pokorny, L., Siniscalchi, C.M., Tellería, M.C., Leitch, I.J., and Hidalgo, O. (2022). Asteraceae as a model system for evolutionary studies: from fossils to genomes. Bot. J. Linn. Soc. 200: 143–164.

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. (2016). Evolution of gene duplication in plants. Plant Physiol. 171: 2294–2316.

Panero, J.L. and Funk, V.A. (2008). The value of sampling anomalous taxa in phylogenetic studies: Major clades of the Asteraceae revealed. Mol. Phylogenet. Evol. 47: 757–782.

Pei, Q. et al. (2021). Comparative analysis of the *TCP* gene family in celery, coriander and carrot (family Apiaceae). Veg. Res. 2021 11 1: 1–12.

Pertea, G. and Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. F1000Research 9: 304.

Pinyopich, A., Ditta, G.S., Savidge, B., Liljegren, S.J., Baumann, E., Wisman, E., and Yanofsky, M.F. (2003). Assessing the redundancy of *MADS-box* genes during carpel and ovule development. Nature 424: 85–88.

Pramanik, D., Dorst, N., Meesters, N., Spaans, M., Smets, E., Welten, M., and Gravendeel, B. (2020). Evolution and development of three highly specialized floral structures of bee-pollinated Phalaenopsis species. Evodevo 11: 1–20.

Rice, P., Longden, L., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16: 276–277.

Rosvall, M. and Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. U. S. A. 105: 1118–1123.

Ruelens, P., De Maagd, R.A., Proost, S., Theißen, G., Geuten, K., and Kaufmann, K. (2013). *FLOWERING LOCUS C* in monocots and the tandem origin of angiosperm-specific *MADS-box* genes. Nat. Commun. 4: 1–8.

Sato, S. et al. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485: 635–641.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13: 2498–2504.

Shen, C.Z., Zhang, C.J., Chen, J., and Guo, Y.P. (2021). Clarifying recent adaptive diversification of the Chrysanthemum-group on the basis of an updated multilocus phylogeny of subtribe Artemisiinae (Asteraceae: Anthemideae). Front. Plant Sci. 12: 874.

Smaczniak, C. et al. (2012). Characterization of MADS-domain transcription factor complexes in *Arabidopsis* flower development. Proc. Natl. Acad. Sci. U. S. A. 109: 1560–1565.

Smit, A., Hubley, R., and Green, P. (2019). 2013–2015. RepeatMasker Open-4.0.

Smith, R.I.L. and Richardson, M. (2010). Fuegian plants in Antarctica: Natural or anthropogenically assisted immigrants? Biol. Invasions 13: 1–5.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics 24: 637–644.

Stevens, P.F. (2017). Angiosperm Phylogeny Website. Version 14. Missouri Bot. Gard. St. Louis, Missouri USA.

Stuessy, T.F. and Garver, D. (1996). The defensive role of pappus in heads of Compositae. Compos. Biol. Util. 2: 81–91.

Susanna, A., Baldwin, B.G., Bayer, R.J., Bonifacino, J.M., Garcia-Jacas, N., Keeley, S.C., Mandel, J.R., Ortiz, S., Robinson, H., and Stuessy, T.F. (2020). The classification of the Compositae: A tribute to Vicki Ann Funk (1947–2019). Taxon 69: 807–814.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34.

Tahtiharju, S., Rijpkema, A.S., Vetterli, A., Albert, V.A., Teeri, T.H., and Elomaa, P. (2012). Evolution and diversification of the *CYC/TB1* Gene Family in Asteraceae--A Comparative Study in Gerbera (Mutisieae) and Sunflower (Heliantheae). Mol. Biol. Evol. 29: 1155–1166.

Theissen, G., Becker, A., Rosa, A. Di, Kanno, A., Kim, J.T., Münster, T., Winter, K.-U., and Saedler, H. (2000). A short history of *MADS-box* genes in plants. Plant Mol. Biol. 42: 115–149.

Theißen, G., Kim, J.T., and Saedler, H. (1996). Classification and phylogeny of the *MADS-box* multigene family suggest defined roles of *MADS-box* gene subfamilies in the morphological evolution of eukaryotes. J. Mol. Evol. 43: 484–516.

Theißen, G., Melzer, R., and Ruümpler, F. (2016). MADS-domain transcription factors and the floral quartet model of flower development: linking plant development and evolution. Dev. 143: 3259–3271.

Underwood, C.J. et al. (2022). A *PARTHENOGENESIS* allele from apomictic dandelion can induce egg cell division without fertilization in lettuce. Nat. Genet. 54: 84–93.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27: 737–746.

Verhoeven, K.J.F. and Biere, A. (2013). Geographic parthenogenesis and plant-enemy interactions in the common dandelion. BMC Evol. Biol. 13: 1–8.

Vijverberg, K., Van Der Hulst, R.G.M., Lindhout, P., and Van Dijk, P.J. (2004). A genetic linkage map of the diplosporous chromosomal region in *Taraxacum officinale* (common dandelion; Asteraceae). Theor. Appl. Genet. 108: 725–732.

Vijverberg, K., Milanovic-Ivanovic, S., Bakx-Schotman, T., and van Dijk, P.J. (2010). Genetic fine-mapping of *DIPLOSPOROUS* in *Taraxacum* (dandelion; Asteraceae) indicates a duplicated *DIP*-gene. BMC Plant Biol. 10: 154.

Vijverberg, K., Ozias-Akins, P., and Schranz, M.E. (2019). Identifying and engineering genes for parthenogenesis in plants. Front. Plant Sci. 10: 128.

Vijverberg, K., Welten, M., Kraaij, M., van Heuven, B.J., Smets, E., and Gravendeel, B. (2021). Sepal identity of the pappus and floral organ development in the common dandelion (*Taraxacum officinale*; Asteraceae). Plants 10: 1682.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: an Integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963.

Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., Guo, H., Kissinger, J.C., and Paterson, A.H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40: e49–e49.

Won, S.Y., Jung, J.A., and Kim, J.S. (2021). Genome-wide analysis of the *MADS-box* gene family in Chrysanthemum. Comput. Biol. Chem. 90: 107424.

Yeo, S., Coombe, L., Warren, R.L., Chu, J., and Birol, I. (2018). ARCS: scaffolding genome drafts with linked reads. Bioinformatics 34: 725–731.

Zhang, T., Zhao, Y., Juntheikki, I., Mouhu, K., Broholm, S.K., Rijpkema, A.S., Kins, L., Lan, T., Albert, V.A., Teeri, T.H., and Elomaa, P. (2017). Dissecting functions of *SEPALLATA*-like *MADS-box* genes in patterning of the pseudanthial inflorescence of *Gerbera hybrida*. New Phytol. 216: 939–954.

Zhao, T., Holmer, R., Bruijn, S. de, Angenent, G.C., van den Burg, H.A., and Schranz, M.E. (2017). Phylogenomic synteny network analysis of *MADS-box* transcription factor genes reveals lineage-specific transpositions, Ancient Tandem Duplications, and Deep Positional Conservation. Plant Cell 29: tpc.00312.2017.

Zhao, T. and Schranz, M.E. (2017). Network approaches for plant phylogenomic synteny analysis. Curr. Opin. Plant Biol. 36: 129–134.

Zhao, Y., Broholm, S.K., Wang, F., Rijpkema, A.S., Lan, T., Albert, V.A., Teeri, T.H., and Elomaa, P. (2020). TCP and *MADS-box* transcription factor networks regulate heteromorphic flower type identity in *Gerbera hybrida*. Plant Physiol. 184: 1455–1468.

4

7. Supplementary materials

Data is available via: 10.4121/21803436

Supplementary Figures

Supplementary Figure S1. Quality measurements of the T. officinale genome assembly.
Supplementary Figure S2. Quality measurements of the T. officinale genome assembly.
Supplementary Figure S3. Analysis of the T. officinale plastome assembly.
Supplementary Figure S4. Genome comparison of T. officinale genome versus the genome sequence of T. koksaghyz (top) and the genome sequence of T. mongolicum.
Supplementary Figure S5. Quality control of transcriptome data of dandelion floral developmental replicates.
Supplementary Figure S6. RNAseq analysis.
Supplementary Figure S7. Heatmaps of genes in genes associated to floral organ initiation and development in T. officinale.
Supplemental Figure S8. Gene trees of MADS-box and TCP.

Supplementary Tables

Supplementary Table S1. Statistics of Taraxacum officinale assembly.
Supplementary Table S2. Genome repeat characteristics of T. officinale.
Supplementary Table S3. Gene prediction statistics of T. officinale.
Supplementary Table 4. Genome comparison between three Taraxacum species: T. officinale, T. monogolicum and T. koksaghyz.
Supplementary Table S5. Read and mapping statistics of transcriptomes of floral tissues of T. officinale plant FCh72.
Supplementary Table S6. Characteristics of the T. officinale plant FCh72 transcriptome.
Supplementary Table S7. MADS-box and TCP gene expression in floral tissues of T. officinale, averaged per tissue type.
Supplementary Table S8. Species used for Synteny analysis.
Supplementary Table S9. Reference genes for MADS-box and TCP genes used in this study.
Supplementary Table S10a. Count of identified MADS-box and TCP genes.
Supplementary Table S10b. Summary statistics of identified *MADS-box* and TCP genes.
Supplementary Table S11. Examples of tandem duplicate of SEP3-FLC and SEP3-AS-MADS.
Supplementary Table S12. MADS-box and TCP Gene info.

Supplementary Data

Supplementary Data 1. Gene description.
Supplementary Data 2. Gene expression.
Supplementary Data 3. Identification and classification of MAD-box and TCP gene.
Supplementary Data 4. Synteny Network and profiling of identified MADS-box and TCP.
Supplementary Data 5. Genome-wide search for MADS-box tandem duplicates.

Chapter 5

General discussion

1. The era of plant genome sequencing

Most plant genomes were challenging to sequence and assemble due to their typically large sizes, high repetitive content and variable ploidy levels. In **Chapters 2, 3** and **4**, I created and analyzed *de novo* assemblies of two *Lactuca* wild relatives (*Lactuca saligna* and *Lactuca virosa*) and a common dandelion (*Taraxacum officinale*). The sequencing and assembling employed many state-of-art techniques, at least when the work was begun in 2017. I followed a general plan (or recipe) that combined PacBio long-reads and Illumina short-reads first to construct high-quality contig assemblies (Table 1). Then contigs were reoriented and concatenated to scaffolds (of contigs) and later super-scaffolds (of scaffolds) by different mapping technologies, including linked reads from 10x Genomics, optical mapping of Bionano and chromatin ligation of Dovetail (except for dandelion). These sequencing projects aimed to deliver high-quality chromosome-level assemblies for genome structure and sequence diversity studies, and test new strategies for genome reconstruction including optical mapping data (mentioned in **Chapter 1**).

Table 1. Summary of sequencing and scaffolding techniques used in each chapter.

Action	**Platform**	**Genome assemblies**		
		L. saligna **Chapter 2**	***L. virosa*** **Chapter 3**	***T. officinale*** **Chapter 4**
Sequencing	Illumina	x	x	x
	Pacbio	x	x	x
Scaffolding	10x Genomics	x	x*	x
	Bionano	x	x	x
	Dovetail	x	x	

* 10x Genomics reads were generated but not used in the final assembly of *L. virosa*.

Using the *de novo* assemblies, I then studied structural variation and gene family and sequence diversity related to important traits. Below, I will systematically discuss the collective insights gained from my sequencing projects and relate my work to future plant genome assembly projects and phylogenomics.

1.1 Assessment of sequencing strategies

To start with, the sequencing techniques plus mapping data produced genome assemblies of high quality and continuity for all three species (*L. saligna*, *L. virosa* and *T. officinale)*. This can be seen by the N50/L50 values and gene space (i.e., BUSCO values) of scaffolds (Table 2). Similar to lettuce (*L. sativa)* (Reyes-Chin-Wo et al., 2017), my assemblies of *Lactuca* wild relatives are also chromosome- or near-chromosome level after super-scaffolding. Regarding genome completeness, the BUSCO values for the *L. sativa* assembly and annotation are higher than for *L. saligna* and *L. virosa*.

However, they still both possess relatively complete gene spaces, illustrated by their high assembly (92 – 96%) and protein (89 – 90%) BUSCO values. It is worth noting that the gene space of *L. virosa* draft genome assembly was particularly low (~75%) before integrating the sequence data from Wei et al. (2021) (**Chapter 3**). I believe the reason for such incompleteness was due to the relatively low sequencing coverage of *L. virosa* (depth ~90 x), especially for the PacBio long reads (20x), which profoundly affected my ability to fully assemble the genome. Also, somehow the initial genome build was not polished and fixed later (Table 2: superscript 2). Explicitly, the sequencing depth of *L. virosa* is almost halved comparing to other mentioned *Lactuca* and *Taraxacum* species (Table 3), while it has the largest genome size with the highest amount of repetitive content (**Chapter 3**). Thus, the low coverage was caused by the larger genome size of *L. virosa* compared to the other species, which is due to budget constraints.

Table 2. Statistics of genome assemblies for *Lactuca* and *Taraxacum* species.

Species	Ploidy level	Het %	# seq contig	Seq length	N50/L50	Gen. BUSCO%	Pro. BUSCO%
*L. sativa**	2x	-	8,325	2.39 Gb	258 Mb/4	97.8	98.5
L. saligna	2x	0.12	10	2.16 Gb	239 Mb/4	92.4	88.8
*L. virosa***	2x	0.17	29	3.30 Gb	317 Mb/5	75.2	-
L. virosa	2x	0.17	5,855	3.45 Gb	317 Mb/5	96.2	90.2
T. officinale	2x	1.5	4,059	936 Mb	757 kb/286	97.2	90
*T. kok-saghyz****	2x	-	160	1.10 Gb	132 Mb/4	85.5	74
*T. mongolicum****	3x	1.3	65	790 Mb	97 Mb/4	92.9	69

Het%, heterozygosity rate; Gen., Genome; Pro., Proteome.

* Genome assembly (version 8) from Reyes-Chin-Wo et al. in 2017.

** Draft genome before extra polishing and assembling combination.

*** Genome assemblies from Lin et al. in 2022.

Although the *T. officinale* assembly had generally large (N50 = 756 Kb, L50 = 286) super-scaffolds, it is rather fragmented compared to the chromosome-level assemblies of two related species (*Taraxacum kok-saghyz* and *Taraxacum mongolicum*) (Table 2; Lin et al., 2022). While the sequencing depth was comparable to other two *Taraxacum* species (Table 3), *T. officinale* in my study didn't use Hi-C data after Bionano optical mapping as I did so for my two *Lactuca* species with larger scaffolds (Table 1), as the same in *T. kok-saghyz* and *T. mongolicum* (Lin et al., 2022). Nevertheless, the lack of this final mapping step didn't undermine contig assembling nor the annotation quality. Our *T. officinale* genome assembly has almost complete gene space (BUSCO = 97%) with 90% of BUSCO annotated, which are both significantly higher than its two *Taraxacum* relatives.

Table 3. Summary of sequencing depth for each genome assembly.

Species	Estimated size	Pacbio depth (x)	Illumina depth (x)
L. saligna	2.3 Gb	41	175
L. virosa	3.7 Gb	20	69.3
T. officinale	831 Mb	75	161
T. kok-saghyz[*]	1.1 Gb	55	133
T. mongolicum[*]	790 Mb	61	264

[*] Genome assemblies from Lin et al. in 2022.

1.2 Specificity and order of scaffolding techniques

As summarized by Sedlazeck et al. (2018), each scaffolding approach has its application strength and limits. In my work, the combination of three scaffolding tools was applied to *L. saligna* (**Chapter 2**) and *L. virosa* (**Chapter 3**) assemblies (Table 1: Scaffolding partition). 10x Genomic data was first generated to create separate *de novo* genome assemblies in addition to the hybrid assembly based on Pacbio and Illumina reads. Next, labelled Bionano DNA molecules was produced and assembled for another round of hybrid scaffolding. Finally, Hi-C data created by Dovetail was used to reconstruct the Bionano assembly into longer-range sequences (e.g. chromosomes). After each round of scaffolding, both *Lactuca* genome assemblies were significantly improved, except the 10x Genomic scaffolding for *L. virosa*, which is why 10x Genomic data was ultimately disregarded in the final assembly (Table 1). This insensitivity or specificity shows that the power of 10x Genomics scaffolding is dependent on the targeted genomes. The characteristic, that 10x Genomic is sparse sequencing of (i.e., not true long read) might explain the ineffectiveness for a large and highly repetitive genome like *L. virosa*, due to its poor resolution on cis-repetitive regions (Sedlazeck et al., 2018).

As for Bionano and Dovetail techniques, they can both effectively extend and improve the contigs/scaffolds of the *Lactuca* assemblies. However, from the Bionano to Dovetail assemblies, many scaffolds were broken and rejoined by HiRise pipeline using Hi-C data (Dovetail platform). These conflicts between the two types of mapping data highlight the accuracy differences and extrapolate that the order of scaffolding techniques can affect the final assembly quality. To further explore this phenomenon, I compared synteny between *L. sativa* and *L. saligna*, before and after Dovetail scaffolding (Figure 1). The synteny plot illustrates that better interspecific collinearity was achieved after Dovetail scaffolding broke the scaffold 13 and relocated them in two new Hi-C scaffolds for *L. saligna*. This result suggests that Dovetail have a better resolution for genome scaffolding than Bionano besides their complementary role to each other. My speculation is consistent with a previous research on the genome of *Medicago truncatula*, where it thoroughly describes the impact of reversing Dovetail and Bionano techniques (Moll et al., 2017). This study confirmed that the order of scaffolding tools clearly affects the continuity and completeness of genome assembly. Moreover, the authors suggest implementing Dovetail before Bionano because Dovetail can break chimeric scaffolds and better scaffold tiny sequences.

Our sequencing project of lettuce wild relatives provide knowledge on improving the sequencing and mapping strategy from the following aspects: i) enough sequencing depth is the cornerstone of a successful *de novo* assembly project, and ii) the effect of scaffolding platform varies in different genomes, and iii) ordering of scaffolding technologies makes a difference.

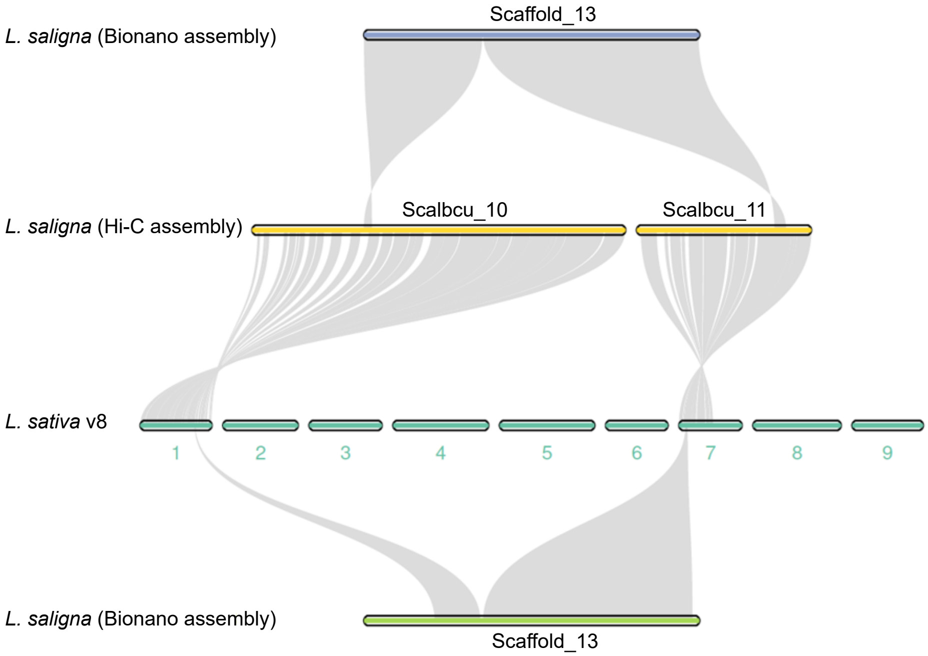

L. saligna (Bionano assembly)
Scaffold_13
L. saligna (Hi-C assembly)
Scalbcu_10
Scalbcu_11
L. sativa v8
1
2
3
4
5
6
7
8
9
L. saligna (Bionano assembly)
Scaffold_13

Figure 1. Synteny between *L. saligna* scaffold 13 and *L. sativa* chromosomes after Hi-C re-construction. Scaffold 13 of *L. saligna* Bionano assembly was broken and rejoined in two new super-scaffolds by Hi-C sequencing data. This synteny plot shows an mis-joined scaffold fixed by Hi-C data from Bionano scaffolding.

1.3 Long-range sequencing technologies now in 2022

All sequencing and scaffolding techniques applied for *de novo* assemblies in this project were advanced when the projects were originally conceived. However, sequencing technologies have continued to develop rapidly. I would like to briefly describe sequencing advances and give an outlook to the future and how my work, or projects like it, will be impacted. There are two major platforms of single-molecule long-read sequencing: Pacbio (SMRT) and Oxford Nanopore (ONT) as mentioned in the general introduction (**Chapter 1**). Compared to Pacbio used in this study, Nanopore released its first commercial sequencer later than Pacbio in 2014. The nanopore read (10 – 100 kb) is longer than Pacbio single-molecule real time read (multi-kb) with a high error rate (up to ~10%). However, both techniques suffered from the high error rate at the beginning, and companies have strived to improve them by product iteration. Later, Oxford Nanopore released nanopore ultra-long sequencing which can span extremely large distances (>100 kb). This technique was later used in many projects, for example, genome assemblies of human genome assembly (Jain et al., 2018), and rice (Tanaka et al., 2020). As the main competitor of Nanopore, in 2019, Pacbio released the improved SMRT platform named accurate circular consensus sequencing (CCS) that produces HiFi reads (Wenger et al., 2019). The CCS provides 20 kb-long reads with amazingly low error rate (0.1%), which was used as the main approach to construct the complete sequence of a human genome recently (Nurk et al., 2022). HiFi reads also have been used to construct plant genomes, like maize and Arabidopsis (Hon et al., 2020; Wang et al., 2022). Although nanopore ultra-long read can facilitate complete assemblies as well, the high error rate (> 5%) still imposes difficulties on assembling long and, near-identical repeat arrays (Nurk et al., 2022). For nanopore ultra-long reads, additional Illumina data need to be incorporated to gain a comparable accuracy (99.8%) of Pacbio HiFi read (99.9%; Jain et al., 2018). For plant genomes, HiFi read with a compromised read length but the highest accuracy, is more suitable considering their large size and high repetitiveness.

1.4 Long-read based haplotype phasing of the *T. officinale* genome

While *Lactuca* species are natural self-pollinators and thus highly homozygous, the sexual *Taraxacum* spp. are obligate outcrossers and thus can be heterozygous (~1.5%; Table 2). In **Chapter 4**, the *T. officinale* assembly is a mix of collapsed contigs (e.g. duplicate contigs were purged) from the two parental chromosomes. In other words, the genome is unphased (Figure 2, Top). Hence, the incomplete genetic makeup will limit the study of allelic variants of interesting genes and traits in dandelion. Also, some copy number and/or structural variants may complicate the assembly and the purging steps. Being a mixed genome also explains the larger assembly size (936 Mb) than the expected size based on C-value (831 Mb). After collapsing haplo-contigs, about 20 % of genes still remained as duplicated alleles, which could have complicated the study of gene variation, illustrated by the studies of *MADS-box* and *TCP* gene families in **Chapter 4**. Nowadays, the latest Pacbio CSS and nanopore ultra-long sequencing can phase haplotypes to complete the genetic variation (Figure 2, Bottom). It is worth noting that the other two chromosome-level Taraxacum are also unphased. In the future, it will be valuable to build a new fully phased genome assembly for *T. officinale* by adding extra HiFi or ultra-long reads to the current dataset.

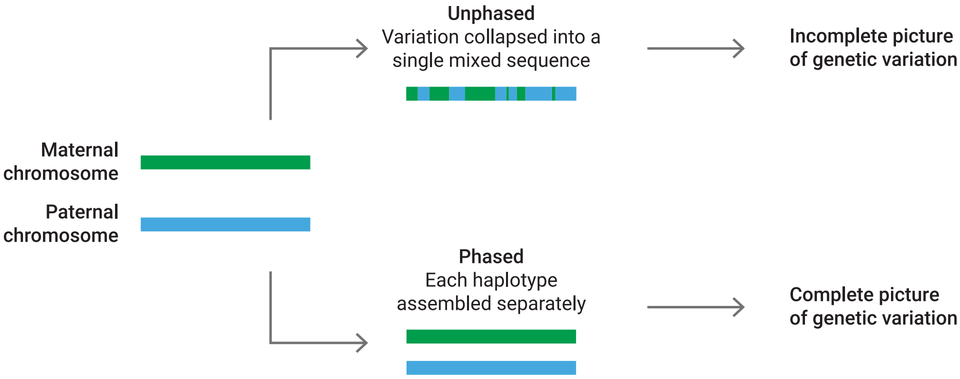

Unphased
Variation collapsed into a single mixed sequence
Incomplete picture of genetic variation
Maternal chromosome
Paternal chromosome
Phased
Each haplotype assembled separately
Complete picture of genetic variation

Figure 2. Haplotype phasing of diploid. Phasing separates the maternal and paternal chromosomes into individual haplotypes reaching a complete genetic makeup. Figure retrieved from Pacbio website: https://www.pacb.com/blog/ploidy-haplotypes-and-phasing/.

2. From the model species level to the family level for the Asteraceae

2.1 Arabidopsis as a dominant model plant in the past

In the last decades, *Arabidopsis thaliana* has served as the most important plant model species facilitating and dominating a vast swath of plant biology. Its annotated assembly has been used for functional prediction in all subsequent plant genome sequencing projects. One of the characteristics of *A. thaliana* genome is its small size (~135 Mb) and low chromosome numbers (1x = 1n = 5), which makes it easier to be sequenced, annotated, and validated. Gradually, the *A. thaliana* model started to reveal its limitations. One limit is that the genetic makeup of *A. thaliana* cannot represent the whole plant kingdom. After several rounds of updates, the current *A. thaliana* annotation (version 11) still contains more than 1,000 (out of 27,655) gene models with unknown functions. Based on *A. thaliana* data, other plants, with diverse genomic content cannot be fully annotated, thus a significant proportion of plant genes remain as "unknown". These limitations of *A. thaliana* restrict studies of biological understanding, potentially related to relevant traits in crops. Therefore, more crop models are required to not only complete the oversights of genetic variation in *A. thaliana*, but better transfer its well-studied biology to breeding industry.

Another limitation of focusing on *A. thaliana* is that it is phylogenetically limiting. More distant species are needed to understand trait evolution. For example, polyploidy events can drive trait diversification and species radiations. An excellent example can be found in Cleomaceae family, which is the sister of Brassicaceae (*A. thaliana*). In Cleomaceae, the whole-genome duplication (WGDs) *Gg-α* (Green star in Figure 3) occurred before the divergence of C_3 *Tarenaya hassleriana* and C_4 *Gynandropsis gynandra* after *Cleome violacea* divergence (Figure 3; Hoang et al., 2022). The Cleomaceae family plus Arabidopsis (C_3) therefore makes an outstanding model to study the evolutionary trajectory of C_4 photosynthesis. Accordingly, without proper phylogeny outgroup, it is difficult for comparative genomics analysis to study the gene and trait evolution when model and non-model species experienced independent ancient polyploidy events, like *A. thaliana* and *G. gynandra*. I specifically use this as an example, since I was responsible for the gene annotation of *G. gynandra*. In conclusion, it is essential to expand the scope of model plants to better understand genetics and biology under different evolutionary contexts.

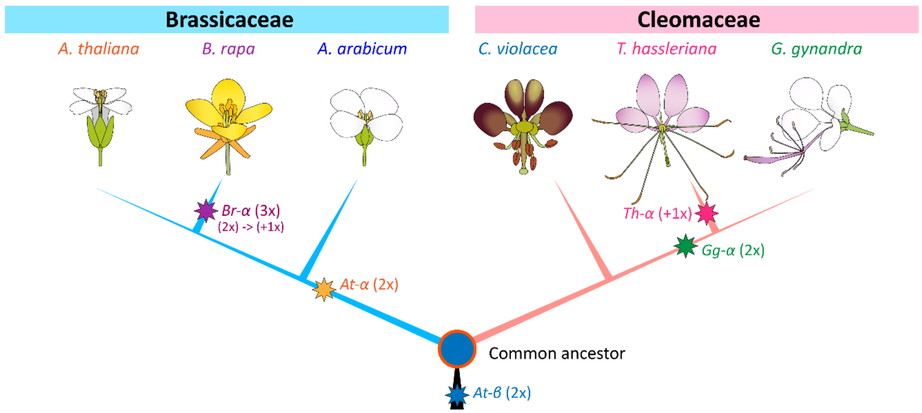


Figure 3. Evolutionary relationship between Cleomaceae and Brassicaceae species. This species tree summarizes the independent ancient ploidy events in these two families. The figure is adjusted based on the Figure 3 in *Gynandropsis gynandra* genome paper (Hoang et al., 2022).

2.2 Lettuce as a promising model for leafy crops

Currently, we don't yet have a robust model for leafy green vegetable crops that are comparable to model fruit (e.g., tomato) and seed (e.g., rice/maize) crops. By contrast, the production of leaves requires distinct traits such as late flowering. Lettuce, as a worldwide leafy vegetable, is a logical and emerging candidate to fill this gap. Over the last decades, extensive genomic data has been generated by many global sequencing efforts. In the 2010s, the Compositae Genome Project and the Lettuce Genome Sequencing Project generated the data for Crisphead lettuce (*L. sativa* L., cv Salinas), and the resulting genome was published in 2017 (Reyes-Chin-Wo et al., 2017) and deposited on the Lettuce Genome Resource website (https://lgr.genomecenter.ucdavis.edu/Home.php). Later, extensive re-sequencing data of 445 *Lactuca* accessions substantially broadened our knowledge of the domestication history of lettuce (Wei et al., 2021). In addition, a large-scale of RNA-seq data for 240 *Lactuca* accessions was also produced to study the lettuce evolution by Zhang et al. (2017).

The *de novo* sequencing of two wild lettuce species and re-sequencing of 100 selected *Lactuca* accessions were conceived in 2014 within the framework of the International Lettuce Genomics Consortium (ILGC mentioned in **Chapter 1**) co-funded by TKI (Netherlands) and breeding companies. Subsequently, the generated data from ILGC constitutes the backbone of this PhD project. Previously, the reference genomes with good quality (i.e., high completeness and continuity) of lettuce wild relatives didn't exist, as opposed to the genome reference for domesticated lettuce. Many genetic variants of breeding interest would therefore remain obscured in these wild lettuce species. From now, the reported *L. saligna* (**Chapter 2**) and *L. virosa* (**Chapter 3**) genome reference assemblies can be integrated with further large-scale genotyping and phenotyping projects, which lay a strong foundation for biological studies to develop lettuce into a model for leafy crops. For example, two ongoing research programs directly benefit from my work: i) the LettuceKnow project (https://lettuceknow.nl/) where genetic and phenotypic data will be generated for 500 *Lactuca* accessions, and ii) the Lettuce Genome Database (Guo et al., 2022), which is an integrative lettuce database ranging from genome, genotype, germplasm, phenotype and other omics data (https://www.lettucegdb.com/). Lastly, **Chapter 2** presents the study of non-host resistance phenotype against *Bremia lactucae*, which causes downy mildew disease that is responsible for the yield loss of lettuce leaves (discussed in next section of trait) using *L. saligna*. Future work can be done with this species and with crosses to *L. sativa* to more precisely study NHR.

2.3 Sequencing Cichorioideae members of the Asteraceae

Besides *Lactuca* wild relatives, another sequencing project of sexual common dandelion initiated in 2015 is described in **Chapter 4**. Because the co-existing sexual and asexual reproduction (i.e., apomixis), the common dandelion (*Taraxacum officinale*) has been a famous model for the studies of ecological evolution of the components of apomixis such as parthenogenesis (Underwood et al., 2022). It is worth mentioning, that I was co-author on this study (Underwood et al., 2022) due to my contribution on the synteny of the *PAR* genes between lettuce and *Taraxacum*. Beyond that, the *Lactuca* and *Taraxacum* spp. together represent the Cichorioideae, known as one of two major sub-families of the largest flowering family Asteraceae in addition to Asteroideae (Vijverberg et al., 2021). As mentioned in **Chapter 1** (General introduction), the Asteraceae family has extremely diverse habitats and characteristics implying its phenomenal genomic variation, which makes it an outstanding model for biological questions of evolution and diversification. To date, 39 genomes (majorly with economic importance) were sequenced and assembled of Asteraceae members including 14 species of Cichorioideae and 20 species of Asteroideae (Palazzesi et al., 2022). This project now added genome assemblies of three Cichorioideae species to the evolutionary framework of Asteraceae. Compared to economic species, these wild plants can also contribute to ecological studies in future researches. To illustrate the model function of Asteraceae (**Chapter 4**), I conducted a broad range of comparative genomics analysis to study the gene evolution of flower development for this family (discussed in next section of trait), including the three novel and other 9 released Asteraceae genomes.

3. Trait evolution in *Lactuca* species and Asteraceae family

3.1 Extracellular recognition of *Bremia lactucae* for lettuce downy mildew disease

In plants, pattern recognition receptors (PRRs) perceive the molecular patterns or effectors caused by microbe invasion and activate downstream defenses. There are two types of PRRs, receptor-like kinases (RLKs) and receptor-like proteins (RLPs) localized on the cell membrane. An RLK contains an extracellular domain (i.e., ectodomain), a transmembrane domain and an intracellular kinase domain, while an RLP is lack of the kinase domain. The extracellular domain of RLKs and RLPs are variable and essentially recognize the ligands from pathogens like bacteria, fungi, oomycetes, or nematodes. Although without a kinase domain, RLPs can recognize pathogen and associate with RLKs to form a heteromeric kinase for signal transduction. For example, the SOBIR1 (RLK) is a common partner for many RLPs (Gust and Felix, 2014).

In **Chapter 2**, I investigated the genetic basis of the non-host resistance (NHR) phenotype in wild lettuce *L. saligna* to lettuce downy mildew disease caused by the oomycete (*Bremia lactucae*). The combination of genome-wide search and expression analysis suggests a potential role of *RLK* genes in defense of *L. saligna* against *Bremia*. Several up-regulated *RLK*s on the previously mapped NHR locus (chromosome 8) were identified, including a tandem array of wall-associated kinases (WAK). WAKs could potentially recognize an endogenous pattern-derived effector from a degraded cell wall caused by oomycete enzyme(s) (endogenous pattern in Fig. 4). However, many elicitors secreted by the oomycete (exogenous patterns in Fig. 4) are perceived by RLPs (Raaymakers and Van Den Ackerveken, 2016), for example, a 20 amino-acid fragment (nlp20) common in pathogens can be perceived by RLP23 with the help of RLKS (BAK1 and SOBIR1; Albert et al., 2015), as demonstrated in Figure 4. Thus, it is critical to perform additional genome-wide searches and expression analysis for the RLP encoding genes in *L. saligna* genome, to depict a complete distribution of the extracellular recognition system. Based on **Chapter 2**, an independent RNA-seq project of a *Bremia*-infection bioassay was also conducted within my PhD framework, including three *Lactuca* genotypes (*L. saligna*, *L. sativa* and resistant introgression line) at different time-points. However, the results of this analysis are not included in the thesis, but will be later prepared for publication. A comprehensive comparative transcriptomic analysis is being performed to capture and validate all related isoforms (e.g., RLK/RLP) among the three materials. In addition to the pattern-triggered immunity (PTI) activated by PRRs or effector-triggered immunity (ETI) by nucleotide-binding leucine-rich repeat receptors (NLRs), plant hormones or phytohormones (small molecules) are also widely involved in plant defense. Among them, jasmonate (JA) and salicylic acid (SA) are two major phytohormones associated with plant defense (Berens et al., 2017).

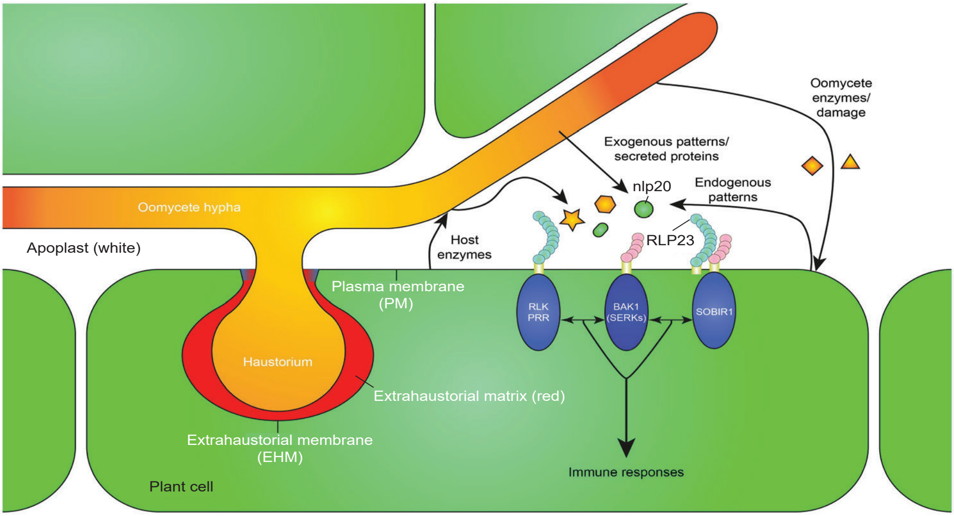


Figure 4. Plant immunity against oomycete infection induced by the recognition of exogenous and endogenous patterns. Oomycete can cause two types of molecules which can be recognized by pattern recognition receptors (PRRs). Cell wall that damaged by oomycete releases fragment as endogenous pattern activating immunity. Receptor-like protein (RLP) like RLP23 can perceive the exogenous pattern secreted by oomycete (e.g., proteins with nlp20 pattern), and trigger the immunity with the association of SUPPRESSOR OF BIR1 1 (SOBIR1) and BRI1-ASSOCIATED RECEPTOR KINASE 1 (BAK1; Albert et al., 2015). Figure is adjusted from the review by Raaymakers and Van Den Ackerveken (2016).

3.2 The Defense hormone salicylic acid and its regulation in *L. saligna* upon *Bremia* infection

The plant hormone salicylic acid (SA) is a critical regulator in defense against pathogenic microbes. It is required for both resistance at the local infected-sites and distant uninfected-sites (Ding and Ding, 2020). For example, in Arabidopsis SA accumulation is associated with hypersensitive response (HR) at the infected location (Devadas and Raina, 2002). In TMV-infected tobacco, the depletion of SA level will lead to reduced systemic resistance at distal tissues (Gaffney et al., 1993; Yalpani et al., 1991), where such immunity is also known as systemic acquired resistance (SAR). In plants, endogenous SA and exogenous application of SA can both induce defense (Ding and Ding, 2020). For lettuce growing, applying exogenous SA has proved to enhance yield (Youssef et al., 2017) and resilience to abiotic stresses like salinity (Babaousmail et al., 2022). However, the understanding of both endogenous and exogenous SA is lacking for response to biotic stress in lettuce. As mentioned, **Chapter 2** has described the potential effect of PRRs (RLP/RLK) in the immune response (e.g., NHR) after *Bremia* infection, where the regulatory role of plant hormones like SA was not highlighted. In hindsight, my differential expression analysis of RNA-seq also identified many genes relevant to the up- and down-stream SA pathway. For example, Table 4 lists four SA-related genes that were differentially expressed after *Bremia-infection* (**Chapter 2**), which are the homologs of *SARD1*, *UGT76B1*, *ABCG40* and *PEN3* in Arabidopsis.

Table 4. Salicylic acid (SA) related DEGs upon *Bremia*-infected *L. saligna*.

Gene ID	TAIR homolog	Other name
Lsal_1_v1_gn_2_00003439	AT3G11340	UGT76B1
Lsal_1_v1_gn_8_00004656	AT1G73805	SARD1
Lsal_1_v1_gn_1_00001172	AT1G15520	ABCG40
Lsal_1_v1_gn_4_00003154	AT1G59870	PEN3

* This table is extract from the Supplemental Table 28 in **Chapter 2**.

In terms of synthesis and metabolism (Figure 5), SARD1 activates the translation of key factor *ICS1* for IC pathway of SA biosynthesis, while *UGT76B1* encodes a glycosyltransferase that converts the defense hormone SA to immunity-inactive SA-glycoside (SAG). The expression of *SARD1* was stably up-regulated at moderate levels at 8- and 24-hours post-infection (8phi and 24hpi), while the translation of *UGT76B1* spiked at 8hpi and mitigated at 24hpi. For post-transcription, the SA-dependent *ABCG40* and *PEN3* were both up-regulated after inoculation at two time points. Interestingly, the activity of *PEN3* gene still amplified compared to its high-expression in control plants. This new result strongly suggests that SA is involved in the resistance of *L. saligna* to *Bremia*. Moreover, the simultaneous up-regulation of SARD1 and UGT76B1 demonstrates a precise modulation of SA via positive and negative transcriptional regulation to optimize the plant defense.

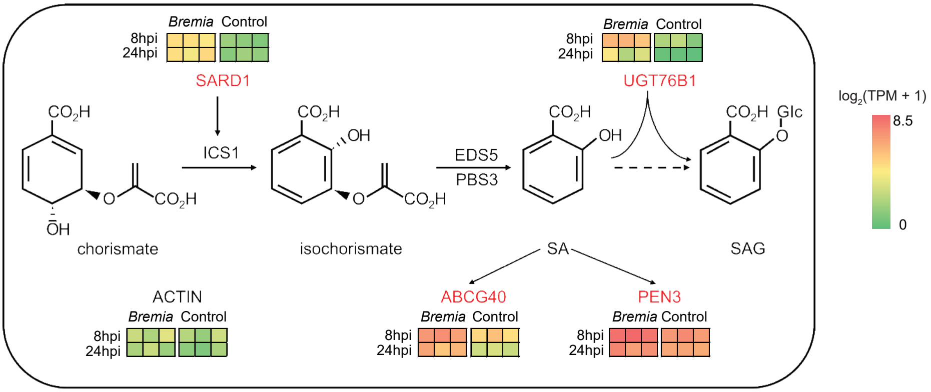


Figure 5. Differentially expressed genes involved in regulation of salicylic acid (SA) biosynthesis and metabolism tuning on its defense responses. Differential expression (DE) analysis was carried out in **Chapter 2** upon a Bremia infection bioassays at two time points after inoculation (8 hpi and 24 hpi). By revisiting the DE data, fours genes related to biological activity of SA were identified: *SARD1* involved in SA synthesis, *UGT76B1* involved in SA depletion, while *ABCG40* and *PEN3* respond to SA regulation. The proteins encoded by the four genes were displayed on the SA biosynthesis pathway with their expression in different treatments and time points. The expression of housekeeping gene ACTIN was used as a reference.

3.3 Outlook the role of acquired resistance in non-host resistance of *L. saligna* to Bremia

Plant immunity can be classified into innate (e.g., PTI induced by RLK/RLP and ETI induced by NLR) and acquired resistance (e.g., SAR mentioned in previous section). Many studies have shown that the trigger of SAR is also dependent on the same elements of the innate resistance system, for example, an elicitor of *Phytophthora* can activate both HR and SAR in tobacco and Chinese cabbage (Wang et al., 2003). While in some cases, the NLRs can induce SAR via the SA-dependent path without causing HR. Moreover, the SAR of the non-host Arabidopsis is found to be enhanced after the recognition of a *Pseudomonas syringae* elicitor by PRRs, which doesn't cause necrosis in Arabidopsis. As for the interaction between *Bremia* and *L. saligna*, the former cannot establish haustoria in its non-host *L. saligna* post-penetration (den Boer, 2014). A previous study by Giesbers et al. (2017) showed that HR induced by *NLR* genes is beneficial but not required for NHR in *L. saligna*, while **Chapter 2** of this project has suggested the potential role of *RLK/RLP* genes. However, the effect of SAR on NHR in *L. saligna* and its relationship with important immune proteins (i.e., NLR and RLK/RLP) were barely discussed. Here, I propose to expand the scope of the potential NHR mechanism in *L. saligna* to acquired resistance. First, the SAR responses should be described after *Bremia* infection. Then, the relationship between immune proteins and SAR activation needs to be checked: SAR can be modulated by the interplay of N-hydroxypipecolic acid (NHP) and SA (Hartmann and Zeier, 2019), therefore the induction of gene expression for central enzymes involved in

SA (e.g., *ICS1*, *EDS5* and *PBS3*) and NHP (*ALD1*, *SARD4* and *FMO1*) biosynthesis should be closely monitored after elicitor or effector treatments.

3.4 Missing elements in comparative genomics analysis of Asteraceae floral development

In **Chapter 4**, I performed a broad range of phylogenomic analyses to assess the evolution of *MADS-box* and *TCP* gene families and their potential roles in floral development. My study mainly focused on *Taraxacum* and the Asteraceae family. Genes located in lineage-specific contexts (i.e., genomic or evolutionary position) were identified and their expression across floral development was analyzed by RNAseq. However, some interesting results were not mentioned in Chapter 4, which I would like to present and discuss now. One of the main goals of this PhD project was to provide biological knowledge for lettuce breeding. For lettuce, the late-flowering (i.e., transition from vegetative to reproductive phase) trait is favorable because bolting (rapid stem elongation) before flowering brings the bitter taste to lettuce leaves, which is undesirable to consumers. In Arabidopsis, the *FLOWERING LOCUS T* (*FT*) is one of the major regulators promoting flowering (i.e., after bolting; Kardailsky et al., 1999), which is inhibited by the complex of *SHORT VEGETATIVE PHASE (SVP)* and *FLOWERING LOCUS C (FLC*) (Adrian et al., 2010). Similar to Arabidopsis, a previous study has shown the correlation between the *FT* homolog in lettuce (LsFT) and bolting induced by high temperature (Fukuda et al., 2017). In my synteny network of **Chapter 4**, a lineage-specific (100%) synteny cluster array was revealed for the *MADS-box* genes in three *Lactuca* species (Figure 6: *MADS-box*). This cluster array contains three synteny clusters, which are annotated as *FLC*-like (Cluster 1 and 2) and Mα (Cluster 3). Potentially, future studies of these *FLC* genes in *Lactuca*-specific synteny cluster would be valuable, as they may have antagonistic roles with *FT* genes. The gained knowledge of phylogeny and expression pattern of mentioned *FLC*s therefore might provide novel insights for the bolting resistance in lettuce breeding.

Besides *MADS-box* and *TCP* genes, there are other transcription factors (TFs) involved in the shaping of floral morphology. For example, the *RADIALIS* (*RAD*) and *DIVARICATA* (*DIV*) from *MYB* gene family. They were identified in *Antirrhinum* and found to play an essential role in ventral and dorsal development determining flower symmetry (Galego and Almeida, 2002; Corley et al., 2005). Moreover, RAD and DIV can also interact with *TCP* members, like the specialized *CYC-RAD* transcription factor network in *Antirrhinum, which differs from Arabidopsis (Costa et al., 2005). In addition, TOPLESS* (*TPL*), with encodes a co-repressor family with a broad range of targets, can inhibit the expression of TCP genes (Wei et al., 2015). Moreover, TPL interacts with the EAR domain (Martin-Arevalillo et al., 2017). While the *Parthenogenesis* gene (*PAR*) identified in apomictic dandelion (*T. officinale*) also contain EAR (Underwood et al., 2022), which makes it a potential target of TPL. This could be further supported by the interaction of *PAR* homologs (e.g., *DAZ1*, *2*, *3*, and *TREE1*) with TPL in Arabidopsis (Borg et al., 2014; Wang et al., 2020). Therefore, a synteny network plus phylogeny analysis for *MYB* and *TPL* will be highly valuable to dissect the regulation network of Asteraceae and trait evolution like apomixis (illustrated by Figure 6: *TOPLESS* clusters). Given the potential network, additional experimental assay like yeast on-hybrid (Y1H) can further recognize cis-regulatory TFs promoting *TCP* expression, like the interaction between *TCP* and *MADS-box* in gerbera (Zhao et al., 2020).

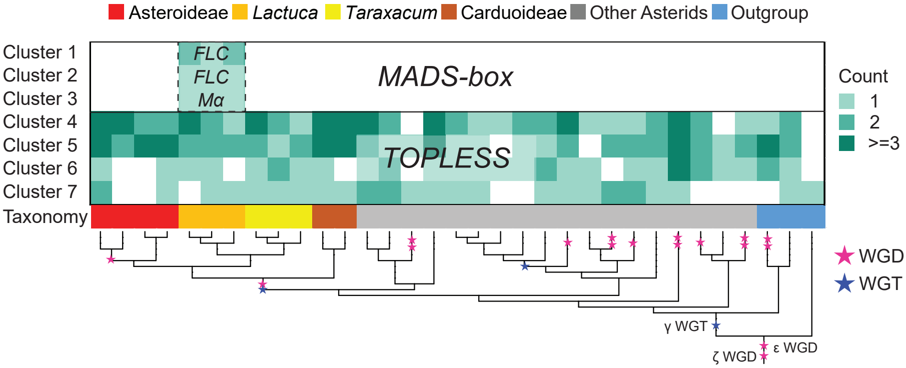


Figure 6. Phylogenetic profiling illustrates the specific-context of *MADS-box* in *Lactuca* and general-context of *TOPLESS* in Asteraceae. Gradient green cells show the number of syntelog (syntenic homolog) for each cluster in the different species. Species were divided into 6 groups based on taxonomy and indicated by different colors: *Asteroideae* sub-family (red); *Taraxacum* genus (yellow); *Lactuca* genus (orange); *Carduoideae* (brown); other species from Asterids clade (grey); three basal species as outgroup (blue). Phylogenetic profiling identified *Lactuca*-specific *MADS-box* clusters, and demonstrated the syntenic relationship of *TOPLESS* gene across Asterids. Pink and blue stars represent known ancient polyploidy events.

4. Relationship between duplication mechanism and gene types

Gene duplication can provide the natural genetic variation needed for adaption and trait. For plant genomes, the percentage of duplicate genes is certainly high, where whole-genome duplication (WGD) acts the most dramatically contributing the gene content from the 25% in Arabidopsis up to 67% in soybean (Blanc and Wolfe, 2004; Schmutz et al., 2010). Apart from WGD, duplicated genes may also be produced by tandem duplication, transposon-mediated duplication, segmental duplication and retroduplication (reviewed in Panchy et al., 2016). Through all research projects (**Chapters 2, 3 and 4**), I studied the variation and evolution of several genes for different traits in *Lactuca* and *Taraxacum* species, while duplication has consistently been an crucial element in a rather implicit manner. Following that, I would like to discuss the effect of duplication types on different scenarios for gene evolution.

4.1 The tangle of tandem and transposed duplicated immune genes

Chapter 3 showed the relationship between immune genes (i.e., *NLR* and *RLK)* expansion and tandem duplication in *Lactuca sativa*, compared to the other two *Lactuca* species. Moreover, *L. virosa* underwent a genome expansion induced by TE proliferation (**Chapter 3**). It would be worthwhile to examine how transposon expansions have affected the expansion of immune genes *NLR* and *RLK*. In **Chapter 3**, an extra counting was done to calculate the LTR density at the flanking regions (+/- 5 kb) in three *Lactuca* species genome references for different types of genes. The results are summarized by the line charts in Figure 7. Because of the high N content in genome assemblies, the flanking density of *L. virosa* for complete gene set (top), *RLK* (middle) and *NLR* (bottom) were smaller than the *L. sativa* and *L. saligna* as expected. Therefore, I won't include *L. virosa* in the following discussion of LTR density. Interestingly, the line charts illustrate that *L. sativa* has a higher LTR density than *L. saligna* at both *NLR* and *RLK* flanking regions, while the reverse trend was observed in the complete gene set. These opposing results imply that the LTR/transposon activity might also partially explain the expansion of *NLR* and *RLK* among *Lactuca* species (**Chapter 4**). Such a hypothesis is further supported by the positive correlation between tandem duplication and transposition in Arabidopsis for *NLR* duplication (Freeling et al., 2008). Therefore, if the expansion in *L. sativa* is valid, I extrapolate that the *LTR* density near *L. virosa* immune genes is also lower than *L. sativa*. To conclude, the tandem and transposon-derived duplication should be collectively researched for specific genes, like *NLR*s and *RLK*s.

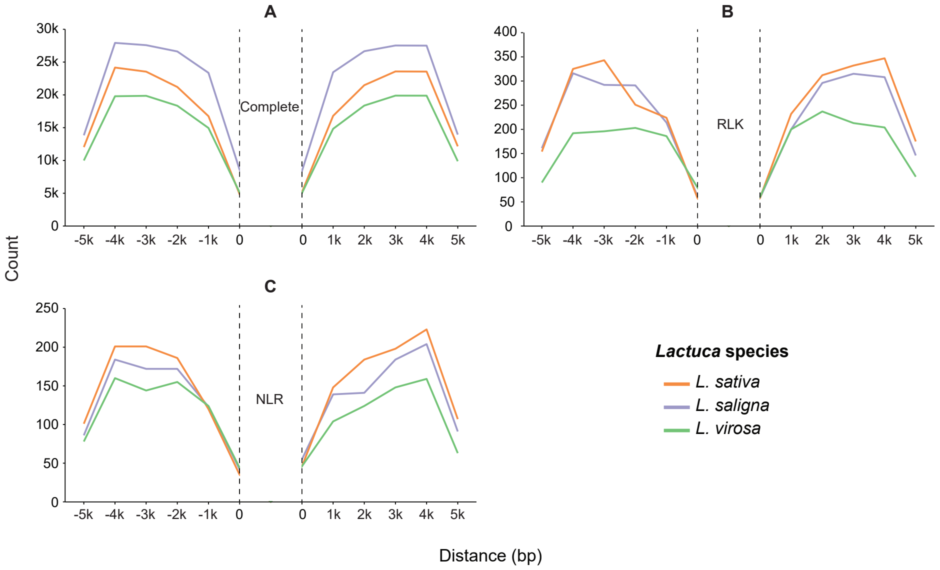


Figure 7. LTR distribution at flanking regions (+/- 5kb) of various genes on *Lactuca* spp. genomes. LTR number was counted and compared to three *Lactuca* genomes via a custom script base on bedmap tool. Density of LTR for up and down 5 kb intergenic regions were visualized by line plots (bin = 1000 bp) targeting different gene datasets: A for complete gene sets; B for *RLK* genes; and C for *NLR* genes.

4.2 Unbalanced gene retention after WGDs: *MADS-box* vs *SARD4*

In **Chapter 4**, I evaluated the impact of polyploidy events on *MADS-box* evolution for flower development. *MADS-box*, as the important gene family for plant morphology structure, is comparable to the role of Hox genes in animals (Theißen et al., 1996; Ng and Yanofsky, 2001). While Hox genes in animals are organized into tandem arrays, *MADS-box* genes spread across plant genomes after whole-genome duplication and fractionation. For example, Arabidopsis has over 100 *MADS-box* homologs spread across its five chromosomes. This distribution of *MADS-box* genes reflects the overall plasticity of plant genomes and the importance of WGD for plant evolution. However, there are also extremely conserved genes in plants that always return to single-copy, even after multiple rounds of WGD and fractionation, many of which function is often essential to all eukaryotes (De Smet et al., 2013). One incredible case is the *SARD4* mentioned in the last section,, which is an important enzyme for NHP biosynthesis. To investigate this scenario, a homolog search of SARD4 was conducted by an MSc student, Kooistra (2022), which I supervised. With my help, he analyzed more than 200 plant species using a previously published dataset from Pancaldi *et al.*, 2022. The result on the plant species tree stunningly demonstrates the 1:1 homologous relationship for *SARD4* across plant kingdom apart from a few exceptions (Figure 8), compared to the *MADS-box* tree for 33 species (**Chapter 4**).

The extreme contrast between *MADS-box* (e.g., Arabidopsis: ~100 copies) and *SARD4* (e.g., Arabidopsis: 1 copy) gene(s) resembles the comparison between Brassicaceae (~3,700 species) vs. Aethionemeae (~60 species) families. Such significant size-difference in two families was first explained by the Whole Genome Duplication Radiation Lag Time model (WGD-RLT; Eric Schranz et al., 2012), where the model claims a diversification lags after a shared WGD event. In addition, Tank *et al.* (2015) emphasized the importance of natural selection (e.g., geological or climate changes) for species radiation on the basis of WGDs. For gene duplication after WGD, gene loss is unbalanced by selection pressure biased towards conserved function (De Smet et al., 2013). Consequently, the single-copy scenario of *SARD4* in plants could be explained by the extremely severe selection or conserved preservation after re-occurring WGD events, which stresses its highly indispensable role and specialized function. To conclude, it is essential to take the gene type and function into account when studying different duplication mechanisms for gene evolution.

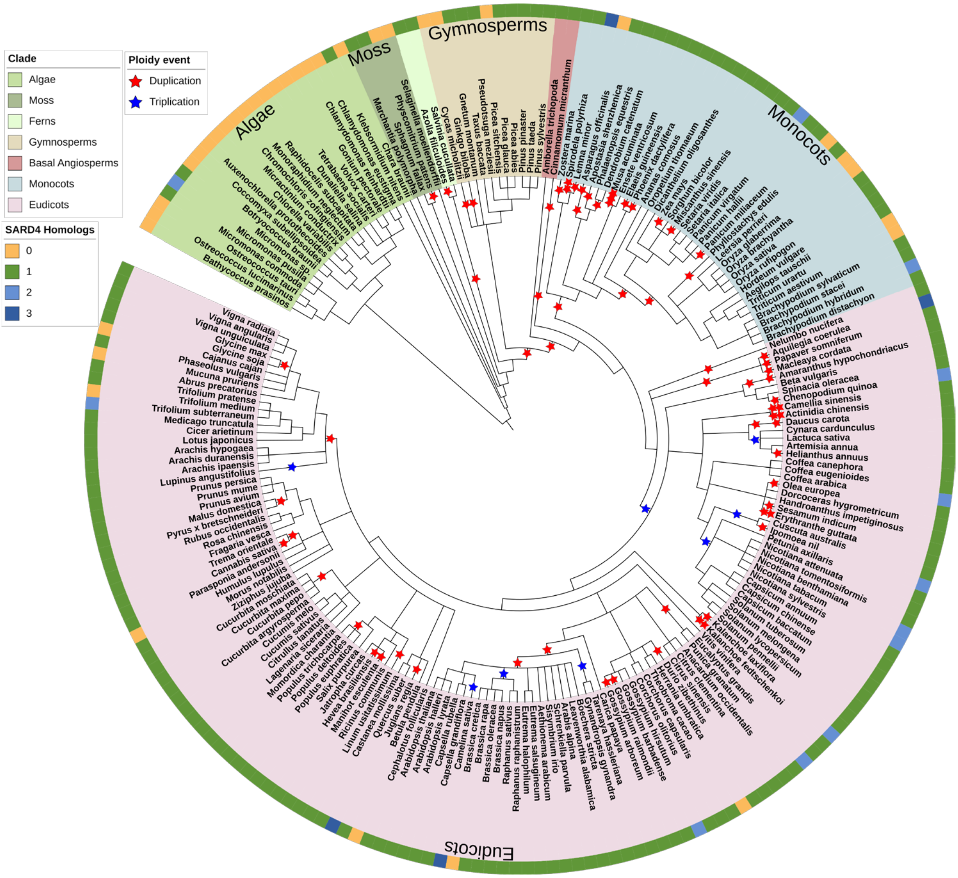


Figure 8. *SARD4* variation across the plant kingdom. The homolog(s) of *SARD4* were searched in 211 plant species including algae, moss, ferns gymnosperm, and angiosperm (monocots and eudicots). The color of outside strip indicates the copy number of *SARD4* homolog in each species. The evolutionary tree was based on the NCBI Common Tree (Schoch et al., 2020). Red and blue stars on the species tree represent shared and lineage-specific polyploidy events. Figure was retrieved from the MSc research practice report of Kooistra (2022).

5. If I Could Travel by in Time

> "Like all vain men, he had moments of unreasonable confidence."
> — Warren Eyster, The Goblins of Eros

As mentioned at the beginning of this chapter, we are now in the great era of genome sequencing. Diverse systems, organisms, tissues, and cells were and are being sequenced at an unprecedented speed. Looking back to the starting point of my PhD, I was thrilled and felt blessed to already have access to such incredibly large and diverse sequencing datasets (Figure 9). I was confident that my success was secured and pictured great publishments with these data. Though, I was not ready to carry the weight of the datasets due to my lacking of knowledge, skills and experience.

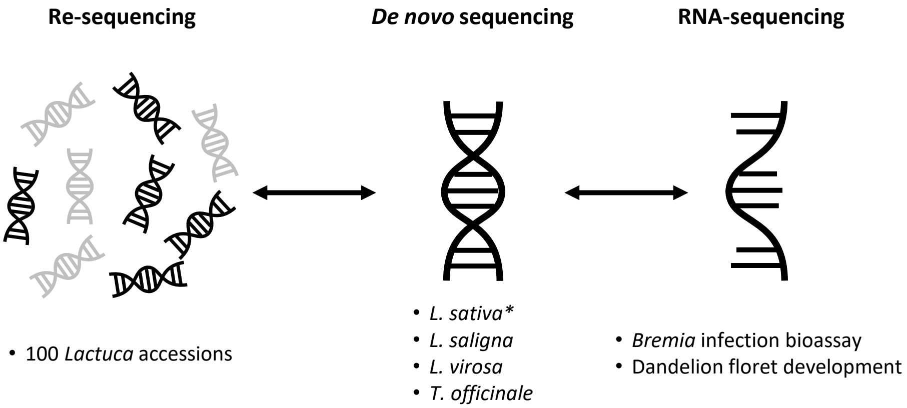


Figure 9. Summary of sequencing data used in my PhD thesis. Three main types of sequencing data were generated as described in the general introduction (**Chapter 1**) and the sequencing section start the general discussion (**Chapter 5**). They are: re-sequencing data (left) of 100 selected *Lactuca* accessions including the three *Lactuca* species studied; whole-genome sequencing (middle) and RNA-sequencing data (of different tissues) for the assembling and annotation of *L. saligna* (**Chapter 2**), *L. virosa* (**Chapter 3**) and *T. officinale* (**Chapter 4**); additional RNA-sequencing data (right) for the functional validation of identified genes for lettuce downy mildew resistance against Bremia (**Chapter 2**) and developmental regulation of dandelion flower (**Chapter 4**). * Genome assembly (version 8) retrieved from previous research (Reyes-Chin-Wo et al., 2017).

My over-confidence reminds me of the Dunning-Kruger effect (Dunning, 2011), which defines the overestimation of performance by a group that lack key competences (Figure 10 left: red area). This can be transformed to a more understandable curve illustrating the personal-development (Figure 10 right): A person starts with potential over-confidence but low competence, and then falls into the valley of despair when he/she encounters real situations. Yet, here is also where they can start to develop their skills and thus (re-)gain confidence. After climbing the learning curve, finally one will arrive at the plateau of sustainability where skilled people tend to underestimate their competences (Figure 10 left: green area). For me, who just got out of despair valley, it is time to rest on the hillside and take a reflection moment. I cannot help wondering what I will do differently to the datasets if I can time travel for a better project.

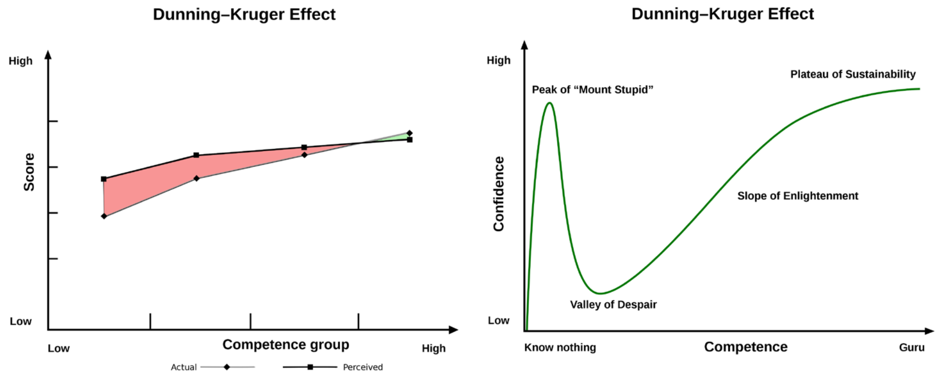


Figure 10. Dunning-Kruger effect: "Original science" Vs "Understanding curve". Left figure is based on the research of Dunning (2001), where shows the relation between self-perceived performance and actual performance on exam. Right figure of curve is a reading-friendly version to illustrate Dunning-Kruger effect in a understanding way.
Left figure created by 忍者猫(https://commons.wikimedia.org/wiki/File:Dunning%E2%80%93Kruger_Effect_01.svg), and right figure created by Diego Moya (https://en.wikipedia.org/wiki/File:Dunning%E2%80%93Kruger_Effect2.svg).

Figure 11A summarizes the diverse analyses and data used in my thesis. It reveals a significant overlap of biological questions (reflected by analyses) and objectives (reflected by sequencing data) between **Chapter 2** and **Chapter 3**. My personal development explains such planning, where I first learnt different bioinformatic skills and explored diverse possibilities for *L. saligna* genome before reusing them for the *L. virosa* project. Such planning touched on many biological questions at the same time, and repeated them in separate genome projects. Efficiency-wisely, it is hard to dig deep and take full advantage of the sequencing data. Moreover, a reliable genomic study is based on solid genome assembly and annotation. During my thesis, several times I had to improve or fix something in my assemblies, which thus forced me to redo downstream analyses.

Nevertheless, I am grateful for the various acquired skills and perspectives. Hence, I came up with an alternative PhD plan (Fig. 11B), hoping it can help the past me from a parallel universe. The new plan is based on two main criteria aiming for an efficient data usage: i) one biological question or goal per time, and ii) study *L. sativa*, *L. saligna*, and *L. virosa* as a whole. The new plan contains five projects. ***Project 1*** exploits the published *L. sativa* genome (Reyes-Chin-Wo et al., 2017) as reference for read mapping of TKI-100 re-seq data to study the domestication history and genetic diversity of important traits (GWAS) for lettuce, like the whole-genome sequencing project of 445 *Lactuca* accessions published by Wei et al. in 2021. After finalizing the genome and annotation of *L. saligna* and *L. virosa* together, I would have run a three-way homology group done in **Chapter 2** and detect synteny (like **Chapter 2 and 3**) thereafter to unveil the interspecific genome rearrangements (i.e., structural variation), as demonstrated by ***Project 2***. In this way, I can describe the homologous relationship and genome collinearity for three representative *Lactuca* spp simultaneously. Besides, the finalized genome assemblies and annotations would be ready for downstream analyses on different topics. As resistance is the major goal of lettuce breeding, I would have identified and classified the *NLR* and *RLK* for all three *Lactuca* species in ***Project 3*** like I did in **Chapter 3**, and search candidate genes for NHR using the RNA-seq of infected and un-uninfected *L. saligna*. Based on homologous relationships, I could have speculated whether the genetic determinant is also present in *L. virosa*. For ***Project 4***, I would have only focused on the genome evolution (i.e., genome size difference) among *Lactuca* spp., like the comparative repeatomics analysis in **Chapter 3.** Finally, ***Project 5*** is same as **Chapter 4**. Following this proposed project, I believe one could fully exploit the potential of existing sequencing data and save time to study every topic in profound depth. In addition to all the data and methods, it is valuable to create more sequencing data and genetic markers for *L. virosa* to elevate the assembly to chromosome-level, which can bring all mentioned projects to a higher level. By imagining this alternative me, I hope to help the future and real me to thrive in this great era of next-generation sequencing!

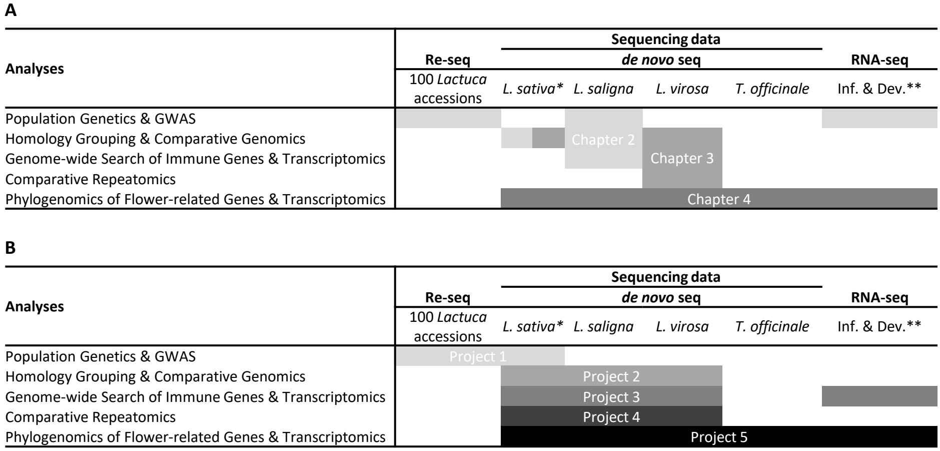


Figure 11. "Current project structure" Vs "Time-travelled structure". This figure shows my idea of how to restructure my PhD project with datasets and analyses remaining the same. **A,** current plan of data and methods used in each chapter. **B,** proposed plan using the same data and methods in a novel orientation, represented by 5 projects. * *L. sativa* genome assembly (version 8) from previous research (Reyes-Chin-Wo et al., 2017). ** Inf. for *Bremia* Infection and Dev. for floral development.

6. Reference

Adrian, J., Farrona, S., Reimer, J.J., Albani, M.C., Coupland, G., and Turck, F. (2010). Cis-regulatory elements and chromatin state coordinately control temporal and spatial expression of *FLOWERING LOCUS T* in *Arabidopsis*. Plant Cell 22: 1425–1440.

Albert, I. et al. (2015). An RLP23-SOBIR1-BAK1 complex mediates NLP-triggered immunity. Nat. Plants 1: 1–9.

Babaousmail, M., Nili, M.S., Brik, R., Saadouni, M., Yousif, S.K.M., Omer, R.M., Osman, N.A., Alsahli, A.A., Ashour, H., and El-Taher, A.M. (2022). Improving the tolerance to salinity stress in lettuce plants (*Lactuca sativa* L.) using exogenous application of salicylic acid, Yeast, and Zeolite. Life 12: 1538.

Berens, M.L., Berry, H.M., Mine, A., Argueso, C.T., and Tsuda, K. (2017). Evolution of hormone signaling networks in plant defense. Annu. Rev. Phytopathol. 55: annurev-phyto-080516-035544.

Blanc, G. and Wolfe, K.H. (2004). Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell 16: 1679–1691.

den Boer, E. (2014). Genetic investigation of the nonhost resistance of wild lettuce, *Lactuca saligna*, to lettuce downy mildew, *Bremia lactucae*. edepot.wur.nl.

Borg, M., Rutley, N., Kagale, S., Hamamura, Y., Gherghinoiu, M., Kumar, S., Sari, U., Esparza-Franco, M.A., Sakamoto, W., Rozwadowski, K., Higashiyama, T., and Twell, D. (2014). An EAR-dependent regulatory module promotes male germ cell division and sperm fertility in *Arabidopsis*. Plant Cell 26: 2098–2113.

Corley, S.B., Carpenter, R., Copsey, L., and Coen, E. (2005). Floral asymmetry involves an interplay between TCP and MYB transcription factors in *Antirrhinum*. Proc. Natl. Acad. Sci. U. S. A. 102: 5068–5073.

Costa, M.M.R., Fox, S., Hanna, A.I., Baxter, C., and Coen, E. (2005). Evolution of regulatory interactions controlling floral asymmetry. Development 132: 5093–5101.

Devadas, S.K. and Raina, R. (2002). Preexisting systemic acquired resistance suppresses hypersensitive response-associated cell death in *Arabidopsis* hrl1 Mutant. Plant Physiol. 128: 1234–1244.

Ding, P. and Ding, Y. (2020). Stories of Salicylic Acid: a Plant defense hormone. Trends Plant Sci. 25: 549–565.

Dunning, D. (2011). The dunning-kruger effect. On being ignorant of one's own ignorance (Academic Press).

Eric Schranz, M., Mohammadin, S., and Edger, P.P. (2012). Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Curr. Opin. Plant Biol. 15: 147–153.

Freeling, M., Lyons, E., Pedersen, B., Alam, M., Ming, R., and Lisch, D. (2008). Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. Genome Res. 18: 1924–1937.

Fukuda, M., Yanai, Y., Nakano, Y., Sasaki, H., Uragami, A., and Okada, K. (2017). Isolation and gene expression analysis of flowering-related genes in lettuce (*Lactuca sativa* L.). Hortic. J. 86: 340–348.

Gaffney, T., Friedrich, L., Vernooij, B., Negrotto, D., Nye, G., Uknes, S., Ward, E., Kessmann, H., and Ryals, J. (1993). Requirement of salicylic acid for the induction of systemic acquired resistance. Science 261: 754–756.

Galego, L. and Almeida, J. (2002). Role of *DIVARICATA* in the control of dorsoventral asymmetry in *Antirrhinum* flowers. Genes Dev. 16: 880–891.

Giesbers, A.K.J., Pelgrom, A.J.E., Visser, R.G.F., Niks, R.E., Van den Ackerveken, G., and Jeuken, M.J.W. (2017). Effector-mediated discovery of a novel resistance gene against *Bremia lactucae* in a nonhost lettuce species. New Phytol. 216: 915–926.

Guo, Z. et al. (2022). LettuceGDB: The community database for lettuce genetics and omics. Plant Commun.: 100425.

Gust, A.A. and Felix, G. (2014). Receptor like proteins associate with SOBIR1-type of adaptors to form bimolecular receptor kinases. Curr. Opin. Plant Biol. 21: 104–111.

Hartmann, M. and Zeier, J. (2019). N-hydroxypipecolic acid and salicylic acid: a metabolic duo for systemic acquired resistance. Curr. Opin. Plant Biol. 50: 44–57.

Hoang, N. V. et al. (2022). The genome of Gynandropsis gynandra provides insights into whole-genome duplications and the evolution of C4 photosynthesis in Cleomaceae. bioRxiv: 2022.07.09.499295.

Hon, T. et al. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. Sci. Data 7: 1–11.

Jain, M. et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. 36: 338–345.

Kardailsky, I., Shukla, V.K., Ahn, J.H., Dagenais, N., Christensen, S.K., Nguyen, J.T., Chory, J., Harrison, M.J., and Weigel, D. (1999). Activation tagging of the floral inducer FT. Science 286: 1962–1965.

Lin, T. et al. (2022). Extensive sequence divergence between the reference genomes of *Taraxacum kok-saghyz* and *Taraxacum mongolicum*. Sci. China Life Sci. 65: 515–528.

Martin-Arevalillo, R., Nanao, M.H., Larrieu, A., Vinos-Poyo, T., Mast, D., Galvan-Ampudia, C., Brunoud, G., Vernoux, T., Dumas, R., and Parcy, F. (2017). Structure of the Arabidopsis TOPLESS corepressor provides insight into the evolution of transcriptional repression. Proc. Natl. Acad. Sci. U. S. A. 114: 8107–8112.

Moll, K.M., Zhou, P., Ramaraj, T., Fajardo, D., Devitt, N.P., Sadowsky, M.J., Stupar, R.M., Tiffin, P., Miller, J.R., Young, N.D., Silverstein, K.A.T., and Mudge, J. (2017). Strategies for optimizing BioNano and Dovetail explored through a second reference quality assembly for the legume model, *Medicago truncatula*. BMC Genomics 18: 578.

Ng, M. and Yanofsky, M.F. (2001). Function and evolution of the plant *MADS-box* gene family. Nat. Rev. Genet. 2: 186–195.

Nurk, S. et al. (2022). The complete sequence of a human genome. Science 376: 44–53.

Palazzesi, L., Pellicer, J., Barreda, V.D., Loeuille, B., Mandel, J.R., Pokorny, L., Siniscalchi, C.M., Tellería, M.C., Leitch, I.J., and Hidalgo, O. (2022). Asteraceae as a model system for evolutionary studies: from fossils to genomes. Bot. J. Linn. Soc. 200: 143–164.

Pancaldi, F., Vlegels, D., Rijken, H., van Loo, E.N., and Trindade, L.M. (2022). Detection and analysis of syntenic quantitative trait loci controlling cell wall quality in Angiosperms. Front. Plant Sci. 13.

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. (2016). Evolution of gene duplication in plants. Plant Physiol. 171: 2294–2316.

Raaymakers, T.M. and Van Den Ackerveken, G. (2016). Extracellular recognition of oomycetes during biotrophic infection of plants. Front. Plant Sci. 7: 906.

Reyes-Chin-Wo, S. et al. (2017). Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. Nat. Commun. 8: 14953.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., and others (2010). Genome sequence of the palaeopolyploid soybean. Nature 463: 178–183.

Schoch, C.L. et al. (2020). NCBI Taxonomy: A comprehensive update on curation, resources and tools. Database 2020.

Sedlazeck, F.J., Lee, H., Darby, C.A., and Schatz, M.C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat. Rev. Genet. 19: 329–346.

De Smet, R., Adams, K.L., Vandepoele, K., Van Montagu, M.C.E., Maere, S., and Van De Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc. Natl. Acad. Sci. U. S. A. 110: 2898–2903.

Tanaka, T., Nishijima, R., Teramoto, S., Kitomi, Y., Hayashi, T., Uga, Y., and Kawakatsu, T. (2020). *De novo* genome assembly of the indica rice variety IR64 using linked-read sequencing and Nanopore sequencing. G3 Genes|Genomes|Genetics 10: 1495–1501.

Tank, D.C., Eastman, J.M., Pennell, M.W., Soltis, P.S., Soltis, D.E., Hinchliff, C.E., Brown, J.W., Sessa, E.B., and Harmon, L.J. (2015). Nested radiations and the pulse of angiosperm diversification: Increased diversification rates often follow whole genome duplications. New Phytol. 207: 454–467.

Theißen, G., Kim, J.T., and Saedler, H. (1996). Classification and phylogeny of the *MADS-box* multigene family suggest defined roles of *MADS-box* gene subfamilies in the morphological evolution of eukaryotes. J. Mol.

Evol. 43: 484–516.

Underwood, C.J. et al. (2022). A *PARTHENOGENESIS* allele from apomictic dandelion can induce egg cell division without fertilization in lettuce. Nat. Genet. 54: 84–93.

Vijverberg, K., Welten, M., Kraaij, M., van Heuven, B.J., Smets, E., and Gravendeel, B. (2021). Sepal identity of the pappus and floral organ development in the common dandelion (*Taraxacum officinale*; Asteraceae). Plants 10: 1682.

Wang, B., Yang, X., Jia, Y., Xu, Y., Jia, P., Dang, N., Wang, S., Xu, T., Zhao, X., Gao, S., Dong, Q., and Ye, K. (2022). High-quality *Arabidopsis thaliana* genome assembly with Nanopore and HiFi long Reads. Genomics, Proteomics Bioinforma. 20: 4–13.

Wang, L., Ko, E.E., Tran, J., and Qiao, H. (2020). TREE1-EIN3–mediated transcriptional repression inhibits shoot growth in response to ethylene. Proc. Natl. Acad. Sci. U. S. A. 117: 29178–29189.

Wang, Y.C., Hu, D.W., Zhang, Z.G., Ma, Z.C., Zheng, X.B., and Li, D.B. (2003). Purification and immunocytolocalization of a novel *Phytophthora boehmeriae* protein inducing the hypersensitive response and systemic acquired resistance in tobacco and Chinese cabbage. Physiol. Mol. Plant Pathol. 63: 223–232.

Wei, B., Zhang, J., Pang, C., Yu, H., Guo, D., Jiang, H., Ding, M., Chen, Z., Tao, Q., Gu, H., Qu, L.J., and Qin, G. (2015). The molecular mechanism of SPOROCYTELESS/NOZZLE in controlling *Arabidopsis* ovule development. Cell Res. 25: 121–134.

Wei, T. et al. (2021). Whole-genome resequencing of 445 *Lactuca* accessions reveals the domestication history of cultivated lettuce. Nat. Genet. 53: 752–760.

Wenger, A.M. et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 37: 1155–1162.

Yalpani, N., Silverman, P., Wilson, T.M., Kleier, D.A., and Raskin, I. (1991). Salicylic acid is a systemic signal and an inducer of pathogenesis-related proteins in virus-infected tobacco. Plant Cell 3: 809–818.

Youssef, S., Abd Elhady, S., Abu El-Azm, N., and El-Shinawy, M. (2017). Foliar application of salicylic acid and calcium chloride enhances growth and productivity of lettuce (*Lactuca sativa*). Egypt. J. Hortic. 44: 1–16.

Zhang, L. et al. (2017). RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. Nat. Commun. 8: 2264.

Zhao, Y., Broholm, S.K., Wang, F., Rijpkema, A.S., Lan, T., Albert, V.A., Teeri, T.H., and Elomaa, P. (2020). TCP and *MADS-box* transcription factor networks regulate heteromorphic flower type identity in *Gerbera hybrida*. Plant Physiol. 184: 1455–1468.

Summary

Improved crops are needed to provide healthy and nutritious food for the growing human population. A genome assembly depicts a map for exploring genetic diversity and chromosome organization of plants. The knowledge gained from genome sequencing can help breed improved crop plant cultivars. Lettuce (*Lactuca sativa* L.) is an important leafy green vegetable and constitutes a significant component of a healthy diet of many people worldwide, which makes it and its wild relatives important targets for genome sequencing. **Chapter 1** presented the development and types of sequencing technologies required for genome assembly. Then, three important species for lettuce breeding were sequentially introduced. Firstly, the two wild lettuce species *L. saligna* and *L. virosa*, primarily used in hybridization breeding, were introduced with a focus on their resistances to lettuce downy mildew (*Bremia lactucae*) and aphid (*Nasonovia ribisnigri*), respectively. Genome-editing (GEd) vastly expands the scope of the gene pool enabling more species accessible to plant breeding. Considering this, sexual diploid and asexual triploid dandelion (*Taraxacum officinale*), as the outgroup of lettuce, can be possible donors of important traits, like parthenogenesis (i.e., asexual reproduction via clonal seed). In addition to lettuce breeding, *Lactuca* and *Taraxacum* represent the subfamily Cichorioideae of Asteraceae, the most diverse flowering plant group. The biological understanding of Asteraceae trait evolution, like tof its iconic capitulum flowers, could be advanced by genome sequencing and phylogenomic analyses. This thesis presents the *de novo* genome assemblies for the three species mentioned above and demonstrates their genome compatibility, genetic diversity and trait evolution for varying scopes from *Lactuca* species to the Asteraceae family.

This dissertation starts with *L. saligna*, which is partially interfertile with lettuce and is part of the secondary gene pool of lettuce. Hybrids of *L. saligna* and lettuce show reduced fertility, caused by hybrid incompatibilities (HI). *Lactuca saligna* is an attractive donor of broad and durable resistance to *B. lactucae,* the cause of lettuce downy mildew disease. *Lactuca saligna* displays complete or non-host resistance (NHR) and probably is induced by a combination of epistatic genes for pathogen-recognition receptors, including nucleotide-binding leucine-rich repeats (NLRs) and receptor-like kinases (RLKs). In **Chapter 2**, I presented a high-quality assembly of *L. saligna* with the whole-genome resequencing (re-seq) of 15 accessions across Eurasia and the RNA-sequencing (RNA-seq) of a *Bremia* infection bioassay. Based on SNP variants derived from re-seq data, population genetics analysis classified the 15 accessions into three sub-clades, which are aligned with their geographical origin. Comparative genomic analysis with lettuce revealed two lineage-specific inversions (>50 Mb) in *L. saligna* compared to *L. sativa*. These inversions both overlapped with HI intervals, thus likely hampering the introgression of genes therein. A comprehensive inventory of *NLR*s and *RLK*s was built for the *L. saligna* genome and revealed genomic hotspots. Three hotspots were found to co-locate with previously identified NHR regions. The RNA-seq analysis identified

differentially expressed genes related to *Bremia* infection and plant immunity. Within NHR regions, a significant number of *RLK*s were up-regulated. Among them, three tandem duplicates of W*all-Associated Kinase (WAK)* genes were particularly interesting due to their high and induced expression.

Chapter 3 focused on *L. virosa,* from the lettuce tertiary gene pool. Although *L. virosa* is uncrossable with lettuce in nature, interspecific crossing can still be achieved by the bridge of *L. serriola* or embryo rescue to introduce agronomic or resistance traits. *L. virosa* (3.7 Gbp) has a significantly larger genome size in contrast to *L. sativa* (2.5 Gbp) or *L. saligna* (2.3 Gbp). In this chapter, a *de novo* genome assembly of *L. virosa* was constructed, resulting in high continuity and complete gene space. Homolog grouping of *L. sativa*, *L. saligna* and *L. virosa* classified the core, accessory and unique gene sets. The three-way genome comparison revealed interspecific inversions that would cause recombination suppression during introgression hybridizations. Moreover, reference-based and reference-free repeatome-comparisons collectively validated a proliferation of long-terminal repeat elements in *L. virosa* compared to other two *Lactuca* spp. with smaller genome sizes. Similar to **Chapter 2**, *NLR* and *RLK* genes were also identified and classified forthe *L. virosa de novo* genome. By associating with homology and genomic position, I showed that the *RLK*s are commonly more conserved among *Lactuca* species, while tandem duplication is the main factor driving the expansion of *NLR*s and *RLK*s.

A *de novo* sequencing project of sexual type (diploid) dandelion (*T. officinale*) was described in **Chapter 4**, which adds a valuable outgroup reference for lettuce breeding. The dandelion *de novo* reference was combined with assemblies mentioned in **Chapters 2** and **3**, plus 30 other genomes from public databases, to perform a phylogenomic study of r gene and flower trait evolution in the Asteraceae family. The search and classification of flower-related *MADS-box* and *TCP* genes indicated expansions of *MIKC*^{c} and *CYC* sub-clades. Phylogenomic analysis of synteny further found *SEEDSTICK*-like (*STK*-like), *SEPALATA*-Like (*SEP3*-Like), and *PCF*-Like copies with lineage-specific genomic-contexts (i.e., synteny) in the Asteraceae or dandelion. Subsequently, the RNA-seq data of dandelion demonstrated the different expression patterns between genes in conserved and specific contexts. The phylogenomic study also revised the origin and evolution of the previously identified *Asteraceae Specific MADS-box* genes (*AS-MADS*). The phylogeny suggested a more ancient origin of the *AS-MADS* clade, while synteny showed it is paralogous to the *FLC* genes. In reference to dandelion, the expression of AS-MADS is clearly distinguished from the other *FLC* genes.

In **Chapter 5**, I discussed my thesis as a whole and horizontally dissected it to synthesize perspectives for future studies. The evaluation of sequencing statistics showed that our strategy was robust to construct high-quality Asteraceae plant genome assemblies. The

investigation of scaffolding conflicts and assembly incompleteness further emphasizes the sequence depth and order of different scaffolding platforms as crucial factors for genome assembling. Then, the concept of model plants for biological studies was placed and assessed in the current period. My *de novo* genomes were assessed for their contribution to model systems of leafy crops and ecological evolution integrated with ongoing sequencing projects. For resistance traits, I speculated the potential roles of receptor-like protein (RLP) and salicylic acid (SA) hormone in resistance to lettuce downy mildew. For floral trait evolution, an assessment was performed on the omitted *MADS-box* copies in the current study and other related transcription factors family (*MYB* and *TOPLESS*). The complex of gene duplication patterns was also addressed. The positive correlation of tandem and transposed duplication for immune genes was compared to previous research. Extreme examples of gene retention were observed between *MADS-box* and *SARD4* is raised, and a hypothesis was initiated as its underlying mechanism. Finally, a reflection was taken focusing on the efficiency of resource usage and research depth. Finally, I give a general plan, based on my thesis and its contents, to illustrate an improved template for future studies.

Acknowledgements
About the author
List of publications
Education statement

Acknowledgements

The PhD journey is a story of exploration and growth, finally, I am nearing its curtain call. The path I have traveled has been filled with difficulties, and challenges, but as Winston Churchill once said, 'Success is not final; failure is not fatal: it is the courage to continue that counts.' I am proud of myself for never losing the courage to continue, even when faced with difficult obstacles. However, my completion would not have been possible without the accompany and encouragement I received from countless individuals. Their support has been a beacon of hope, helping me to navigate through moments of vulnerability and uncertainty. Thus, I would like to express my sincere gratitude to all the people who helped me through the journey.

First of all, my most heartfelt thank goes to my promotor Eric Schranz. **Eric**, thank you for giving me the opportunity and providing me with the guidance and resources needed to excel in my research, like the synteny study for the dandelion project. Your expertise, creativity, and rigorous approach to investigating biological questions have been truly inspiring. I appreciate the great extent of liberty you gave me to pursue my research, and your uninterrupted encouragement was instrumental in keeping me motivated. I am grateful for your positive attitude and unwavering support whenever I encountered any obstacles. Thank you for celebrating my achievements, no matter how small, which greatly helps to build my confidence. I appreciate your efforts in creating a stimulating and collaborative research environment. Benefiting from this, I was involved in numerous exciting projects additional to my PhD, including parthenogenesis and gynandropsis papers. These projects have provided me with a unique opportunity to collaborate with many talented scientists and expand my professional network. Furthermore, I am grateful for your understanding and support when I faced personal challenges, especially during the pandemic period. Your kindness and encouragement helped me to stay focused on my research and achieve my goals. Thank you for being an outstanding mentor and role model.

I would like to express my sincere gratitude to the important collaborators who have contributed to my research. Firstly, I extend my thanks to Klaas Bouwmeester, Marieke Jeuken, Sandra Smit, and Kitty Vijverberg, whom I deeply respect and consider as my co-promotors. Their contributions have been invaluable in shaping the direction and scope of my research, and I am grateful for their guidance and expertise throughout the process.

Dear **Klaas**, I want to express my heartfelt appreciation for your unwavering support throughout my PhD journey. You have been like a daily supervisor to me, guiding me through the analysis and writing of the *Lactuca* papers. I learned how to plan and write a scientific and rigorous paper from you, and I am grateful for the challenging questions you asked that pushed me to think deeply. Your expertise and perspective on plant immunity and pathology have been invaluable to me, and I always enjoyed our conversations on science and bird-watching. Most of all, I am grateful for your encouragement and support during the difficult times. You were always there for me, lifting me up with your kind words and suggestions. Thank you for being a mentor, a colleague, and a friend.

Marieke, as my study advisor during my master's degree, I consider myself fortunate to have had your guidance and support throughout my PhD research. I want to express my sincere appreciation for your patience and expertise in the areas of resistance, genetic mapping, and lettuce breeding. Your vast knowledge and experience in breeding complemented my research and helped me overcome many challenges. I am grateful for your quick responses to my requests for data and meetings, and for your calm and kind demeanor that always put me at ease. Furthermore, thank you for teaching me how to design a proper bioassay and for your assistance with the lettuce growth.

Sandra, as my external supervisor and later the role of co-promoter for my *Lactuca virosa* study, I am grateful to have had the opportunity to work with you. Your rigorous attitude and attention to detail have been instrumental in shaping me into a better genomic data scientist. I appreciate your high standards for data management, which have benefited me greatly. Thank you for taking the time to review my manuscript, even when you were super busy. I still remember receiving notifications from you late at night and early in the morning as you reviewed my paper. Your comments were always constructive and objective, and I also appreciate that you initiated equal communication by encouraging me to take the lead and overrule the suggestions you gave. Thank you for your invaluable guidance and support throughout my research.

Kitty, I have known you since the beginning of my PhD. I remember how kind and approachable you were when I was new to the Biosys group. Although your contract ended shortly after, I am grateful that we reconnected through your incredible work on dandelion parthenogenesis published in Nature Genetics, as well as our collaboration on the dandelion flower development paper. Thank you for introducing me to the fascinating world of flower development in dandelions and Asteraceae. I have learned from you the importance of a hands-on approach and determined persistence in research. Your career tips have also been incredibly helpful. Despite the COVID-19 pandemic, our collaboration was a success through online meetings and emails. I enjoyed our conversations and chitchats, which served as a great accompaniment to me during that challenging time.

To **DJ**, **Judith**, and **Siva**, who collaborated closely with me on the de novo genome sequencing projects, I want to express my heartfelt gratitude for your assistance in the genomic analysis. You are all highly experienced and professional bioinformaticians whom I deeply admire and look up to. **DJ**, I am truly grateful for your exceptional organizational skills and efficiency, which were critical in completing the two Lactuca genome papers. **Judith**, thank you for your remarkable dedication in tackling the intricate dandelion genome, which laid the foundation for our later research and analysis. I also appreciate your patience in dealing with my numerous emails about data updates. **Siva**, my dear friend and collaborator, I cannot thank you enough for sharing your invaluable life experience and industrial perspective with me. Our conversations about bioinformatics, food, and culture were enlightening and always enjoyable.

I am grateful to the Chinese Scholarship Council (CSC) for awarding me a PhD fellowship to pursue my studies in the Netherlands.

I would like to express my gratitude to my thesis committee members: Prof. Dr. **Richard Immink**, Dr. **Robert van Loo**, Dr. **Dmitry Lapin**, and Dr. **Saulo Alves Aflitos**. Thank you for the time and effort you devoted to evaluating my work.

To my paranymphs **DJ** and **Xi**, I am fortunate to have such wonderful people like you in my personal and professional life throughout my PhD journey. DJ, as my esteemed collaborator, and Xi, as my dear friend. I want to thank you for supporting me during my PhD defense and celebration event.

I would like to express my gratitude to my Biosystematics group colleagues. Our group is like a diverse family, a melting pot of different cultures and backgrounds. I appreciate the enjoyable vibe and friendly atmosphere and thank you for being a great colleague and friend. I would like to thank the following individuals for their support: **Wilma**, thank you for your help and support throughout my PhD project. Your devoted work and kind caring made each step easier, and I will miss our life stories and snack exchanges. **Patrick**, you are an inspiration with your energy and enthusiasm. I am grateful for your help with RNA-sequencing and material organization. **Freek**, your amazing lecture on phylogeny and subsequent help were invaluable. Your trumpet performance also created many memorable moments for me. **Robin**, thank you for your expertise on phylogeny and population genetics, as well as for your social and humorous nature. Our chats were always enjoyable, and I appreciate your career path suggestions. **Nina**, thank you for being so nice and involving me in group events when I first joined. Our talks about basketball and other hobbies were always fun. **Casper**, you are a warm and gentle person, and it is always a pleasure to chat with you. I really enjoyed our previous squash matches, and I hope we can play again in the future. **Li**, my dear friend, and comrade thank you for your accompany during the toughest times of thesis writing. Our conversations over lunch breaks, about not only our thesis but also life and future, were like shining stars that guided me through the darkest moments. I truly appreciate your friendship and encouragement. I wish you all the best in your future, and I am sure you will achieve great success. **Wouter**, your confidence and positive attitude always lifted me up. I enjoyed our conversations about the economy, society, and philosophy. **Cloe**, I respect you as a great scientist and creative artist. Thank you for helping me with protein modeling and teaching me how to create elegant figures in PowerPoint. **Marieke**, thank you for encouraging me and sharing your invaluable experience with life and research. I wish you all the best in your future career. **Nam**, your work on the "Gynandropsis" paper was impressive, and I learned a lot from your expertise. Thank you for always being fun and nice. **Tao**, I was inspired by your passion for science and your insights as an evolutionary biologist. Your balancing of life and work with your badminton, Chinese flute, and calligraphy practices is admirable. I wish you and **Xiaoyan** a wonderful stay in the Netherlands. **Chao** and **Lei**, thank you for the delightful conversations about science and society, as well as your life experience and suggestions for pursuing an academic career in China. I hope we can work together in the future. **Eva**, **Liana**, and **June**, thank you for your amazing work on the party committee. You made the last stage of my PhD a memorable and restful experience.

The gratitude and miss also go to my former colleagues: Dear **Tao Zhao** and **Zhen Wei**, my big brother and sister in biosystematics, I will never forget the opportunity you gave me to pursue a PhD project in this exciting field. I learned so much from your expertise in lettuce phylogeny and synteny studies, which provided a solid foundation for my own research. Even after your graduation, your support and care have been unwavering and invaluable. I miss the casual time we spent together in Wageningen and I look forward to meeting you both in China soon. Best of luck with your research! **Niccolo**, I appreciate the support and camaraderie we provided each other throughout our PhD journeys. It was great to have you as a member of our silence room gang, and I look forward to trying your Italian food in Utrecht. **Deedi**, thank you for your help with variant calling and GWAS studies. I wish your business in Benin a bright future. **Floris**, your dedication to family and science is admirable. Thank you for your assistance with repeatomic analysis. **Thijmen**, I always think of your fashion and humor, and I appreciate your help with transcriptomic analysis. I hope your start-up business in the UK is successful. **Nora**, thank you for sharing great moments and conversations about digital games. Your assistance with population evolution and synteny studies was invaluable. I cherish the memories of our time together and appreciate your invitation to your home in Wageningen and Germany. Best wishes to you, John, and Fay. I also miss **Lotte** and **Sara**, thank you for the companionship and support. I miss the coffee breaks and enjoyable conversations we shared.

I would like to express my gratitude to several individuals who provided invaluable support and guidance beyond the Biosystematics group. **Frank** from the Genetics group also deserves my thanks for his support during experimental work. Additionally, I am grateful to **Francesco** from the Plant Breeding group for sharing his exceptional genome data with me. I would also like to extend my appreciation to **Susan**, the coordinator from the EPS graduate school, for her tireless efforts in guiding me through the administrative and educational aspects of my journey. I want to thank **Xu**, a senior whose advice and suggestions have been instrumental in my professional development. I would also like to thank **Pingtao** for his encouragement, academic support, and help in career planning. I am indebted to **Chengcheng** from the Plant Breeding group and **Zhaodong**, a guest researcher from Nanjing Forestry University in China, for their invaluable assistance in my bioinformatic analysis. To **Wenqin**, I would like to express my gratitude for sharing your expertise on greenhouses and vertical farming. Finally, I would like to express my gratitude to **Erik** from Proefschriften.nl company and **Ilse** from Ilse Modder Grafisch Ontwerper for their help with printing and designing my thesis, ensuring that the final step of my PhD journey was completed smoothly and professionally.

I would like also to thank my MSc students for the nice experience working with them: **Xiaoyue**, **Iris**, and **Sybren**. Thank you all for trusting me and giving me the chance to grow with you. I wish you all a bright future.

I would like to thank the friends I met during my PhD journey. First and foremost, I want to greet my dearest gang of friends since my master's degree: **Wuhua**, **Wenhao**, **Ying** & **Yang** (aka Yang couple), **Zhengcong**, **Liyou**, **Hao**, **Shuo**, **Weijia**, **Lei**, **Miaoying**, and **Ziwei**. I have countless memories full of laughter and happiness with each one of you, and I don't know where to start. You have always been there for me through the ups and downs, and for that, I cannot express my sincere gratitude enough. I also want to extend great gratitude to **Xi** and **Pepijn**, the best and sweetest couple in the world. Thank you for organizing all the beautiful events. I cherish your company a lot. Our explorations, travels, and culinary hunting activities are the highlights of my Ph.D. Every time I finish dinner with you and lounge on the sofa watching TV, I feel as relaxed as I would at home. To all of you, as the old saying goes, everything I want to say is in the wine. Cheers to our friendship!

I would also like to thank **Da Lao**, not only for being my friend but also for being an elder figure who has shared all your life wisdom with me. I will always cherish the amazing food you have cooked for me whenever I sneak into your kitchen and all the chitchat we've had behind the checkout counter.

I would also like to express my sincere gratitude to the many friends who have supported and accompanied me throughout my PhD journey. **Huaiyu** , **Zilin** & **Yan**, **Rubing**, **Xianhui** (the party queen of Dijkgraaf), **Wenqin** & **Baozi** (the harmony song of science and art), **Xinxin** (my esteemed hair-style director), **Pol** (the king of fermentation), **Chuanbao**, **Yifeng**, **Meixiu**, **Qiaofeng** and **Xunkun** (Versatile chef), thank you all for your support and accompany. I will never forget the wonderful moments we shared. Thank you for being my friends, spending time with me, and helping me to grow and broaden my horizons.

I often joke about myself that I moved six times during my PhD, which was really exhausting. However, I am truly grateful for the wonderful friends I made during those moves. **Yuxi** and **Jun**, we were roommates when I started my PhD while you were both about to finish yours. Witnessing the phases of your PhDs prepared me for my research. Thank you both for being such great role models for me. I also appreciate **Yang** & **Wang's** family for offering me shelter during the final phase of my thesis writing. Your harmonious family atmosphere made me feel a lot of warmth. I would also like to thank **Chengcheng,** watching cartoons with you was one of the happiest moments during my thesis writing. Additionally, I want to thank my three-day roommates, **Momo** & **Dandan**, for all the kind care and suggestions during my thesis writing. I feel honored to have witnessed the birth of your little baby. Thank you for sharing with me the joy of being new parents and a lot of parenting experience.

Last but the most important, to my beloved family. 爸爸妈妈，感谢你们为我成长所付出的辛劳。离家已有十多年，你们一直是我最坚实的后盾。感谢你们对我所做的决定的支持和鼓励，感谢你们的宽容和理解，你们的爱是我前进的力量。谢谢心语，和我一起长大，一起成熟，一起发paper，brother & sister forever! 最后，外公和奶奶，我没有辜负你们的期待, 我做到了。

About the author

Wei Xiong was born in the Yongding village in Fujian, China on November 27th, 1991, and moved to Xiamen city with his parents at 7 years old. In 2010, he began his Bachelor's degree at Nanjing Agricultural University, specializing in Seed Science and Engineering.

But Wei's true adventure began in 2014 when he embarked on a new chapter in the Netherlands. Driven by his passion for plant breeding, Wei pursued a Master's degree in Plant Breeding and Genetic Resources at Wageningen University. It was here that Wei's academic journey set sail as he worked on his thesis at the Bio-based Economy group under the supervision of Prof. Dr. ir Luisa Trindade. His thesis aimed to develop a predictive model for high-throughput phenotyping of hemp fiber composition, utilizing near-infrared spectroscopy and biochemical data. Wei continued his exploration by taking an internship at Thatchtec B.V. in the Netherlands. From seed products to breeding methods, Wei developed a keen interest in various biological topics during his studies. To gather his thoughts and clear his mind for his next move, Wei decided to take a gap year. During this time, he joined the Wageningen Student Council campaign and was elected in 2016. Striving for sustainability and integration, Wei served the students of Wageningen.

Following his stint in the Student Council, Wei made up his mind to pursue a PhD in exploring fundamental questions of biology, particularly in the evolution of plant genomes, genes, and traits. In 2017, he received a grant from the Chinese Scholarship Council and began his doctoral studies at the Biosystematics group at Wageningen University and Research, supervised by the brilliant Prof. Dr. M.E. Schranz. Wei's journey was not without challenges, but he rose to them. During his PhD studies, he also worked as an external researcher of the "LettuceKnow" consortium and co-founded a journal club of Chinese scholars in the plant science field at Wageningen University and Research. Wei's academic and personal growth is a testament to his dedication and passion for academia, and his desire to make an impact in the world of society.

Contact me via:
wei.xiong2023@hotmail.com

List of publications

Wei Xiong, Lidija Berke, Richard Michelmore, Dirk-Jan M. van Workum, Frank F.M. Becker, Elio Schijlen, Linda V. Bakker, Sander Peters, Rob van Treuren, Marieke Jeuken, Klaas Bouwmeester, M. Eric Schranz (2022). The genome of *Lactuca saligna*, a wild relative of lettuce, provides insight into non-host resistance to the downy mildew *Bremia lactucae*. bioRxiv: 2022.10.18.512484. (Accepted by "The Plant Journal")

Wei Xiong+, Dirk-Jan M. van Workum+, Lidija Berke, Linda V. Bakker, Elio Schijlen, Frank F.M. Becker, Henri van de Geest, Sander Peters, Richard Michelmore, Rob van Treuren, Marieke Jeuken, Sandra Smit, M. Eric Schranz (2023). Genome assembly and analysis of *Lactuca virosa*: implications for lettuce breeding. (submitted and under review)

Wei Xiong+, Judith Risse+, Lidija Berke, Tao Zhao, Henri van de Geest, Carla Oplaat, Marco Busscher, Ingrid van der Meer, Koen Verhoeven, M. Eric Schranz, Kitty Vijverberg (2023). Phylogenomic and gene expression analysis of de novo genome and transcriptome sequencing of dandelion (*Taraxacum officinale*) provide insights into *MADS-box* and *TCP* gene diversification and floral development of the Asteraceae. (submitted and under review)

Wei Xiong, Sivasubramani Selvanayagam, Marieke Jeuken, Klaas Bouwmeester, M. Eric Schranz. Comparative transcriptomics to identify and validate non-host resistance genes in *Lactuca saligna* against *Bremia lactucae*. (in preparation)

Charles J. Underwood+, Kitty Vijverberg+, Diana Rigola+, Shunsuke Okamoto, Carla Oplaat, Rik H. M. Op den Camp, Tatyana Radoeva, Stephen E. Schauer, Joke Fierens, Kim Jansen, Sandra Mansveld, Marco Busscher, **Wei Xiong** et al. (2022). A PARTHENOGENESIS allele from apomictic dandelion can induce egg cell division without fertilization in lettuce. Nat. Genet. 54: 84–93

Nam V Hoang+, E O Deedi Sogbohossou+, **Wei Xiong** et al. (2022). The Gynandropsis gynandra genome provides insights into whole-genome duplications and the evolution of C4 photosynthesis in Cleomaceae. The Plant Cell.

+ Contributed equally.

Education Statement of the Graduate School Experimental Plant Sciences

Issued to:	**Wei Xiong**
Date:	**28 March 2023**
Group:	**Biosystematics**
University:	**Wageningen University & Research**

1) Start-Up Phase	*date*	*cp*
▶ **First presentation of your project**		
Genome analysis of Lactuca saligna provides insights into non-host resistance and evolution of lettuce	2018/05/02	1.5
▶ **Writing or rewriting a project proposal**		
Comparative genomics and trait evolution of lettuce wild relatives	2020/05/07	6.0
▶ **Writing a review or book chapter**		
▶ **MSc courses**		
Subtotal Start-Up Phase		7.5

2) Scientific Exposure	*date*	*cp*
▶ **EPS PhD student days**		
EPS Get2Gether 2018	2018/02/15 - 2018/02/16	0.6
EPS Get2Gether 2019	2019/02/11 - 2019/02/12	0.6
EPS Get2Gether 2020	2020/02/10 - 2020/02/11	0.6
EPS Get2Gether 2021	2021/02/01 - 2021/02/02	0.4
▶ **EPS theme symposia**		
EPS theme 4 symposium 'Genome Biology'	2018/09/25	0.3
EPS theme 4 symposium 'Genome Biology'	2019/12/13	0.3
EPS theme 4 symposium 'Genome Biology'	2020/12/11	0.2
EPS theme 4 symposium 'Genome Biology'	2022/01/17	0.3
EPS theme 2 symposium 'Interactions between plants and biotic agents'	2022/02/08	0.2
▶ **Lunteren Days and other national platforms**		
Annual meeting 'Experimental Plant Sciences'	2018/04/09 - 2018/04/10	0.6
Annual meeting 'Experimental Plant Sciences'	2019/04/08 - 2019/04/09	0.6
Annual meeting 'Experimental Plant Sciences'	2021/04/12 - 2021/04/13	0.5
▶ **Seminars (series), workshops and symposia**		
Ritzema Bos Lecture – Prof. Nick Talbot	2019/04/02	0.1
"Population Genomic Analyses Reveal Connectivity via Human-Mediated Transport across Populus Plantations in North America and an Undescribed Subpopulation of Sphaerulina musiva" - Jared LeBoldus	2020/09/15	0.1
"Evolutionary dynamics of NLR immune receptors in plants" - Dr. Sophien Kamoun and Dr. Phil Carella	2020/09/16	0.2
"Challenges as a PhD and postdoc in Science and tips to overcome those" - Stefan Geisen	2021/01/21	0.1
Wageningen Evolution & Ecology Seminar (WEES): "The hunt for our molecular past" - Prof. Eske Willerslev	2021/03/17	0.1
2nd EPSO Seminar: Prof Corné M.J. Pieterse "The plant microbiome and plant health"; Prof Paola Bonfante "Arbuscular mycorrhizal fungi: living in between plants and endobacteria"; Prof Gabriel Castrillo "Coordination between the microbiota and the root endodermis is required for plant mineral nutrient homeostasis"	2021/04/15	0.2
"Modern Science Communication" - Prof. Dr. Elizabeth (Liz) Haswell	2021/05/11	0.1
Novogene Webinar: "RNA-seq Results Explained: what you can expect from the analysis"	2022/02/17	0.1
KeyGene Webinar: "Apomixis: the breakthrough breeding technology for the 2020s"	2022/02/17	0.1
Plantae Presents: "How to Read a Scientific Paper" - Dr. Facette	2023/01/17	0.1
Plant Cell Webinar: "Plant Responses to Abiotic Stress" Leia Colin, Sophie Filleur, and Yong-Fei Wang	2023/02/07	0.2
Symposium: "Land Plant Evolution & Improving Photosynthesis and Crops"	2019/06/20	0.3
▶ **Seminar plus**		
▶ **International symposia and congresses**		
Plant Genome in a Changing Environment, Wellcome Genome Campus, Cambridge, UK	2018/10/24 - 2018/10/26	0.9
EMBL Visualizing biological data (VIZBI) 2021 conference, online	2021/03/24 - 2021/03/26	0.9
▶ **Presentations**		
Poster presentation at VIZBI 2021 conference & Annual meeting 'Experimental Plant Sciences' 2021	2021/03/25 & 2021/04/12	1.0
Oral presentation at EPS theme 4 symposium	2020/12/11	1.0
Oral presentation at LettuceKnow user meeting	2021/10/27	1.0
Oral presentation at LettuceKnow user meeting	2022/05/17	1.0
▶ **IAB interview**		
▶ **Excursions**		
Online Company Visit @ Rijk Zwaan	2021/06/16	0.2
Subtotal Scientific Exposure		12.9

3) In-Depth Studies	*date*	*cp*
▶ **Advanced scientific courses & workshops**		
EPS course 'Phylogenetics: Principles & Methods'	2018/04/23 - 2018/04/26	1.2
Software Carpentry workshop 'Python programming'	2019/01/14 - 2019/01/15	0.6
EPS online workshop 'Snakemake'	2020/05/13	0.3
PE&RC/EPS/WIMEK course 'Introduction to machine learning'	2021/06/28 - 2021/07/02	1.5
EHBIO Gene Technology course 'Advanced analysis of RNA-sequencing data'	2021/11/27 - 2021/11/29	0.8
WUR workshop 'Specificity and side-effects of mutagenesis by CRISPR-Cas in plants'	2022/08/31	0.3
LettuceKnow GWAS Hackathon	2023/01/31	0.3
▶ **Journal club**		
▶ **Individual research training**		
Subtotal In-Depth Studies		5.0

4) Personal Development	*date*	*cp*
▶ **General skill training courses**		
WGS PhD Competence Assessment	2018/02/14	0.3
EPS Introduction Course	2018/03/27	0.3

CONTINUED ON NEXT PAGE

	WGS course 'Brain Training'	2018/04/11	0.3
	WGS course 'Scientific writing'	2019/03/12 - 2019/04/30	1.8
	WGS course 'Infographics and Iconography'	2019/05/07	0.2
	WGS PhD Workshop Carousel 2019	2019/05/24	0.3
	WGS course 'Adobe Indesign Essential Training'	2019/11/04 - 2019/11/05	0.6
	WGS workshop 'Reviewing a Scientific Manuscript'	2020/11/12	0.1
	WGS course 'Efficient Writing strategies'	2021/04/05 - 2021/06/07	1.3
	EPS course 'How to make a movie about your research'	2021/03/02 & 2021/03/30	0.6
	WGS Career assessment	2021/04/23	0.3
►	**Organisation of meetings, PhD courses or outreach activities**		
	Founder and organiser of Seminar for Chinese Plant Scientists in WUR * 10 times	2020/09/06 - 2021/06/13	1.5
►	**Membership of EPS PhD Council**		
		Subtotal Personal Development	7.6

TOTAL NUMBER OF CREDIT POINTS*	**33.0**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.

** A credit represents a normative study load of 28 hours of study.*

This research was performed at the Biosystematics group, Wageningen University & Research, the Netherlands. The work described in this thesis is financed by a grant from the International Lettuce Genomics Consortium (ILGC) funded by the Top Consortium for Knowledge and Innovation Horticultural and Starting Materials (grant number 1406-039), and by grants from the Dutch Research Council Applied and Engineering Sciences for ParTool (grant number 13700). Wei Xiong was financially supported by a PhD fellowship from the China Scholarship Council (CSC). Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design: Wei Xiong, Ilse Modder || www.ilsemodder.nl
Layout: Ilse Modder || www.ilsemodder.nl
Printed by: Proefschriften.nl | | www.proefschriften.nl